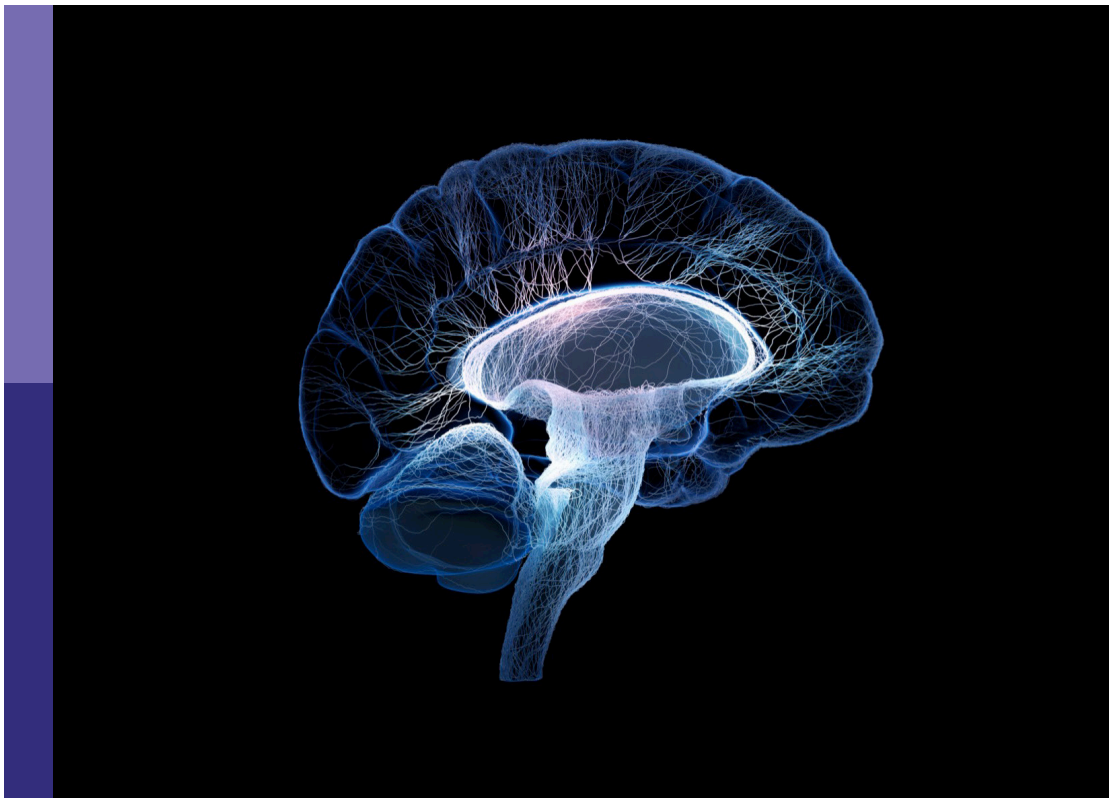
frontiers | Research Topics

# Neuroscience-driven visual representation

**Edited by**
Teng Li, Fudong Nian, Caifeng Shan, Jianfei Liu and Qieshi Zhang

**Published in**
Frontiers in Neuroscience



ISSN 1664-8714
ISBN 978-2-8325-5322-0
DOI 10.3389/978-2-8325-5322-0

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Neuroscience-driven visual representation

**Topic editors**

Teng Li — Anhui University, China

Fudong Nian — Hefei University, China

Caifeng Shan — Shandong University of Science and Technology, China

Jianfei Liu — National Eye Institute (NIH), United States

Qieshi Zhang — Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (CAS), China

**Citation**

Li, T., Nian, F., Shan, C., Liu, J., Zhang, Q., eds. (2024). *Neuroscience-driven visual representation*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5322-0

# Table of contents

frontiers | Frontiers in Neuroscience

TYPE Editorial
PUBLISHED 19 December 2023
DOI 10.3389/fnins.2023.1345688



OPEN ACCESS

EDITED AND REVIEWED BY
Benjamin Thompson,
University of Waterloo, Canada

*CORRESPONDENCE
Teng Li
liteng@ahu.edu.cn

RECEIVED 28 November 2023
ACCEPTED 04 December 2023
PUBLISHED 19 December 2023

CITATION
Li T, Zhang Q, Nian F and Shan C (2023) Editorial: Neuroscience-driven visual representation. *Front. Neurosci.* 17:1345688. doi: 10.3389/fnins.2023.1345688



# Editorial: Neuroscience-driven visual representation

Teng Li[1,2]*, Qieshi Zhang[3], Fudong Nian[4] and Caifeng Shan[5,6]

[1]Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, China, [2]Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei, China, [3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (CAS), Shenzhen, Guangdong, China, [4]School of Advanced Manufacturing Engineering, Hefei University, Hefei, Anhui, China, [5]College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China, [6]School of Intelligence Science and Technology, Nanjing University, Nanjing, China

KEYWORDS

**visual representation, computer vision, neuroscience, learning, deep neural networks**

Editorial on the Research Topic
Neuroscience-driven visual representation

Visual representation learning seeks to mimic the human visual system using deep neural networks, enabling machines to interpret digital images and video for diverse applications from manufacturing to energy. However, major gaps remain compared to biological vision, and most representation learning methods do not sufficiently incorporate neuroscientific and psychological principles. Key open questions persist around designing optimized architectures to extract meaningful representations from complex 2D or 3D scenes containing numerous heterogeneous, unlabeled examples.

While deep learning has achieved state-of-the-art results across various vision tasks like classification, detection and segmentation, core challenges in representation learning need to be tackled to reach human-level visual understanding. For instance, handling unlabeled, unstructured data and generalizing learned patterns to novel datasets continue to pose difficulties. Furthermore, lack of model interpretability is an issue that integration of biological approaches could help address.

This research area aims to advance visual representation learning through synergistic fusion of deep neural networks with psychological and neuroscientific concepts. By providing a platform to exchange cutting-edge techniques spanning both data-driven and theory-driven disciplines, impactful progress can be made toward biomimetic visual systems. Realizing more efficient, generalizable, and explainable visual learning has the potential to profoundly transform capabilities in scientific imaging, manufacturing, transport and healthcare.

Enhanced analysis of facial imagery for health assessment. Building on advanced computer vision techniques, Li et al. present a facial analysis methodology using convolutional neural networks (CNNs) to detect depression. They introduce innovations including multi-head attention modules and region-specific tuning to improve CNN sensitivity in analyzing different facial areas tied to depression. With further research, such AI-based systems could assist in mental health evaluation and screening.

Multi-constraint modeling for 3D shape reconstruction. Reconstructing 3D structure from 2D image sequences is an important but challenging computer vision task. Chen X. et al. put forth a multi-constraint estimation algorithm that first extracts shape bases via sparse coding, then estimates 3D geometry through a penalized least-squares model incorporating orthogonal and similarity constraints. Experiments demonstrated higher accuracy compared to existing methods, showing the value of fusing multiple constraints.

Automated defect recognition for semiconductor quality control. As discussed by Chen Y. et al., precise identification of surface defects in semiconductor wafers is critical for controlling manufacturing quality. They develop a multi-scale visual perception network architecture for automated wafer defect pattern recognition. By effectively integrating fine-grained texture cues across resolutions, their approach achieved state-of-the-art accuracy on a real-world industry dataset, demonstrating feasibility for quality inspection.

Self-supervised representation learning from multimodal data. For human action recognition, Yang et al. present a novel framework applying contrastive self-supervised learning on paired unlabeled data (skeleton sequences and inertial sensor signals). Without requiring negative samples, they show superior cross-dataset retrieval and zero-shot transfer performance compared to previous multimodal methods. This highlights the promise of self-supervised techniques to improve model generalization.

Elucidating audiovisual processing in the brain. Understanding the complex neural mechanisms underlying sensory integration remains a key challenge in neuroscience. Jiang et al. combine functional MRI and EEG to construct brain networks involved in audiovisual processing. Through their novel dynamic analysis approach, they revealed early visual-auditory integration occurring prior to attentional effects. These insights shed light on the nature of inter-sensory interactions within the brain.

AI for detecting overloaded trucks to improve road safety. Excessively overloaded trucks pose critical challenges regarding road damage and traffic safety. Sun et al. develop an AI system to detect truck overloading by recognizing truck models from images and matching against weight data. Achieving 85–100% accuracy on small real-world datasets shows feasibility for automated enforcement on highways to improve infrastructure maintenance and prevent hazardous accidents.

More human-like image captioning via reinforced decoding. Generating textual descriptions for images, known as image captioning, requires modeling both visual concepts and language semantics. Bai et al. introduce techniques including guided decoding connections, DenseNets, and reinforcement learning to enhance contextual modeling and feature extraction. Superior results across standard captioning metrics represent tangible progress toward human-level visual understanding.

Targeted smoke reduction to maintain surgical visualization. As discussed by Wang et al., smoke generated during endoscopic procedures can severely obscure surgical sight. They create an enhanced classifier to detect smoke-filled frames prior to selective image enhancement, maximizing efficiency. Achieving high accuracy and speed shows promise for integrated, real-time de-smoking systems to improve situational awareness.

Intelligent product recognition to enable smart vending. As Xu et al. explored, computer vision powered by deep learning can enable emerging autonomous retail models like smart vending machines to accurately recognize products for automatic checkout and inventory status tracking, reducing overhead costs. Their results demonstrate the feasibility of AI to deliver advanced functionality without constant human intervention.

Multi-Scale adaptive learning for robust driving scene parsing. Liu et al. address core challenges in semantic segmentation for autonomous vehicle perception including variations in scale, occlusions and diverse appearances. Their multi-scale adaptive network dynamically selects the most relevant features across levels to accurately parse complex driving environments. State-of-the-art performance on automotive datasets confirms robustness, advancing safety for self-driving systems.

Group-based sparse modeling for image restoration. Recovering high-quality images from incomplete or corrupted inputs remains an active computer vision research area. Ning et al. propose a multi-scale group sparse residual constraint model exploiting patch correlations to effectively eliminate noise and fill in missing regions. Experiments show marked improvements in restoration fidelity compared to existing methods, enabled by joint image priors.

In conclusion, recent advances in visual representation learning could unlock transformative capabilities in transportation, manufacturing, healthcare, and scientific imaging. While progress has been made in tackling real-world vision tasks, continued research into dynamic models, multimodal fusion, and incorporating domain-specific constraints will be instrumental in achieving human-like scene understanding.

## Author contributions

TL: Writing—original draft, Writing—review & editing. QZ: Writing—original draft. FN: Data curation, Writing—original draft. CS: Writing—review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is supported by the University Synergy Innovation Program of Anhui Province (No. GXXT-2022-037).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

frontiers | Frontiers in Neuroscience

TYPE Brief Research Report
PUBLISHED 19 May 2023
DOI 10.3389/fnins.2023.1191574



OPEN ACCESS

EDITED BY
Fudong Nian,
Hefei University, China

REVIEWED BY
Shengsheng Qian,
Chinese Academy of Sciences (CAS), China
Sisi You,
Nanjing University of Posts and Telecommunications, China
Fan Qi,
Tianjin University of Technology, China

*CORRESPONDENCE
Zhan-Li Sun
✉ zhlsun2006@126.com

RECEIVED 22 March 2023
ACCEPTED 17 April 2023
PUBLISHED 19 May 2023

CITATION
Chen X, Sun Z-L and Zhang Y (2023) 3D shape reconstruction with a multiple-constraint estimation approach.
*Front. Neurosci.* 17:1191574.
doi: 10.3389/fnins.2023.1191574



# 3D shape reconstruction with a multiple-constraint estimation approach

Xia Chen[1,2,3], Zhan-Li Sun[4,5]* and Ying Zhang[4]

[1]School of Information and Computer, Anhui Agricultural University, Hefei, China, [2]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institute of Physical Science and Information Technology, Anhui University, Hefei, China, [3]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei, China, [4]School of Electrical Engineering and Automation, Anhui University, Hefei, China, [5]Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei, China

In this study, a multiple-constraint estimation algorithm is presented to estimate the 3D shape of a 2D image sequence. Given the training data, a sparse representation model with an elastic net, i.e., $l_1$−norm and $l_2$−norm constraints, is devised to extract the shape bases. In the sparse model, the $l_1$−norm and $l_2$−norm constraints are enforced to regulate the sparsity and scale of coefficients, respectively. After obtaining the shape bases, a penalized least-square model is formulated to estimate 3D shape and motion, by considering the orthogonal constraint of the transformation matrix, and the similarity constraint between the 2D observations and the shape bases. Moreover, an Augmented Lagrange Multipliers (ALM) iterative algorithm is adopted to solve the optimization of the proposed approach. Experimental results on the well-known CMU image sequences demonstrate the effectiveness and feasibility of the proposed model.

KEYWORDS

**non-rigid structure from motion, elastic net, similarity constraint, Augmented Lagrange multipliers, 3D reconstruction**

## 1. Introduction

As an important component of computer vision, 3D shape reconstruction has been widely used in many applications (Li et al., 2016, 2018; Adamkiewicz et al., 2022; Chiang et al., 2022; Fombona-Pascual et al., 2022; Jang et al., 2022; Lu et al., 2022; Nian et al., 2022a,b; Wang et al., 2022; Wen et al., 2022). Among the various 3D shape reconstruction methods, non-rigid structure from motion (NRSFM) offers a technique to simultaneously recover the 3D structures and motions of an object, by using the 2D landmarks in a series of images (Graßhof and Brandt, 2022; Kumar and Van Gool, 2022; Song et al., 2022). Nevertheless, NRSFM is still an underconstrained and challenging issue because of lacking any prior knowledge of 3D structure deformation.

To alleviate the uncertainty, the various constraints are exploited constantly. Bregler et al. (2000), proposed a low-rank constraint-based approach to decompose the observation matrix into a motion factor and a shape basis. In order to reduce the number of the unknown variables proposed by Bregler et al. (2000), a point trajectory approach was presented by Akhter et al. (2010) by using the predefined bases of discrete cosine transform (DCT). However, the high-frequency deformation cannot be reconstructed well via this trajectory representation because of the low-rank constraint. Gotardo and Martinez (2011) modeled a smoothly deforming 3D shape as a single point moving along a smooth time trajectory within a linear shape space. In addition to the low-rank constraint, the higher frequency DCT was adopted to capture the high-frequency deformation.

For the low-rank constraint methods, it is difficult to determine the optimal number of shape bases or trajectory bases. To solve this problem, a Procrustean normal distribution (PND) model was presented by Lee et al. (2013) to separate the motion and deformation components strictly, without any additional constraints or prior knowledge. The experimental results demonstrate the performance of PND. Subsequently, the Procrustean Markov Process (PMP) algorithm was proposed by Lee et al. (2014), by combing in a first-order Markov model representing the smoothness between two adjacent frames with PND. Lee et al. (2016) reported a consensus of non-rigid reconstruction (CNR) approach to estimate 3D shapes based on local patches. However, the reconstruction performance of these methods may degrade significantly when the number of images becomes small, especially for a single image.

Referring to the active shape model (Cootes et al., 1995), a limb length constraint-based approach was presented by Wang et al. (2014) to estimate the 3D shape of an object from a single 2D image, by solving a $l_1-$norm minimization problem. Zhou et al. (2013) proposed a sparse representation-based convex relaxation approach (CRA) to guarantee global optimality. The shape bases were extracted from a given training data by using a sparse representation model. The corresponding coefficients were obtained by adopting a convex relaxation assumption. A prominent advantage of CRA is that the algorithm can deal with a single image.

To further enhance the performance of the CRA algorithm, a multiple-constraint-based estimation approach is proposed to estimate the 3D shape of a 2D image sequence. Inspired by Zhang and Xing (2017), a dictionary learning model with $l_1-$norm and $l_2-$norm, i.e., elastic net, is constructed to extract more effective shape bases from a given training set. Referring to (Cheng et al., 2015), a penalized least-square model is constructed to estimate 3D shape and motion, by considering the orthogonal constraint of the transformation matrix and the similarity constraint between the 2D observations and the shape bases. In addition, an augmented Lagrange multipliers (ALM) iterative algorithm is developed to optimize the reconstruction model. The effectiveness and feasibility of the proposed algorithm are verified on the well-known CMU image sequences.

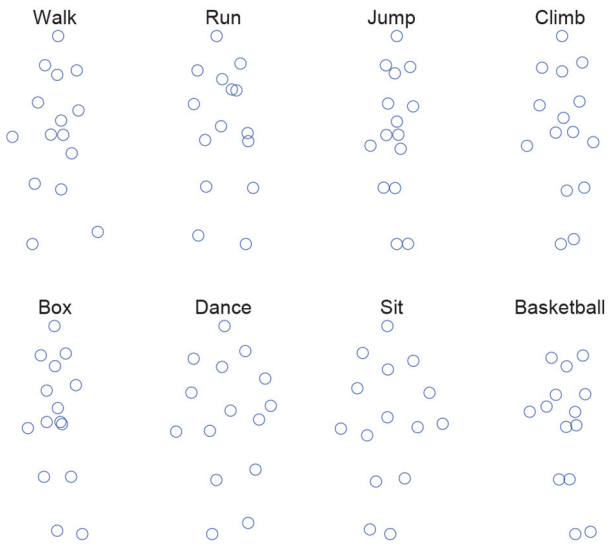



FIGURE 1
One frame of those eight categories.

The rest of this article is organized as follows. A detailed description of the designed MCM-RR approach is introduced in Section 2. In Section 3, we report the experimental results. Finally, the article is concluded in Section 4.

## 2. Methods

According to the shape-space model by Zhou et al. (2013), the unknown 3D shape $\mathbf{S} \in \mathbb{R}^{3\times p}$ is constructed as a linear combination of a few shape bases $\mathbf{B}_i \in \mathbb{R}^{3\times p}$, i.e.,

$$\mathbf{S} = \sum_{i=1}^{K} c_i \mathbf{R}_i \mathbf{B}_i, \tag{1}$$

where $p$ and $K$ are the numbers of feature points and shape bases, respectively. The parameter $c_i$ and $\mathbf{R}_i \in \mathbb{R}^{3\times 3}$ denote the coefficient and rotation matrix, respectively. In terms of the weak-perspective projection model, the corresponding 2D observations are modeled as a matrix $\mathbf{W} \in \mathbb{R}^{2\times p}$,

$$\mathbf{W} = \sum_{i=1}^{K} \mathbf{M}_i \mathbf{B}_i. \tag{2}$$

The matrix $\mathbf{M}_i \in \mathbb{R}^{2\times 3}$ can be represented as

$$\mathbf{M}_i = c_i \tilde{R}_i, \tag{3}$$

where $\tilde{R}_i \in \mathbb{R}^{2\times 3}$ is the first two rows of $\mathbf{R}_i$. Combining the orthogonal constraint, the matrix $\mathbf{M}_i$ satisfies

$$\mathbf{M}_i \mathbf{M}_i^T = c_i^2 \mathbf{I}_2, \tag{4}$$

where $\mathbf{I}_2 \in \mathbb{R}^{2\times 2}$ is an identity matrix. The 3D shape, i.e., $z-$coordinates, and the motion parameters $c_i$ and $\mathbf{R}_i$, are estimated by utilizing the observations $\mathbf{W}$, i.e., the $(x, y)$ coordinates of feature points.

In the proposed method, the shape bases $\mathbf{B} \in \mathbb{R}^{3K\times p}$ are extracted via a sparse model with the elastic net constraint. The $\mathbf{B}$ is the stacking of $\mathbf{B}_i (i = 1, ..., K)$. The matrix $\mathbf{M}$ are solved by a penalized least-square model. Given $\mathbf{M}$, the parameters $c_i$ and $\mathbf{R}_i$ are derived via refinement decompose (Zhou et al., 2013). After obtaining $c_i$, $\mathbf{R}_i$ and $\mathbf{B}_i$, the unknown 3D shape can be computed via (1). The pseudocode of the proposed algorithm is summarized in 1. The pseudocode of the proposed algorithm is summarized in Algorithm 1.

### 2.1. Extraction of shape bases via a sparse model with elastic net constraint

For a given 3D training set $\mathbf{A} \in \mathbb{R}^{3p\times F}$, i.e., the $(x, y, z)$ coordinates of feature points of training images, the shape bases

TABLE 1 Mean and standard deviation ($\mu \pm \sigma$) of the 3D reconstruction errors $\xi$ of eight motion categories for five methods.

| Sequence | PMP | CNR | PND2 | CRA | MCM-RR |
|---|---|---|---|---|---|
| Walk | 97.06 ± 17.35 | 78.28 ± 15.70 | 104.20 ± 26.13 | 38.98 ± 19.64 | 35.37 ± 18.49 |
| Run | 119.37 ± 31.37 | 65.92 ± 23.69 | 124.54 ± 28.82 | 55.69 ± 18.13 | 52.64 ± 17.05 |
| Jump | 102.22 ± 30.74 | 61.66 ± 40.35 | 84.64 ± 41.80 | 57.08 ± 41.56 | 44.56 ± 27.30 |
| Climb | 119.08 ± 39.39 | 69.36 ± 30.21 | 87.72 ± 56.04 | 58.87 ± 24.73 | 50.25 ± 25.88 |
| Box | 252.61 ± 41.28 | 82.83 ± 33.65 | 146.91 ± 45.17 | 72.90 ± 30.64 | 65.28 ± 26.82 |
| Dance | 118.24 ± 35.34 | 105.73 ± 38.81 | 118.52 ± 62.07 | 102.36 ± 44.93 | 83.59 ± 34.88 |
| Sit | 96.31 ± 32.77 | 69.58 ± 42.18 | 73.20 ± 32.47 | 75.68 ± 36.29 | 62.72 ± 26.79 |
| Basketball | 121.26 ± 44.83 | 67.63 ± 38.97 | 105.38 ± 72.17 | 63.66 ± 27.92 | 57.57 ± 22.96 |

TABLE 2 Corresponding 3D reconstruction error decreasing percentage $\xi_p(\%)$ of MCM-RR compared to CRA for eight motion categories.

| Sequence | $\xi_p$ |
|---|---|
| Walk | 9.26 |
| Run | 5.48 |
| Jump | 21.93 |
| Climb | 14.64 |
| Box | 10.43 |
| Dance | 18.34 |
| Sit | 17.12 |
| Basketball | 9.57 |

$\mathbf{N} \in \mathbb{R}^{3p\times K}$ and the coefficient matrix $\mathbf{X} \in \mathbb{R}^{K\times F}$ can be obtained from the following sparse model:

$$\begin{aligned} &\min_{\mathbf{N}_1,\cdots,\mathbf{N}_K} \frac{1}{2}\left\|\mathbf{A}-\mathbf{N}\mathbf{X}\right\|_F^2 + \lambda\left(\tau\left\|\mathbf{X}\right\|_1 + (1-\tau)\left\|\mathbf{X}\right\|_2^2\right) \\ &\text{s.t.}\ \left\|\mathbf{N}_i\right\|_F \le 1, \qquad X_{ij} \ge 0, \forall i \in [1,K], j \in [1,F], \end{aligned} \tag{5}$$

where $F$ and $\tau$ are the number of frames and a weight coefficient, respectively. The $\mathbf{N}_i \in \mathbb{R}^{3p\times 1}$ is the $i$-th column of $\mathbf{N}$. The linear combination of $l_1-$norm and $l_2-$norm, called elastic net constraint, are enforced to constraint the sparsity of coefficients $\mathbf{X}$ as well as scale. The parameter $\lambda$ is a trade-off parameter between the reconstruction error and the elastic net constraint.

For (5), we first compute the partial differentials of $\mathbf{X}$ and $\mathbf{N}$, i.e.,

$$\partial\mathbf{X} = (\mathbf{N}^t)^T\left(\mathbf{A}-\mathbf{N}^t\mathbf{X}\right) + \lambda\left(\tau\mathbf{I}_{KF} + 2(1-\tau)\mathbf{X}\right), \tag{6}$$

$$\partial\mathbf{N} = \left(\mathbf{A}-\mathbf{N}\left(\mathbf{X}^{t+1}\right)^T\right)\left(\mathbf{X}^{t+1}\right)^T, \tag{7}$$

where $\mathbf{I}_{KF}$ is a $K \times F$ identity matrix. Thereafter, $\mathbf{X}$ and $\mathbf{N}$ can be updated alternately as

$$\mathbf{X}^{t+1} = \mathbf{X}^t - \phi_1\partial\mathbf{X}, \tag{8}$$

$$\mathbf{N}^{t+1} = \mathbf{N}^t - \phi_2\partial\mathbf{N}, \tag{9}$$

```
Compute the shape bases B via the elastic net based sparse model (5).
Initialize α, β, γ.
Initialize M^0, Z^0, Y^0, μ^0, t = 0.
while t <= 1000 do
  Compute the optimized M^{t+1} according to (15) by fixing Z^t, Y^t, and μ^t,
  Update Z^{t+1} via (17) by fixing M^{t+1}, Y^t, and μ^t,
  Update Y^{t+1} via (18) by fixing M^{t+1}, Z^{t+1},
  if δ1 < ε & δ2 < ε then
    break,
  else
    if δ1 > 10δ2 then
      μ^{t+1} = 2μ^t,
    else {δ2 > 10δ1}
      μ^{t+1} = μ^t/2.
    end if
  end if
  Update t ← t + 1.
end while
if refinement reconstruction then
  Compute R and c according to (22) via the alternating minimization (Zhou et al., 2013).
end if
Estimate S by using (1)
```

Algorithm 1. Pseudocode of the MCM-RR algorithm.

where $\phi_1$ and $\phi_2$ are the step size of $\partial\mathbf{X}$ and $\partial\mathbf{N}$, respectively. After convergence, the shape bases $\mathbf{B}$ can be obtained by a re-arrangement of $\mathbf{N}$.

## 2.2. 3D shape estimation via a penalized least-square model with similarity constraint

In terms of (2), the proposed penalized least-square model, including a relaxed orthogonality constraint (Zhou et al., 2013) and

a similarity constraint (Cheng et al., 2015) can be formulated as

$$\begin{aligned} \min_{\widetilde{\mathbf{M}},\mathbf{Z}} \frac{1}{2}\left\|\mathbf{W}-\mathbf{Z}\widetilde{\mathbf{B}}\right\|_F^2+\alpha\sum_{i=1}^{K}\left\|\mathbf{M}_i\right\|_2+\frac{\beta}{2}\left\|\mathbf{ZD}\right\|_2^2 \\ \text{s.t. } \widetilde{\mathbf{M}}=\mathbf{Z}, \end{aligned} \tag{10}$$

where $\mathbf{Z}\in\mathbb{R}^{2\times 3K}$ is an auxiliary variable and $\widetilde{\mathbf{M}}=[\mathbf{M}_1,\cdots,\mathbf{M}_K]$, $\widetilde{\mathbf{B}}=\left[\mathbf{B}_1^T,\cdots,\mathbf{B}_K^T\right]^T$. The parameters $\alpha$ and $\beta$ are used to weight the two regularization terms. The diagonal matrix $\mathbf{D}\in\mathbb{R}^{3K\times 3K}$ is represented as

$$\mathbf{D}=\left(\bar{\mathbf{D}}\otimes\mathbf{I}_3\right). \tag{11}$$

For the diagonal similarity matrix $\bar{\mathbf{D}}\in\mathbb{R}^{K\times K}$, the diagonal element $d_i$ is computed as

$$d_i=\exp\left(\frac{\left\|\mathbf{W}-\Pi\mathbf{B}_i\right\|}{2\gamma^2}\right), \tag{12}$$

where $\Pi=[1,0,0;0,1,0]$, $\gamma^2$ is the parameter of an exponential function.

With the ALM iterative algorithm, the penalized least-square model (10) can be reformulated as

$$\begin{aligned} L=&\frac{1}{2}\left\|\mathbf{W}-\mathbf{Z}\widetilde{\mathbf{B}}\right\|_F^2+\alpha\sum_{i=1}^{K}\left\|\mathbf{M}_i\right\|_2+\langle\mathbf{Y},\widetilde{\mathbf{M}}-\mathbf{Z}\rangle \\ &+\frac{\beta}{2}\left\|\mathbf{ZD}\right\|_2^2+\frac{\mu}{2}\left\|\widetilde{\mathbf{M}}-\mathbf{Z}\right\|_F^2, \end{aligned} \tag{13}$$

where $\mathbf{Y}$ and $\mu$ are a dual variable and a weight of penalty term, respectively. In (13), there are four unknown variables $\widetilde{\mathbf{M}}$, $\mathbf{Z}$, $\mathbf{Y}$, and $\mu$. The solutions can be solved by the alternating direction method of multipliers (ADMM).

First, the optimal $\widetilde{\mathbf{M}}$ at the $(t+1)^{th}$ iteration can be formulated as

$$\widetilde{\mathbf{M}}^{t+1}=\arg\min_{\widetilde{\mathbf{M}}}\sum_{i=1}^{K}\frac{1}{2}\left\|\mathbf{M}_i-\mathbf{P}_i^t\right\|_F^2+\frac{\alpha}{\mu}\left\|\mathbf{M}_i\right\|_2, \tag{14}$$

where $\mathbf{P}_i^t$ is the $i^{th}$ column-triple of $\mathbf{Z}^t-\frac{1}{\mu}\mathbf{Y}^t$. According to the proximal problem (Zhou et al., 2013), $\mathbf{M}_i^{t+1}$ can be computed as

$$\mathbf{M}_i^{t+1}=U\text{diag}\left(\Sigma-\frac{\alpha}{\mu}\mathcal{P}_{l_1}\left(\frac{\Sigma\mu}{\alpha}\right)\right)V^T, i\in[1,K], \tag{15}$$

where $U\Sigma V^T=svd(\mathbf{P}_i^t)$. The operation $\mathcal{P}_{l_1}(\cdot)$ denotes the projection of a vector to the unit $l_1-$norm ball (Zhou et al., 2013).

Similarity, the optimal $\mathbf{Z}$ at the $(t+1)^{th}$ iteration can be formulated as

$$\begin{aligned} \mathbf{Z}^{t+1}=&\arg\min_{\mathbf{Z}}\frac{1}{2}\left\|\mathbf{W}-\mathbf{Z}\widetilde{\mathbf{B}}\right\|_F^2+\langle\mathbf{Y}^t,\widetilde{\mathbf{M}}^{t+1}-\mathbf{Z}\rangle \\ &+\frac{\beta}{2}\left\|\mathbf{ZD}\right\|_2^2+\frac{\mu}{2}\left\|\widetilde{\mathbf{M}}^{t+1}-\mathbf{Z}\right\|_F^2. \end{aligned} \tag{16}$$

We compute the one-order partial derivative of (16) with respect to $\mathbf{Z}$ and set it as zero. Thereafter, $\mathbf{Z}^{t+1}$ can be given by

$$\mathbf{Z}^{t+1}=\left(\mathbf{W}\widetilde{\mathbf{B}}^T+\mu\widetilde{\mathbf{M}}^{t+1}+\mathbf{Y}^t\right)\left(\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T+\mu\mathbf{I}+\beta\mathbf{D}\mathbf{D}^T\right)^{-1}. \tag{17}$$

Afterward, the optimal $\mathbf{Y}$ at the $(t+1)^{th}$ iteration can be computed as

$$\mathbf{Y}^{t+1}=\mathbf{Y}^t+\mu\left(\widetilde{\mathbf{M}}^{t+1}-\mathbf{Z}^{t+1}\right). \tag{18}$$

Given a weight $\tau$, the coefficient $\mu$ at the $(t+1)^{th}$ iteration can be given by

$$\mu^{t+1}=\begin{cases}2\mu^t, & if \quad \delta_1>\tau\delta_2,\\ \mu^t/2, & if \quad \delta_2>\tau\delta_1,\end{cases} \tag{19}$$

where

$$\delta_1=\frac{\left\|\widetilde{\mathbf{M}}^{t+1}-\mathbf{Z}^{t+1}\right\|_F}{\left\|\mathbf{Z}^t\right\|_F},\delta_2=\frac{\left\|\mathbf{Z}^{t+1}-\mathbf{Z}^t\right\|_F}{\left\|\mathbf{Z}^t\right\|_F}. \tag{20}$$

The iterations are repeated until

$$\delta_1<\varepsilon \quad \& \quad \delta_2<\varepsilon, \tag{21}$$

where $\varepsilon$ is a small threshold value. After obtaining $\mathbf{M}_i$, the unknown 3D shape can be reconstructed by refinement reconstruction (Zhou et al., 2013).

In the refinement reconstruction, we assume that the rotation matrices of each shape base are equal, denoted as $\bar{\mathbf{R}}$. Thereafter, $c_i$ and $\bar{\mathbf{R}}$ can be estimated by the following rotation synchronization model

$$\begin{aligned} \min_{\mathbf{c},\bar{\mathbf{R}}} \quad & \sum_{i=1}^{k}\left\|\mathbf{M}_i-c_i\bar{\mathbf{R}}\right\|_F^2, \\ \text{s.t.} \quad & \bar{\mathbf{R}}\bar{\mathbf{R}}^T=\mathbf{I}_2, \end{aligned} \tag{22}$$

which can be solved via the alternating minimization (Zhou et al., 2013). Finally, the 3D shape $\mathbf{S}$ can be estimated after $\mathbf{M}_i$ is obtained.

# 3. Experimental results

## 3.1. Experimental comparison of different algorithms

The performance evaluation of the proposed 3D shape reconstruction model (denoted as MCM-RR) is carried out on eight motion categories (walk, run, jump, climb, box, dance, sit, and basketball) from the CMU motion capture dataset (Zhou et al., 2013). Figure 1 shows one frame of those eight categories.

In the experiments, the performance of several state-of-the-art 3D shape estimation methods are used to compare with the presented approach, including PND2 (Lee et al., 2013), CNR (Lee et al., 2016), PMP (Lee et al., 2014), and CRA (Zhou et al., 2013).

Mean error $\xi$ of 3D shapes is calculated as the performance indicator to measure the estimation results:

$$\xi=\frac{1}{F}\sum_{t=1}^{F}\left\|\tilde{\mathbf{S}}_t-\mathbf{S}_t\right\|_F^2, \tag{23}$$

where $\tilde{\mathbf{S}}_t\in\mathbb{R}^{3\times p}$ and $\mathbf{S}_t\in\mathbb{R}^{3\times p}$ are the reconstructed 3D structure and real 3D structure of $t^{th}$ frame, respectively.

Table 1 displays the mean and standard deviation ($\mu\pm\sigma$) of reconstruction errors $\xi$ of eight motion categories for the five

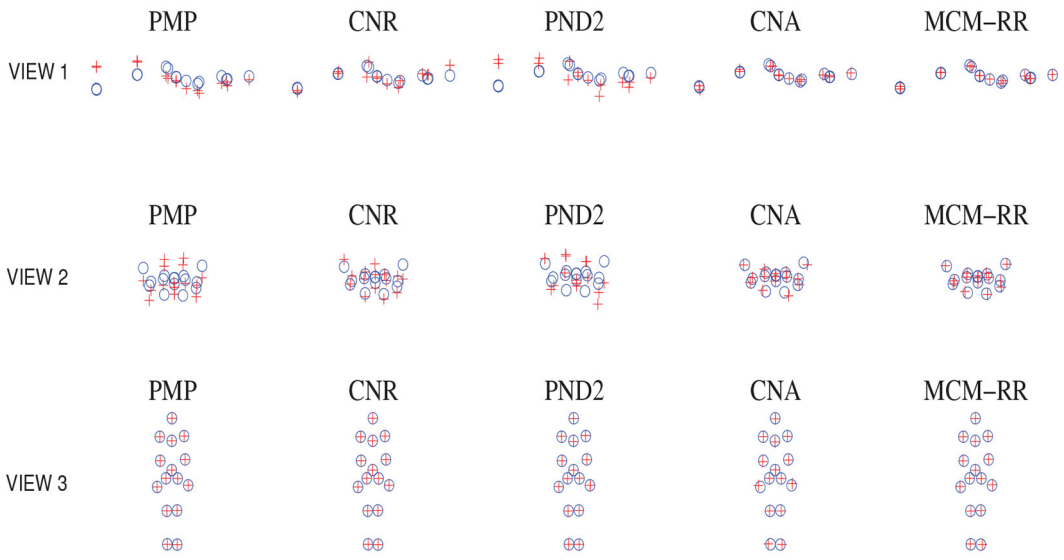



FIGURE 2
Comparisons of estimated shapes for single frame of *Jump* between MCM-RR and other methods from three different viewpoints. The symbol "○" denotes the observed real points, whereas "+" denotes reconstructed points.

**TABLE 3** Mean and standard deviation ($\mu \pm \sigma$) of the 3D reconstruction errors $\xi$ of eight motion categories for four methods.

| Sequence | CRA | CRA-EN | CRA-SC | MCM-RR |
|---|---|---|---|---|
| Walk | 38.98 ± 19.64 | 36.56 ± 19.18 | 38.64 ± 19.03 | 35.37 ± 18.49 |
| Run | 55.69 ± 18.13 | 52.60 ± 16.70 | 56.06 ± 18.03 | 52.64 ± 17.05 |
| Jump | 57.08 ± 41.56 | 46.61 ± 33.79 | 56.42 ± 39.52 | 44.56 ± 27.30 |
| Climb | 58.87 ± 24.73 | 49.99 ± 25.53 | 58.99 ± 24.88 | 50.25 ± 25.88 |
| Box | 72.90 ± 30.64 | 65.32 ± 27.64 | 73.02 ± 30.10 | 65.28 ± 26.82 |
| Dance | 102.36 ± 44.93 | 85.23 ± 35.63 | 101.49 ± 44.01 | 83.59 ± 34.88 |
| Sit | 75.68 ± 36.29 | 63.12 ± 26.79 | 74.92 ± 34.80 | 62.72 ± 26.79 |
| Basketball | 63.66 ± 27.92 | 57.81 ± 22.58 | 63.28 ± 28.29 | 57.57 ± 22.96 |

methods, respectively. The best results are highlighted in red, whereas the second best is in blue.

Table 1 shows the estimation errors of the last two methods are clearly less than that of the first triple algorithms. Among eight categories, the mean reconstruction errors of MCM-RR are the lowest compared to CRA. Moreover, the standard deviations of MCM-RR are less than that of CRA among most categories. Therefore, compared to CRA, both accuracy and robustness are effectively improved for the proposed method.

Compared to CRA, the 3D reconstruction error decreased the percentage $\xi_p(\%)$ of MCM-RR can be computed as

$$\xi_p = \frac{\xi_{\text{CRA}} - \xi_{\text{MCM-RR}}}{\xi_{\text{CRA}}} \times 100\%. \tag{24}$$

From Table 2, we can see that the mean reconstruction errors of MCM-RR decreased about 5.48% $\backsim$ 21.93% compared to CRA. Thus, MCM-RR has a better 3D reconstruction performance than CRA for the eight motion categories.

Take one frame of *Jump* as an example. Figure 2 displays a comparison of reconstructed shapes between MCM-RR and the other methods from three different viewpoints. From Figure 2, we can see that compared to other methods, most estimated shapes of MCM-RR are closer to real points than that of the other methods.

## 3.2. Ablation experiment

In order to verify the feasibility of the proposed two strategies, the elastic net (denoted as CRA-EN) and similarity constraint (denoted as CRA-SC) are separately applied to the original algorithm CRA. Table 3 displays the mean and standard deviation ($\mu \pm \sigma$) of 3D reconstruction errors $\xi$ of eight motion categories for the four methods, respectively. Compared to CRA, both the elastic net and similarity constraint can decrease the 3D reconstruction errors. Therefore, the 3D reconstruction performance can be effectively improved once the two methods are simultaneously designed into CRA.

## 4. Conclusion

In this study, a multiple-constraint algorithm is devised to estimate the 3D shape of a 2D image sequence. Experimental results on the well-known CMU datasets demonstrated that the proposed methods have higher accuracies and more robustness. Compared with CRA, the 3D reconstruction error is decreased by at least 5.48%.

## Data availability statement

The datasets used in this article is from a public datesets, and it can be found in the CMU Graphics Lab Motion Capture Database.

## Author contributions

XC proposed the initial research idea, conducted the experiments, and wrote the manuscript. Z-LS supervised the work and advised the entire research process. YZ collected the dataset, analyzed the formal, and revised the manuscript. All authors reviewed and approved the final manuscript.

## Funding

This work was supported by the National Natural Science Foundation of China (No. 61972002), the University Natural Science Research Project of Anhui Province (No. KJ2021A0180), Natural Science Foundation of Anhui Agricultural University (No. K2148001), Research Talents Stable Project of Anhui Agricultural University (No. rc482004), Key Laboratory of Intelligent Computing & Signal Processing, Ministry of Education (Anhui University) (No. 2020A002), Anhui Provincial Key Laboratory of Multimodal Cognitive Computation (Anhui University) (No. MMC202004), and the Anhui Provincial Natural Science Foundation (No. 2108085MC96).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., et al. (2022). Vision-only robot navigation in a neural radiance world. *IEEE Robot. Automat. Lett.* 7, 4606–4613. doi: 10.1109/LRA.2022.3150497

Akhter, I., Sheikh, Y., Khan, S., and Kanade, T. (2010). Trajectory space: a dual representation for nonrigid structure from motion. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1442–1456. doi: 10.1109/TPAMI.2010.201

Bregler, C., Hertzmann, A., and Biermann, H. (2000). "Recovering non-rigid 3d shape from image streams," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition* (Hilton Head, SC: IEEE), 690–696.

Cheng, J., Yin, F., Wong, D. W. K., Tao, D., and Liu, J. (2015). Sparse dissimilarity-constrained coding for glaucoma screening. *IEEE Trans. Biomed. Eng.* 62, 1395–1403. doi: 10.1109/TBME.2015.2389234

Chiang, F.-K., Shang, X., and Qiao, L. (2022). Augmented reality in vocational training: a systematic review of research and applications. *Comput. Hum. Behav.* 129, 107125. doi: 10.1016/j.chb.2021.107125

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Comput. Vis. Image Understand.* 61, 38–59.

Fombona-Pascual, A., Fombona, J., and Vicente, R. (2022). Augmented reality, a review of a way to represent and manipulate 3d chemical structures. *J. Chem. Inform. Model.* 62, 1863–1872. doi: 10.1021/acs.jcim.1c01255

Gotardo, P. F., and Martinez, A. M. (2011). Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 2051–2065. doi: 10.1109/TPAMI.2011.50

Graßhof, S., and Brandt, S. S. (2022). "Tensor-based non-rigid structure from motion," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition* (Waikoloa, HI: IEEE), 3011–3020.

Jang, H., Sedaghat, S., Athertya, J. S., Moazamian, D., Carl, M., Ma, Y., et al. (2022). Feasibility of ultrashort echo time quantitative susceptibility mapping with a 3d cones trajectory in the human brain. *Front. Neurosci.* 16, 1033801. doi: 10.3389/fnins.2022.1033801

Kumar, S., and Van Gool, L. (2022). "Organic priors in non-rigid structure from motion," in *Proceedings of the European Conference on Computer Vision* (Springer), 71–88.

Lee, M., Cho, J., Choi, C.-H., and Oh, S. (2013). "Procrustean normal distribution for non-rigid structure from motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR: IEEE), 1280–1287.

Lee, M., Cho, J., and Oh, S. (2016). "Consensus of non-rigid reconstructions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 4670–4678.

Lee, M., Choi, C.-H., and Oh, S. (2014). "A procrustean markov process for non-rigid structure recovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 1550–1557.

Li, T., Cheng, B., Ni, B., Liu, G., and Yan, S. (2016). Multitask low-rank affinity graph for image segmentation and image annotation. *ACM Trans. Intell. Syst. Technol.* 7, 1–18. doi: 10.1145/2856058

Li, T., Wang, Y., Hong, R., Wang, M., and Wu, X. (2018). PDisVPL: probabilistic discriminative visual part learning for image classification. *IEEE MultiMedia* 25, 34–45. doi: 10.1109/MMUL.2018.2873499

Lu, W., Li, Z., Li, Y., Li, J., Chen, Z., Feng, Y., et al. (2022). A deep learning model for three-dimensional nystagmus detection and its preliminary application. *Front. Neurosci.* 16, 930028. doi: 10.3389/fnins.2022.930028

Nian, F., Li, T., Bao, B.-K., and Xu, C. (2022a). Relative coordinates constraint for face alignment. *Neurocomputing* 395, 119–127. doi: 10.1016/j.neucom.2017.12.071

Nian, F., Sun, J., Jiang, D., Zhang, J., Li, T., and Lu, W. (2022b). Predicting dose-volume histogram of organ-at-risk using spatial geometric-encoding network for esophageal treatment planning. *J. Ambient Intell. Smart Environ.* 14, 25–37. doi: 10.3233/AIS-210084

Song, J., Patel, M., Jasour, A., and Ghaffari, M. (2022). A closed-form uncertainty propagation in non-rigid structure from motion. *IEEE Robot. Automat. Lett.* 7, 6479–6486. doi: 10.1109/LRA.2022.3173733

Wang, C., Wang, Y., Lin, Z., Yuille, A. L., and Gao, W. (2014). "Robust estimation of 3d human poses from a single image," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 2361–2368.

Wang, T., Chen, B., Zhang, Z., Li, H., and Zhang, M. (2022). Applications of machine vision in agricultural robot navigation: a review. *Comput. Electron. Agric.* 198, 107085. doi: 10.1016/j.compag.2022.107085

Wen, X., Zhou, J., Liu, Y.-S., Su, H., Dong, Z., and Han, Z. (2022). "3d shape reconstruction from 2d images with disentangled attribute flow," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 3803–3813.

Zhang, S., and Xing, W. (2017). "Object tracking with adaptive elastic net regression," in *Proceedings of the IEEE International Conference on Image Processing* (Honolulu, HI: IEEE), 2597–2601.

Zhou, X., Zhu, M., Leonardos, S., and Daniilidis, K. (2013). Sparse representation for 3d shape estimation: a convex relaxation approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1648–1661. doi: 10.1109/TPAMI.2016.2605097

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 24 May 2023
DOI 10.3389/fnins.2023.1188434



OPEN ACCESS

EDITED BY
Caifeng Shan,
Shandong University of Science and Technology, China

REVIEWED BY
Alwin Poulose,
Indian Institute of Science Education and Research, Thiruvananthapuram, India
Huaming Chen,
The University of Sydney, Australia

*CORRESPONDENCE
Zhenyu Liu
✉ liuzhenyu@lzu.edu.cn
Xiping Hu
✉ huxp@lzu.edu.cn
Bin Hu
✉ bh@lzu.edu.cn

RECEIVED 17 March 2023
ACCEPTED 02 May 2023
PUBLISHED 24 May 2023

CITATION
Li Y, Liu Z, Zhou L, Yuan X, Shangguan Z, Hu X and Hu B (2023) A facial depression recognition method based on hybrid multi-head cross attention network.
*Front. Neurosci.* 17:1188434.
doi: 10.3389/fnins.2023.1188434



# A facial depression recognition method based on hybrid multi-head cross attention network

Yutong Li, Zhenyu Liu*, Li Zhou, Xiaoyan Yuan, Zixuan Shangguan, Xiping Hu* and Bin Hu*

Gansu Provincial Key Laboratory of Wearable Computing, Lanzhou University, Lanzhou, China

**Introduction:** Deep-learn methods based on convolutional neural networks (CNNs) have demonstrated impressive performance in depression analysis. Nevertheless, some critical challenges need to be resolved in these methods: (1) It is still difficult for CNNs to learn long-range inductive biases in the low-level feature extraction of different facial regions because of the spatial locality. (2) It is difficult for a model with only a single attention head to concentrate on various parts of the face simultaneously, leading to less sensitivity to other important facial regions associated with depression. In the case of facial depression recognition, many of the clues come from a few areas of the face simultaneously, e.g., the mouth and eyes.

**Methods:** To address these issues, we present an end-to-end integrated framework called Hybrid Multi-head Cross Attention Network (HMHN), which includes two stages. The first stage consists of the Grid-Wise Attention block (GWA) and Deep Feature Fusion block (DFF) for the low-level visual depression feature learning. In the second stage, we obtain the global representation by encoding high-order interactions among local features with Multi-head Cross Attention block (MAB) and Attention Fusion block (AFB).

**Results:** We experimented on AVEC2013 and AVEC2014 depression datasets. The results of AVEC 2013 (RMSE = 7.38, MAE = 6.05) and AVEC 2014 (RMSE = 7.60, MAE = 6.01) demonstrated the efficacy of our method and outperformed most of the state-of-the-art video-based depression recognition approaches.

**Discussion:** We proposed a deep learning hybrid model for depression recognition by capturing the higher-order interactions between the depression features of multiple facial regions, which can effectively reduce the error in depression recognition and gives great potential for clinical experiments.

KEYWORDS

**facial depression recognition, convolutional neural networks, attention mechanism, automatic depression estimation, end-to-end network**

## 1. Introduction

Major depressive disorder (MDD), also called depression, is one of the most common mental and mood disorders. It presents itself through depressed mood, pessimism, loss of attention and memory, self-denial, poor appetite, and decreased activity, among other symptoms. In addition, it can severely impact a person's thoughts, behaviors, work-life, and eating habits (Belmaker and Agam, 2008). With the increasing pressure of life, many people are suffering from depression. The World Health Organization (WHO) released data in 2007 stating that 350 million people worldwide suffered from depression. Moreover, in 2030, depression may overtake cardiovascular disease as the number one cause of disability,

TABLE 1 The relation between the BDI-II cut-off scores and the depression severity level.

| BDI-II score | Severity level |
|---|---|
| 0–13 | None or minimal |
| 14–19 | Mild |
| 20–28 | Moderate |
| 29–63 | Severe |

which means that depression has become a severe social health problem (World Health Organization, 2017). Unfortunately, there are no impactful clinical patterns for the diagnosis of depression due to personal and social development and other factors, which makes the diagnosis of depression complicated and subjective (Maj et al., 2020). Meanwhile, there are few professional psychiatrists in some developing countries, and the insufficient ratio of doctors to patients has become a major problem faced by mental health workers as well. Therefore, it is necessary to find objective parameter indicators to assist doctors in improving the current medical situation.

Studies have shown that depression alters various non-verbal behaviors (Ellgring, 2007), including psychomotor delays, insensitivity to emotional stimuli, and diminished positive and negative emotional responses, all of which can transfer information about depression levels (Cohn et al., 2009; Michalak et al., 2009; Canales et al., 2017). Especially, the face presents most of the people's non-verbal information, which leads to that as a characteristic indicator with high information content in the diagnosis of depression. Clinically, patients with depression often have reduced facial expression richness, drooping eyes, frowning, drooping mouth corners, reduced smile, and easy crying (Pampouchidou et al., 2020). Therefore, various researchers from the affective computing field have attempted to use facial changes as a biomarker to analyze the individual depression level and measured by the Beck Depression Inventory-II (BDI-II) score (McPherson and Martin, 2010), as presented in Table 1.

Estimating the level of depression from facial images usually includes the following steps: (1) feature extraction and (2) regression (or classification). Among them, the task of feature extraction involves designing an effective depression representation that plays a significant role in facial depression recognition. At present, there are two main methods of feature extraction as follows: hand-crafted (Valstar et al., 2013, 2014; Wen et al., 2015) and deep-learned (Jan et al., 2017; Zhu et al., 2017; Al Jazaery and Guo, 2018; Zhou et al., 2020; Guo et al., 2021). For hand-crafted features, Local Phase Quantization (LPQ) and Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) are adopted as visual features for predicting the scale of depression (Valstar et al., 2013, 2014). However, these features are difficult to obtain accurate and subtle facial information (Song et al., 2018). Meanwhile, hand-crafted methods often involve a complex set of image processing steps, leading to relying heavily on expert knowledge (Ojala et al., 2002; Laptev et al., 2008; Meng and Pears, 2009). On the contrary, deep learning features do not rely on expert knowledge and complex manual design, which can capture and reveal high-level semantic features of faces. Zhou et al. (2020) propose a deep regression network to learn a depressive feature representation visually interpretably, and the result shows that the area near the eyes plays a crucial role in recognizing depression. Al Jazaery and Guo (2018) have automatically learned spatiotemporal features of facial regions at two different scales by using three-dimensional convolutional neural network (3D-CNN) and recurrent neural network (RNN), which can model the local and global spatiotemporal information from continuous facial expressions to predict depression levels.

However, most of the above methods do not further explore the local details. One unique aspect of facial depression recognition lies in the delicate contention between capturing the subtle local variations and obtaining a unified, holistic representation. Some recent studies focus on attention mechanisms to balance the local details and unified, holistic representation. For instance, He et al. (2021a) propose an integrated architecture called Deep Local-Global Attention Convolutional Neural Network (DLGA-CNN), which utilizes Convolutional Neural Network (CNN) with attention mechanism and weighted spatial pyramid pooling (WSPP) to model a local-global facial feature. Liu et al. (2023) design a global region-based network with part-and-relation attention, which learns the relation between part and global features. Niu et al. (2022) introduce an architecture using CNN and attention mechanism for automatic depression recognition by facial changes, and the performance surpasses most facial depression recognition methods. These methods focusing on attention mechanisms have achieved promising results by paying attention to facial details. Nevertheless, as shown in Figure 1, it is difficult for a model with only a single attention head to concentrate on various parts of the face simultaneously and just concentrate on one coarser image region, missing other important facial locations. Existing research results show that the differences in facial changes between depressed patients and healthy people are simultaneously manifested in multiple parts of the face (Schwartz et al., 1976; Scherer et al., 2013), such as eyebrows, eyes, cheeks, and mouth. Therefore, to mitigate the problems mentioned above, we propose a Hybrid Multi-Head Cross-Attention Network (HMHN), which implements multiple attention mechanisms to capture the high-order interactions between the local features of multiple facial regions by instantiating multiple attention heads.

More specifically, as shown in Figure 2, the HMHN consists of four components as follows: (1) Grid-Wise Attention Module (GWA), (2) Deep Feature Fusion Block (DFF), (3) Multi-head cross Attention Block (MAB), and (4) Attention Fusion Block (AFB). Concretely, GWA and DFF are designed to model the long-range dependencies among different regions of the low-level facial image. Next, MAB further measures the high-level detail features from multiple facial regions by combining multiple attention heads, consisting of spatial and channel attention. At the same time, the AFB module makes the attention maps extracted by the MAB focus on different regions, which enables the HMHN to capture several depression-related face regions simultaneously. Finally, AFB outputs the depression severity (BDI-II Score).

The main contributions of this study can be summarized as follows:

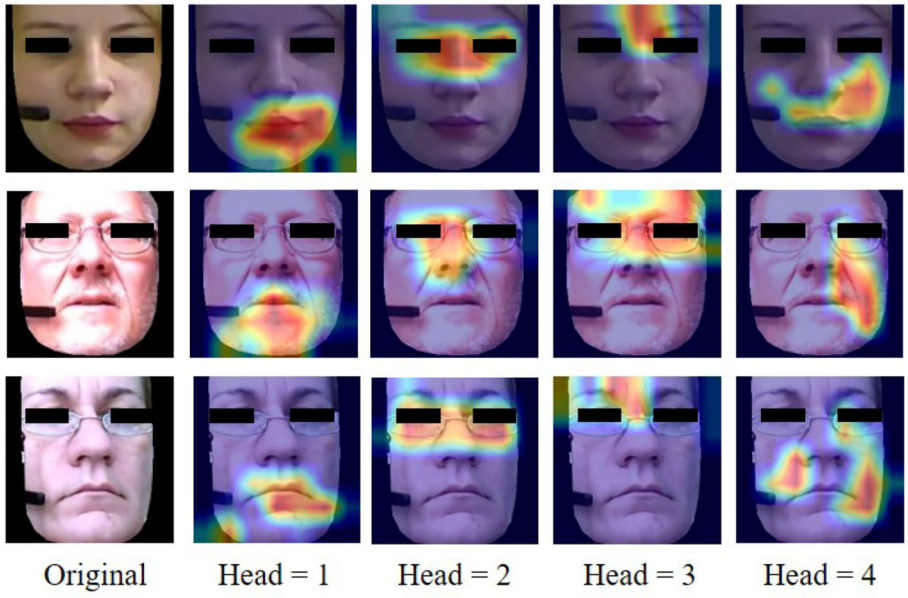



FIGURE 1
Example cases of visualization facial images with different cross-attention head. The first column is original facial images (BDI-II scores of 3, 16, and 44 from top to bottom), and the rest of the columns are generated by four cross-attention heads from HMHN.

- We propose an integrated end-to-end framework HMHN, which effectively captures the facial dynamics information from multi-region as a non-verbal behavior measure for estimating the severity of depression scale.
- To regularize the convolutional parameter learning in the low-level feature extraction for facial depression recognition, we design grid-wise attention and DFF block, which can model long-range dependencies between different facial regions.
- To address the problem that a single attention module cannot adequately capture the subtle depression features of faces, we propose MAB and AFB. On the one hand, MAB further extracts high-level detail features. On the other hand, AFB is designed to capture multiple non-overlapping attention regions and fuse them to encode high-order interactions among local features.
- We conduct the compared experiments on two publicly benchmark depression datasets [i.e., AVEC 2013 (Valstar et al., 2013) and AVEC 2014 (Valstar et al., 2014) depression datasets]. The results demonstrate that our method is promising against several state-of-the-art alternative methods. Moreover, we also do an ablation study that specifically demonstrates the effectiveness of each component in our model.

The structure of the remaining chapters is provided as follows. We, first, briefly discussed the related work in Section 2, and the proposed depression recognition method is described in Section 3. Section 4 demonstrates the dataset and experimental settings. The results and discussions are presented in Section 5, and Section 6 concludes the study.

## 2. Related work

### 2.1. Hand-engineered methods

In the third and fourth Audio-Visual Emotion recognition Challenge depression sub-challenges (AVEC 2013/14), the datasets for depression level prediction are publicly released, which contributed notably to research on automatic depression detection. In the AVEC 2013 depression sub-challenges, they use the Local Phase Quantization (LPQ; Ojansivu and Heikkilä, 2008) feature descriptor as visual features to predict the BDI-II score. Cummins et al. (2013) investigate Space-Time Interest Points (STIP; Laptev et al., 2008) and Pyramid of Histogram of Gradient (PHOG; Bosch et al., 2007) descriptors for extraction of behavioral cues for depression analysis. Meng et al. (2013) propose to use Motion History Histogram (MHH) feature (Meng and Pears, 2009) to model motion in videos by improving the Motion History Image (MHI) in the field of action recognition, and the Partial Least Squares (PLS; De Jong, 1993) is employed for regression learning. Wen et al. (2015) propose to encode temporal information based on Local Phase Quantization from Three Orthogonal Plane (LPQ-TOP) features from sub-volumes of the facial region through discriminative mapping and decision fusion, and the recognition performance is further improved. The following research on the AVEC 2013 dataset relies on Median Robust Local Binary Patterns from Three Orthogonal Planes (MRLBP-TOP; He et al., 2018) and Local Second-Order Gradient Cross Pattern (LSOGCP; Niu et al., 2019). In the AVEC 2014 depression sub-challenges, the author extracted the Local Gabor Binary Pattern (LGBP; Zhang et al., 2005) feature from the XY-T place of video to predict the

BDI-II score. In the study by Dhall and Goecke (2015), Local Binary Patterns (LBP) from three orthogonal plane (TOP) feature descriptors have been considered effective for predicting the scale of depression. In the study by Pérez Espinosa et al. (2014), they use dynamic facial features extracted by LGBP from Three Orthogonal Planes (LGBP-TOP) to predict depression level, another variant of LBP-TOP.

The above methods based on hand-crafted feature descriptors have some positive results in the field of depression recognition. However, they still have some limitations. For instance, hand-crafted features are highly dependent on expert knowledge and cannot extract complex semantic information.

## 2.2. Deep learning methods

As deep networks can extract deeper and more spatial inductive biases information, deep learning methods have gained their prevalence in facial depression recognition tasks. According to combined facial appearance with dynamic features (optical flow) in fully connected layers, Zhu et al. (2017) fine-tune to adopt deep models (GoogleNet), pre-trained on the CASIA (Yi et al., 2014) large facial dataset for predicting BDI scores from video data, and achieve positive performance on AVEC 2013 and AVEC 2014 depression datasets. Zhou et al. (2020) propose a multi-region DepressNet neural network by blending different facial regions on the basis of ResNet-50 (He et al., 2016), proving that the combination of multiple sub-models can improve the performance of depression recognition. In the study by De Melo et al. (2019), Melo et al. adopt a 2D-CNN and distribution learning to predict the BDI-II score from facial images. Similarly, many of the following works using pre-trained CNNs fine-tune their deep architectures on the AVEC 2013 and AVEC 2014 datasets to estimate and prediction (e.g., Kang et al., 2017; De Melo et al., 2020; He et al., 2022a). He et al. (2021a) combine the attention mechanism with CNN to construct an end-to-end depression recognition model named LGA-CNN. He et al. (2022b) also designed an end-to-end framework called the SAN to re-label the uncertain labels for automatic depression estimation. Niu et al. (2022) utilize a pre-trained ResNet-50 model to process video clips. They employed a graph convolution embedding block and a multi-scale vectorization block to capture and represent facial dynamics for predicting BDI-II scores, which reflect the severity of depression. Liu et al. (2023) propose an end-to-end depression recognition model called PRA-Net. They divide the input facial images into parts and calculate the feature weight of each part. Then, they combine the parts using a relation attention module. PRA-Net utilizes part-based and relation-based attention mechanisms to improve the model's performance.

To extract depression cues from the perspective of spatial structure and temporal changes, various studies have been proposed to model spatio-temporal information for depression recognition. Al Jazaery and Guo (2018) have automatically learned spatio-temporal features of face regions at two different scales by using 3D Convolutional Neural Network (C3D) and Recurrent Reural Network (RNN), which can model the local and global spatio-temporal information from continuous facial expressions to predict depression levels. De Melo et al. (2020) designed a novel 3D framework to learn spatio-temporal patterns by combining the full-face and local regions. Uddin et al. (2020) introduce a new two-stream network to model the sequence information from video data. In addition, the 3D-CNN is also used in the study by De Melo et al. (2021) and He et al. (2021b) to capture informative representations for analyzing the severity of depression. In contrast to the above methods, our HMHN achieves comparable results using only facial visual information.

As mentioned above, the existing approaches extract high-level representations of depression cues through CNN, but there are still some problems. First, most of these depression estimation methods are not end-to-end schemes, which increases the difficulty of clinical application. Second, most of these models do not consider convolutional filters' properties in different feature learning stages. This would generally lead the model to pay attention to a single rough area of the face while ignoring other important areas contributing to depression identification. Therefore, to address these problems, we propose a multi-stage hybrid attention structure that considers the long-range inductive biases in low-level feature learning and high semantic feature representation. Multiple non-overlapping attention regions could be activated simultaneously to capture fined-grain depression features from different facial regions. Experimental results on AVEC 2013 and AVEC 2014 depression datasets illustrate the effectiveness of our method.

# 3. Methodology

## 3.1. Framework overview

The proposed end-to-end depression recognition framework HMHN is presented in Figure 2. To learn high-discriminative attentional features with facial depression details, we first extract the long-range biases between different facial regions by GWA and DFF. Second, the MAB takes the features from the DFF module as input and captures several facial regions with depression information. Then, the AFB module attempts to train these attention maps (i.e., outputs from the MAB module), to focus on non-coincident facial areas and merge these attention maps, which predicts the BDI-II score. In the following, we will describe each component in HMHN detail.

## 3.2. Grid-wise attention

To learn long-range bias in low-level feature extraction of facial images and mine discriminative features with facial depressive patterns without relying on large-scale datasets, motivated by Huang et al. (2021), we introduce the grid-attention mechanism, which mainly includes two parts, local grid feature extraction and grid-wise attention calculation. The details are presented in the following sections.

### 3.2.1. Local grid feature extraction network

The facial images are cropped and aligned according to their eye positions and resized to $224 \times 224 \times 3$ by the machine

TABLE 2 The configuration of local grid feature extraction network.

| Input | Operator | Kernel | Output |
|---|---|---|---|
| $C \times \frac{H}{h} \times \frac{W}{w}$ | Convolution | 1 × 1, Stride 1 | $(Ck) \times \frac{H}{h} \times \frac{W}{w}$ |
| $Ck \times \frac{H}{h} \times \frac{W}{w}$ | BatchNorm | / | $(Ck) \times \frac{H}{h} \times \frac{W}{w}$ |
| $Ck \times \frac{H}{h} \times \frac{W}{w}$ | LeakyRelu | / | $(Ck) \times \frac{H}{h} \times \frac{W}{w}$ |
| $Ck \times \frac{H}{h} \times \frac{W}{w}$ | Convolution | 1 × 1, Stride 1 | $C \times \frac{H}{h} \times \frac{W}{w}$ |
| $C \times \frac{H}{h} \times \frac{W}{w}$ | BatchNorm | / | $C \times \frac{H}{h} \times \frac{W}{w}$ |
| $C \times \frac{H}{h} \times \frac{W}{w}$ | LeakyRelu | / | $C \times \frac{H}{h} \times \frac{W}{w}$ |

learning toolkit Dlib (King, 2009). Then, it divided into $h \times w$ grids before being forwarded to the local grid feature extraction network (LGFE), to extract the depression discrimination information in each grid. The details are as follows:

$$\mathbf{Grid}(g, h, w) = \left\{ g_{1,1}^{C\times\frac{H}{h}\times\frac{W}{w}}, \ldots, g_{i,j}^{C\times\frac{H}{h}\times\frac{W}{w}}, \ldots \right\} \tag{1}$$

$$\hat{\mathcal{G}}^{hw\times C\times\frac{H}{h}\times\frac{W}{w}} = \mathbf{LGFE}\left(g^{hw\times C\times\frac{H}{h}\times\frac{W}{w}}\right), \tag{2}$$

$$\widehat{\mathcal{G}}_{i,j} = \mathbf{LGFE}\left(g_{i,j}\right) \tag{3}$$

where $H$, $W$, and $C$ are the height, width, and channels of the original image, respectively. $g_{i,j}^{C\times\frac{H}{h}\times\frac{W}{w}}$ represents that the input image $g$ is divided into $h \times w$ grids, every grid is with a shape of $C \times \frac{H}{h} \times \frac{W}{w}$ and locates in the $i$ th row and the $j$ th column in $g$. Next, as shown in the Equations (2) and (3), each grid will be forwarded to the LGFE, and the local depression feature of the facial region learned is defined as $\widehat{\mathcal{G}}_{i,j}$. We believe that every grid features a respective contribution to depression recognition. Therefore, these feature maps are forwarded to the grid-wise attention calculation to weight their importance. The structure of the LGFE is shown in Table 2.

### 3.2.2. Grid attention calculation

To better extract the depressive features of facial regions, after the LGFE block, the relationship between different facial regions is constructed through grid attention calculation, which is defined as follows:

$$\mathbf{Att}_{q,k} = \delta\left(\frac{q \cdot k}{d_k}\right) \tag{4}$$

where $d_k = \frac{W}{w}$, $q = \hat{\mathcal{G}}^{hw\times C\times\frac{H}{h}\times\frac{W}{w}}$, and $k = \hat{\mathcal{G}}^{hw\times C\times\frac{W}{w}\times\frac{H}{h}}$, and $\delta$ stand for the softmax operation.

Then, the adaptive average pooling is used to squeeze each channel into a scalar after an attention mechanism and expand the channel back to the original shape. The process is formulated as follows:

$$\widetilde{\mathcal{G}}^{hw\times C\times\frac{H}{h}\times\frac{W}{w}} = Aavp\left(\mathbf{Att}_{q,k}\right) * Ones\left(\frac{H}{h}, \frac{W}{w}\right) \tag{5}$$

where "$*$" represents the scalar matrix product between matrices with a broadcasting property. $Aavp(\cdot)$ denoted an adaptive average pooling technique that converts an operand matrix into a scalar and $Ones\left(\frac{H}{h}, \frac{W}{w}\right)$ is to generate a matrix with all elements being equal to 1 in the shape of $\left(\frac{H}{h}, \frac{W}{w}\right)$.

$$\widetilde{\mathcal{G}}^{C\times H\times W} = Ungrid\left(\widetilde{\mathcal{G}}^{hw\times C\times\frac{H}{h}\times\frac{W}{w}}\right) * g^{C\times H\times W} \tag{6}$$

where $Ungrid(\cdot)$ is the reverse operation of Equation (1), which is used to convert these grid attention maps back to the shape of the original facial image and concat these weights back to the shape of the original matrix.

Thus, the resulting $\widetilde{\mathcal{G}}^{C\times H\times W}$ is a feature map that takes into account the long-range bias between different facial regions in the low-level visual depression feature learning stage.

## 3.3. Deep feature fusion

To further extract the depressive features of the face, we fuse the features between the original image $g$ and the weighted feature map $\widetilde{\mathcal{G}}$ of the backbone network by applying residual network technology. In particular, based on the experimental results in Section 5, we choose to remove the average pooling, flattening, and fully connected layer from ResNet-18 (He et al., 2016) as the backbone. The overall structure of the deep feature fusion block is shown in Figure 3. It mainly includes two feature transformation networks and one feature fusion network. These two feature transformation networks share the structure but not the learning parameters. The mathematical definition is as follows:

$$\bar{\mathcal{G}}^{C\times H\times W} = \mathbf{DFF}(\mathbf{FT}_1\left(g\right) + \mathbf{FT}_2\left(\widetilde{\mathcal{G}}\right)) \tag{7}$$

where $\mathbf{FT}_i(\cdot)$ ($i$=1,2) is the feature transformation network of the original facial image $g$ and the weight feature $\widetilde{\mathcal{G}}$ extracted from the GWA module, respectively. **DFF** denotes the deep feature fusion network. Finally, the obtained feature map $\bar{\mathcal{G}}^{C\times H\times W}$ is forwarded to the candidate backbone network.

## 3.4. Multi-head cross attention block

Facial depression behavior is usually manifested by multiple facial regions simultaneously. The GWA module first extracts the low-level local features of the face in HMHN. Then, we need to encode the high-level interactions between local features by multi-head cross-attention block to achieve a holistic approach. The detailed structure of the MAB block is shown in Figure 4. It is composed of parallel cross-head attention units, which are combinations of spatial and channel attention units that remain independent.

More concretely, The spatial attention unit is shown in the left part of Figure 4. We first feed the input features into the 1 × 1 convolution layer to reduce the channel number. Next, we construct the 3 × 3, 1 × 3, and 3 × 1 convolution kernels to efficiently capture spatial relationships. In general, the spatial attention unit consists of four convolution layers and one activation function to capture local features at multiple scales. The channel attention unit shown in the right part of Figure 4 consists of two linear layers and one activation function. We take advantage of

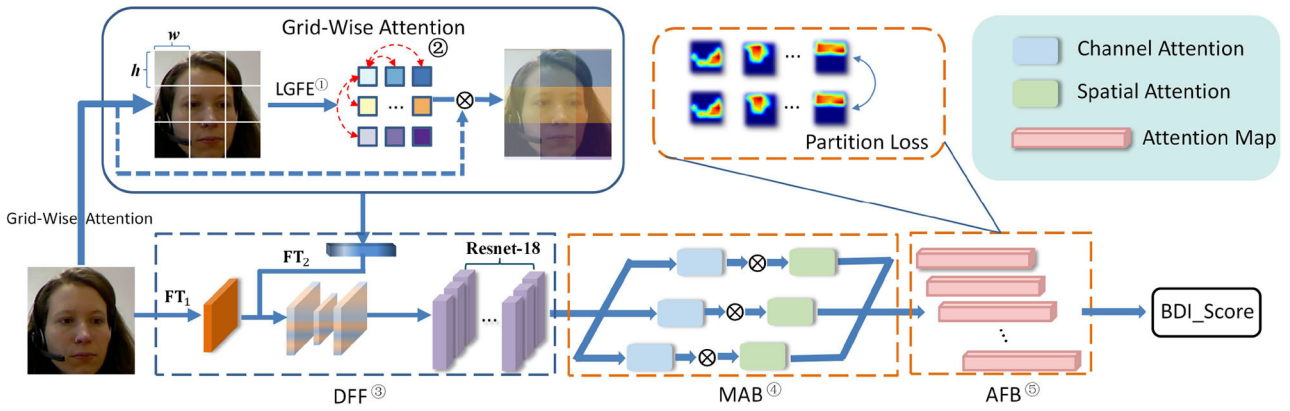



FIGURE 2
The framework of HMHN. Notations: ① Local Grid Feature Extraction, ② Grid-Wise Attention Calculation, ③ Deep Feature Fusion Block, ④ Multi-head cross Attention Block, and ⑤ Attention Fusion Block. DFF, deep feature fusion block; $FT_i$, feature transformation network in DFF; MAB, multi-head cross attention block; AFB, attention fusion block.

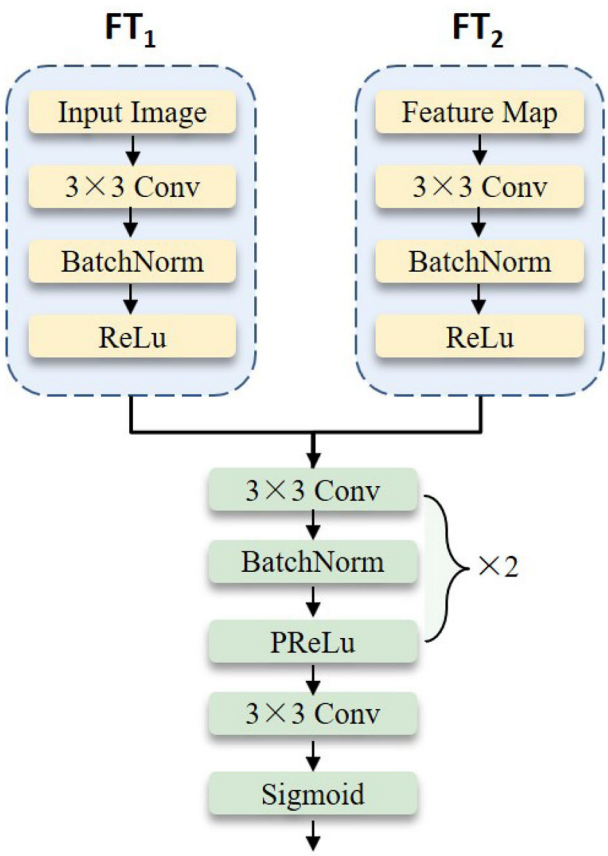



FIGURE 3
The detailed illustration of the deep feature fusion block.

two linear layers to achieve a mini autoencoder to encode channel information.

Mathematically, the above process can be formulated as follows:

$$S_i = \bar{\mathcal{G}} \times H_i\left(\theta_s, \bar{\mathcal{G}}\right), i \in \{1, k\} \tag{8}$$

$$C_i = S_i \times H_i'\left(\theta_c, S_i\right), i \in \{1, k\} \tag{9}$$

where $k$ is the number of cross attention heads. $H_i$ and $H_i'$ are defined as the spatial attention head and the channel attention head, respectively, $\theta_s$ and $\theta_c$ are their parameters. $S_i$ and $C_i$ represent the output of the $i$-h spatial attention and channel attention, separately.

## 3.5. Attention fusion block

After going through several modules above, our HMHN has been able to capture subtle facial depression features, but the multi-head construction could not learn attention maps in an orchestrated fashion. In other words, we hope that different branches can focus on different facial regions as much as possible and fuse the depression feature information of each head. To achieve this aim, we propose that the AFB enhance further the features learned by MAB. In the meantime, the cross-attention heads are supervised to center on different critical regions and avoid overlapping attention using the partition loss, which is defined as follows:

$$\mathcal{L}_{sum} = \mathcal{L}_{att} + \mathcal{L}_{mse} \tag{10}$$

$$\mathcal{L}_{att} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C}\log\left(1 + \frac{k}{\sigma_{ij}^2}\right) \tag{11}$$

This loss contains two components, where $\mathcal{L}_{mse}$ is the square loss for regression and $\mathcal{L}_{att}$ is partition loss to maximize the variance among the attention maps, $k$ is the number of cross attention, $N$ is the number of samples, $C$ is the channel size of the attention maps, and $\sigma_{ij}^2$ is denoted the variance of the $j$-th channel on the $i$-th sample. The merged attention map is then used for computing the BDI-II score with a regression output layer. Finally, we learn the deep discriminative features by jointly minimizing the unified loss functions $\mathcal{L}_{sum}$.

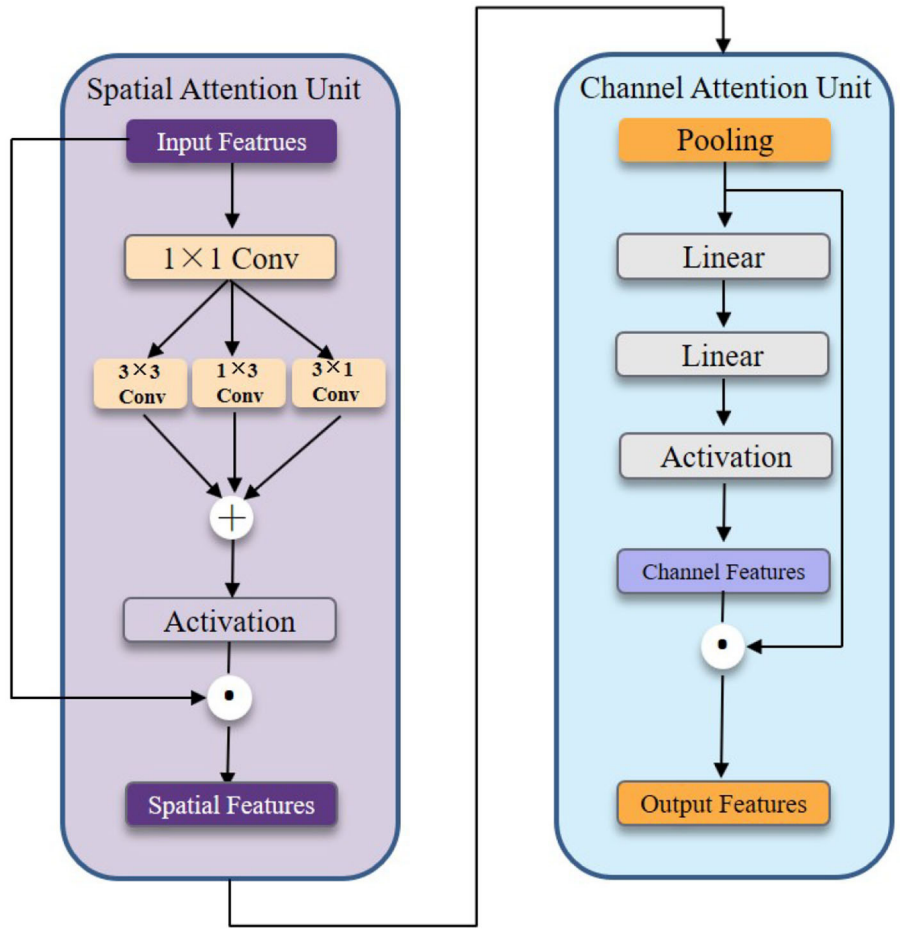



FIGURE 4
The detailed illustration of the cross attention head.

TABLE 3 Ablation study of the individual components on the test set of AVEC 2013.

| Combination | Evaluation metrics | |
|---|---|---|
| | MAE | RMSE |
| **A1**: Resnet18 (backbone) | 8.47 | 9.32 |
| **B1**: Resnet18+GWA | 7.68 | 8.31 |
| **C1**: Resnet18+GWA+DFF | 7.49 | 8.29 |
| **D1**: Resnet18+MAB+AFB | 6.88 | 7.91 |
| **E1**: Resnet18+DFF+GWA+MAB+AFB (Ours) | **6.05** | **7.38** |

TABLE 4 Ablation study of the individual components on the test set of AVEC 2014.

| Combination | Evaluation metrics | |
|---|---|---|
| | MAE | RMSE |
| **A2**: Resnet18 (backbone) | 8.38 | 9.13 |
| **B2**: Resnet18+GWA | 7.57 | 8.47 |
| **C2**: Resnet18+GWA+DFF | 7.41 | 8.46 |
| **D2**: Resnet18+MAB+AFB | 6.90 | 8.13 |
| **E2**: Resnet18+DFF+GWA+MAB+AFB (Ours) | **6.01** | **7.60** |

The bold values indicate the best results.

## 4. Experiments

In order to demonstrate the effectiveness of our depression recognition approach, we conducted experiments on two publicly available datasets, namely AVEC 2013 and AVEC 2014. Compare our performance with start-of-the-art methods, and demonstrate the effectiveness of each component in our model by an ablation study. This section presents a description of the dataset, data pre-processing, experimental setting and evaluation metrics.

### 4.1. AVEC 2013 and AVEC 2014 datasets

In the present paper, all experiments are evaluated on AVEC 2013 and AVEC 2014 depression datasets. The distribution of the

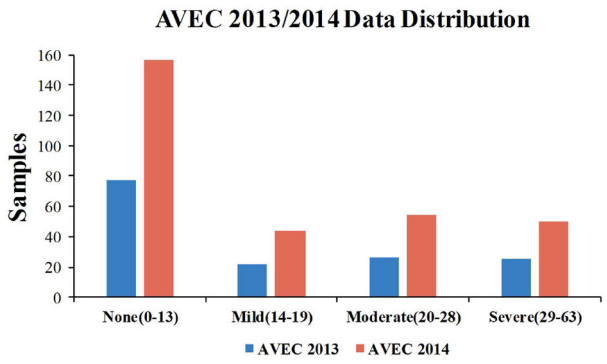



FIGURE 5
The distribution of BDI-II scores in the AVEC 2013 and AVEC 2014 datasets.

BDI-II scores in both the AVEC 2013 and AVEC 2014 datasets is shown in Figure 5.

For the AVEC 2013 depression dataset, there are 150 video clips recorded by 82 subjects participating in human-computer interaction (HCI) task with a microphone and a webcam to record the information. The age range for all subjects in the dataset is 18–63 years old, with an average age is 31.5 years old and a standard deviation of 12.3 years. These video recordings are set to 30 frames per second (fps) with a resolution of 640 × 480 pixels. This depression dataset has been divided into three partitions by the publisher, i.e., training, development, and test set. For every partition, it has 50 videos, and each video has a label corresponding to its depression severity level, which is assessed based on the BDI-II questionnaire.

The AVEC 2014 depression dataset is a subset of the AVEC 2013 dataset. There are two tasks included: FreeForm and Northwind, both of which have 150 video clips. Specifically, in the "FreeForm" task, the subjects responded to several questions, such as describing a sad childhood memory or saying their favorite dish. In the "Northwind" task, the subjects are required to read an excerpt audibly from a fable. The same as AVEC 2013, it also has three partitions, i.e., training, development, and test sets. We perform experiments employing training and development sets from both tasks as training data, and the test sets are used to measure the performance of the model.

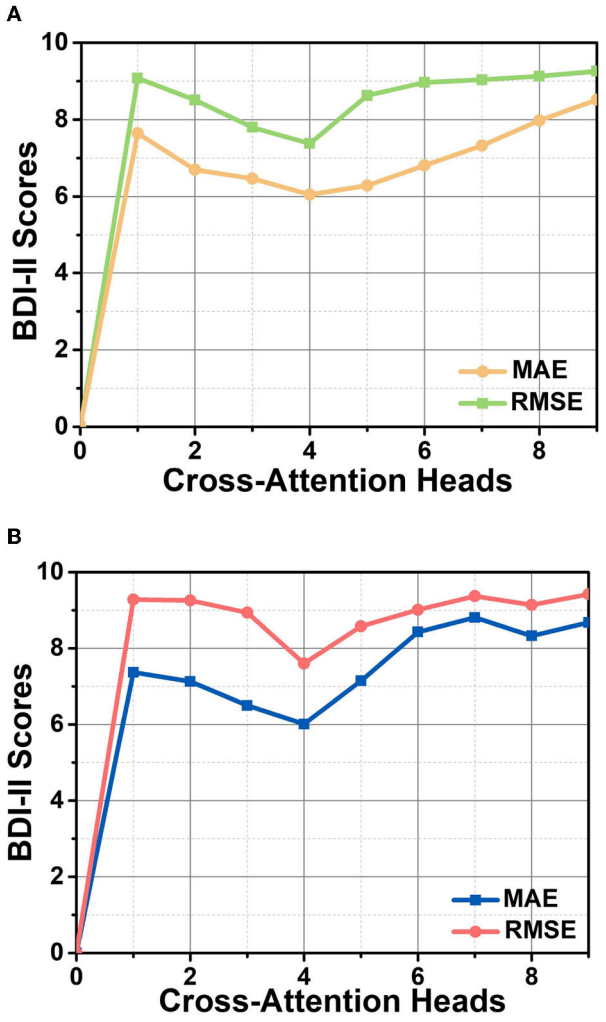



FIGURE 6
The performance of the HMHN architecture in terms of RMSE and MAE for various sizes of cross-attention head on AVEC 2013 **(A)** and AVEC 2014 **(B)** datasets.

## 4.2. Experimental settings and evaluation metrics

### 4.2.1. Experimental settings

The overall framework of HMHN is shown in Figure 2. A machine learning toolkit DliB (King, 2009) is adopted to resize the generated facial images to 224 × 224 with RGB color channels. Instead of using a pre-trained architecture to predict depression severity, we directly train the whole framework in an end-to-end fashion. To be specific, our experimental code is implemented with Pytorch (Paszke et al., 2019), and the models are trained on a local GPU server with a TESLA-A100 GPU (40 G global memory). In order to obtain fast convergence, we use the AdamW (Loshchilov and Hutter, 2017) optimizer with an adaptive learning rate strategy, and its initial learning rate is 0.001, The batch size is 64, the dropout rate is 0.2, and the learning factor is set to 0.1.

### 4.2.2. Evaluation metrics

The performance of the baseline models is assessed on AVEC 2013 and AVEC 2014 datasets in terms of two evaluation metrics—Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Afterward, many researchers have been adopting these two metrics to evaluate the prediction accuracy of their works. This study also regards RMSE and MAE as the metrics during testing to make an equitable comparison, which details are defined as:

$$MAE = \frac{1}{M}\sum_{j=1}^{M}\left|\hat{\ell}_j - \ell_j\right| \tag{12}$$

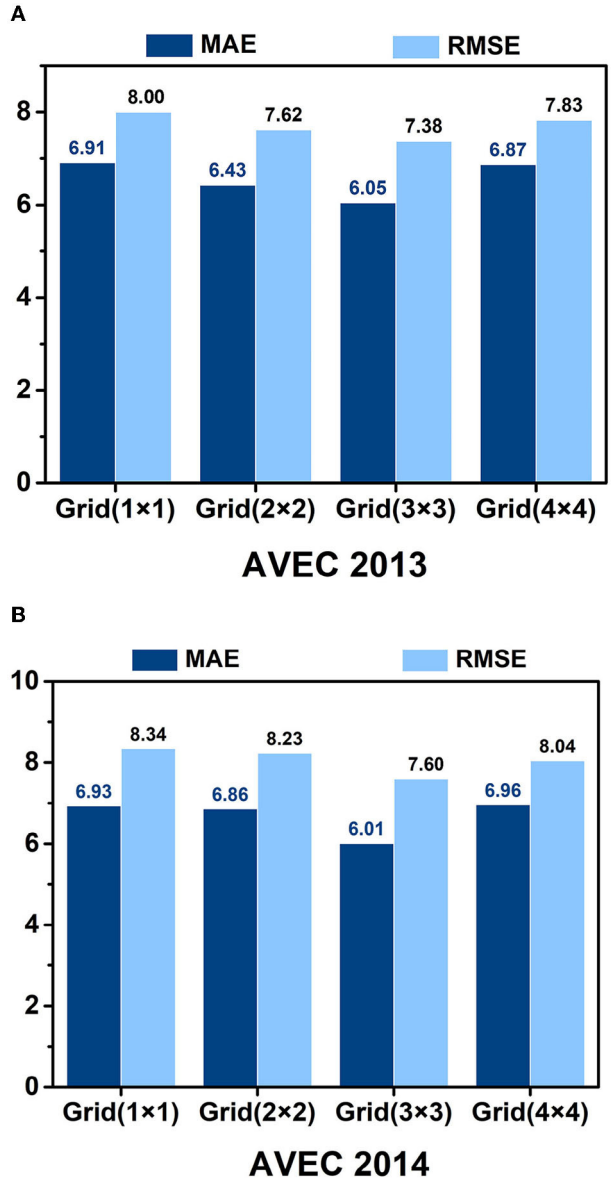



FIGURE 7
Recognition results of different grid parameters of the HMHN model on the AVEC 2013 **(A)** and AVEC 2014 **(B)** datasets.

$$RMSE = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\hat{\ell}_j - \ell_j\right)^2} \tag{13}$$

where $M$ is the total number of video samples, $\ell_j$ and $\hat{\ell}_j$ are the ground truth and the predicted BDI-II score of the $j$-th subject, respectively.

# 5. Experimental results and discussion

In this section, we first perform an ablation study to examine the effectiveness of individual components in the propose framework. Then, we compare the architecture with several state-of-the-art vision-based depression analysis methods to show its promising performance.

## 5.1. Ablation study

In order to verify the effectiveness of the proposed HMHN, we carry out the ablation studies on AVEC 2013 and AVEC 2014 datasets to assess the efficacy of critical components in our method. The results are shown in Tables 3, 4. Specifically, Resnet18+GWA (*B*1,*B*2) outperforms the backbone network (*A*1,*A*2) on both datasets owing to GWA can learn long-range bias in low-level features of facial images. *D*1 and *D*2 are improved by MAB and AFB, which capture multiple non-overlapping attention simultaneously. *E*1 and *E*2 integrate all modules, yielding better results than using them separately. This observation demonstrates that the multi-stage attention mechanism performs better than the one-stage attention mechanism. The prediction accuracy of depression level can be effectively improved by encoding the low-level to high-level interactions between depression discriminative features of multiple facial regions.

## 5.2. Number of the cross attention heads

We opt different numbers of cross-attention heads to observe their effect on the depression recognition performance of the model, allowing us to select an optimal cross-attention head size. The results are shown in Figure 6, where the lines with different colors represent the two evaluation metrics, RMSE and MAE, respectively. The top and bottom figures indicate experimental results on two different datasets, AVEC 2013 and AVEC 2014. It is apparent that the increasing number of layers does not imply an improvement in the performance, and equipping four cross-attention heads maximizes the model's performance. It is probably related that facial depression recognition is affected by multiple facial regions. The single attention head cannot sufficiently capture all the subtle and complex appearance variations, while too many attention heads make the attention regions overly distributed. As shown in Figure 6, our method explicitly learns to attend to multiple local image regions for facial depression recognition.

## 5.3. Impact of the grid size

We examine the impact of grid parameters on the model's performance, as evidenced in Figure 7. Our findings indicate that utilizing a grid strategy generally leads to improved performance over not using a grid strategy. The **Grid**($3 \times 3$) achieves the best results among the tested grid parameters, with an MAE of 6.05 and an RMSE of 7.38 on the AVEC 2013 dataset, and MAE = 6.01 and RMSE = 7.60 on the AVEC 2014 dataset. This phenomenon may be related to the spatial position and size of the grid, as an overly large or small grid size may limit the expression ability of the receptive field and interfere with the acquisition of depression information across facial regions.

TABLE 5 Kernel size of separable convolution on AVEC 2013 and AVEC 2014 datasets.

| Kernel settings | Params(M) | AVEC 2013 | | AVEC 2014 | |
|---|---|---|---|---|---|
| | | MAE | RMSE | MAE | RMSE |
| Standard Conv | 29.33 | 6.07 | 7.43 | 6.09 | 7.66 |
| $(1 \times 7, 7 \times 7, 7 \times 1)$ | 26.57 | 6.16 | 7.49 | 6.19 | 7.78 |
| $(1 \times 5, 5 \times 5, 5 \times 1)$ | 22.63 | 6.14 | 7.51 | 6.12 | 7.71 |
| $(3 \times 1, 1 \times 3)$ | 17.78 | 6.21 | 7.56 | 6.27 | 7.83 |
| $(3 \times 3, 1 \times 3, 3 \times 1)$ | 19.72 | **6.05** | **7.38** | **6.01** | **7.60** |

The bold values indicate the best results.

## 5.4. Kernel size of separable convolutions

We conduct experiments to evaluate the effect of separable convolutions in MAB modules. We test standard convolutions and separable convolutions with different kernel sizes. According to our experimental results, as shown in Table 5, using a separable convolution model with a smaller kernel size ($1 \times 3$, $3 \times 3$, $3 \times 1$) performs better than using a larger kernel size such as ($1 \times 7$, $7 \times 7$, $7 \times 1$) and ($1 \times 5$, $5 \times 5$, $5 \times 1$). In addition, we also find that separable convolutions can achieve similar performance with fewer parameters than standard convolutions. For example, on the AVEC 2013 dataset, the MAE of the separable convolution model with convolution kernel sizes ($1 \times 3$, $3 \times 3$, $3 \times 1$) is 6.05, and the RMSE is 7.38. Compared with using standard convolution, the number of separable convolution parameters is reduced by 32.8%.

## 5.5. Comparison with state-of-the-art methods

In order to further demonstrate the depressive recognition performance of the proposed model, We present the quantitative performance comparison results in Tables 6, 7 for AVEC 2013 and AVEC 2014, respectively. Specifically, models in Valstar et al. (2013, 2014), Wen et al. (2015), He et al. (2018), and Niu et al. (2019) are based on hand-crafted representations. Our method outperforms all other methods, mainly because hand-crafted features rely on researchers' experiences, and it is difficult to characterize depression cues fully. At the same time, our HMHN uses deep neural networks and the multi-attention stage mechanism, which can capture complete semantic information, thereby improving the prediction performance.

For the methods using deep neural networks, Zhu et al. (2017), Al Jazaery and Guo (2018), Zhou et al. (2020), and He et al. (2022a) train the deep models on a large dataset and then fine-tune on the AVEC 2013 and AVEC 2014 datasets. HMHN is an end-to-end scheme for depression recognition and achieves an impressive performance even without a pre-trained model on other large-scale datasets. As shown in Tables 6, 7, we achieve the best performance among end-to-end methods on the AVEC 2013 (MAE = 6.05, RMSE = 7.38) and AVEC 2014 (MAE = 6.01, RMSE = 7.60) datasets. We also achieve the second-best performance compared to other methods pre-trained on large-scale datasets. Specifically, Zhou et al. (2020) propose a CNN-based visual depression recognition model by roughly dividing the facial region into three parts and then combined with the overall facial image to improve the recognition performance of the model. Our better performance is due to the multi-stage attention mechanism for the extraction of depressive features, and Zhou et al.'s visualization results show that their model focuses attention on only one region and ignores other facial details that contribute to depression recognition. In contrast, He et al. (2021a) achieves a passable performance without a pre-trained model. The authors divide the facial region by facial landmark points, then block the feature map to extract local feature information. Finally, the feature aggregation method is used to automatically learn the facial region's local and global feature information. He et al. (2021b, 2022b) and Liu et al. (2023) are also end-to-end methods. Our HMHN outperforms those methods by a significant margin. One important reason is that we consider the long-range inductive biases in both low-level feature learning and high-semantic feature representation. At the same time, Niu et al. (2022) improve the prediction accuracy of depression levels by investigating the correlation between channels and vectorizing the tensors along the time and channel dimensions. De Melo et al. (2020) to encode the smooth and sudden facial expression variations to assess individual BDI-II scores. These two methods model the spatio-temporal information of facial regions; our propose is trained from scratch using only facial visual information and achieves comparable results.

## 5.6. Visual analysis

In order to intuitively observe how the model predicts depression scores from facial images, we present the visualized facial images with different cross-attention heads in Figure 1. The first column of Figure 1 shows the original images, and the second to fifth columns represent the attention regions of different cross-attention heads. The heatmap in the faces is the focus area learned by the model. Our model can attend to multiple locations simultaneously before fusing the attention maps. Our HMHN model specifically focuses on the facial muscle movement regions related to depression, such as the mouth, eyebrows, and eyes, while suppressing irrelevant regions.

# 6. Conclusion

In this paper, an end-to-end two-stage attention mechanism architecture named HMHN for predicting an individual's

TABLE 6 Depression level prediction performance compared with different methods on the AVEC 2013 test set.

| Category | Methods | MAE | RMSE |
|---|---|---|---|
| Pre-trained | Valstar et al. (2013)/LPQ | 10.88 | 13.61 |
| | Cummins et al. (2013)/PHOG | / | 10.45 |
| | Wen et al. (2015)/LPQ-TOP | 8.22 | 10.27 |
| | He et al. (2018)/MRLBP-TOP, DPFV | 7.55 | 9.20 |
| | Niu et al. (2019)/LSOGCP | 6.97 | 9.17 |
| | Zhu et al. (2017)/Optical Flow, 2D-CNN | 7.58 | 9.82 |
| | Al Jazaery and Guo (2018)/C3D, RNN | 7.37 | 9.28 |
| | De Melo et al. (2019)/ResNet-50 | 6.30 | 8.25 |
| | Zhou et al. (2020)/2D-CNN | 6.20 | 8.28 |
| | De Melo et al. (2020)/Two-Stream | **5.96** | 7.97 |
| | Uddin et al. (2020)/LSTM | 7.04 | 8.93 |
| | De Melo et al. (2021)/MDN | 6.59 | 8.39 |
| | Niu et al. (2022)/2D-CNN | 6.12 | 7.49 |
| | He et al. (2022a)/2D-CNN | 7.36 | 9.17 |
| End-to-end | He et al. (2021a)/2D-CNN, Attention | 6.59 | 8.39 |
| | He et al. (2021b)/3D-CNN | 6.83 | 8.46 |
| | He et al. (2022b)/2D-CNN | 7.02 | 9.37 |
| | Liu et al. (2023)/2D-CNN, Attention | 6.08 | 7.59 |
| | Ours | **6.05** | **7.38** |

The bold values indicate the best results.

TABLE 7 Depression level prediction performance compared with different methods on the AVEC 2014 test set.

| Category | Methods | MAE | RMSE |
|---|---|---|---|
| Pre-trained | Valstar et al. (2014)/LGBP-TOP | 8.86 | 10.86 |
| | Dhall and Goecke (2015)/LBP-TOP | 7.08 | 8.91 |
| | He et al. (2018)/MRLBP-TOP, DPFV | 7.21 | 9.01 |
| | Niu et al. (2019)/LSOGCP | 7.19 | 9.10 |
| | Zhu et al. (2017)/Optical Flow, 2D-CNN | 7.47 | 9.55 |
| | Al Jazaery and Guo (2018)/C3D, RNN | 7.22 | 9.20 |
| | De Melo et al. (2019)/ResNet-50 | 6.13 | 8.23 |
| | Zhou et al. (2020)/2D-CNN | 6.21 | 8.39 |
| | De Melo et al. (2020)/Two-Stream | 6.20 | 7.94 |
| | Uddin et al. (2020)/LSTM | 6.86 | 8.78 |
| | De Melo et al. (2021)/MDN | 6.06 | 7.65 |
| | Niu et al. (2022)/2D-CNN | **6.01** | **7.56** |
| | He et al. (2022a)/2D-CNN | 7.26 | 9.03 |
| End-to-end | He et al. (2021a)/2D-CNN, Attention | 6.51 | 8.30 |
| | He et al. (2021b)/3D-CNN | 6.78 | 8.42 |
| | He et al. (2022b)/2D-CNN | 6.95 | 9.24 |
| | Liu et al. (2023)/2D-CNN, Attention | 6.04 | 7.98 |
| | Ours | **6.01** | **7.60** |

The bold values indicate the best results.

depression level by facial images is proposed. HMHN can focus on multiple depression feature-rich areas of the face yet is remarkably capable of recent works in recognition. Specifically, this model mainly includes four blocks: the grid-wise attention block (GWA), deep feature fusion block (DFF), multi-head cross attention block (MAB), and attention fusion block (AFB). GWA and DFF are the first stages to capture the dependencies among different regions from a facial image in a way that the parameter learning of convolutional filters is regularized. In the second stage, the MAB and AFB block is composed of parallel cross-head attention units, which combine spatial and channel attention units to obtain final facial depression features bbsy encoding higher-order interactions between local features. Experimental results on AVEC 2013 and AVEC 2014 depression datasets show the effectiveness of video-based depression recognition of the proposed framework when compared with most of the state-of-the-art approaches.

In the future, we will collect and build a dataset with more depression patients to learn more robust feature representations from the images of diverse appearances. In addition, investigation of the multi-modal (audio, video, text, etc.) depression representation learning appears to be an attractive topic.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

## Author contributions

YL and ZL conceived the study design. YL analyzed the experimental data and drafted the manuscript. ZS, LZ, and XY helped to interpret the data analysis. XH and BH were responsible for the overall planning of the dissertation. All authors agree to be accountable for the content of the work. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2019YFA0706200), in part by the National Natural Science Foundation of China (Grant Nos. 61632014, 61627808, 61802159, and 61802158), in part by Fundamental Research Funds for Central Universities (lzujbky-2019-26 and lzujbky-2021-kb26).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Al Jazaery, M., and Guo, G. (2018). Video-based depression level analysis by encoding deep spatiotemporal features. *IEEE Trans. Affect. Comput.* 12, 262–268. doi: 10.1109/TAFFC.2018.2870884

Belmaker, R. H., and Agam, G. (2008). Major depressive disorder. *N. Engl. J. Med.* 358, 55–68. doi: 10.1056/NEJMra073096

Bosch, A., Zisserman, A., and Munoz, X. (2007). "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (Amsterdam), 401–408. doi: 10.1145/1282280.1282340

Canales, J. Z., Fiquer, J. T., Campos, R. N., Soeiro-de-Souza, M. G., and Moreno, R. A. (2017). Investigation of associations between recurrence of major depressive disorder and spinal posture alignment: a quantitative cross-sectional study. *Gait Posture* 52, 258–264. doi: 10.1016/j.gaitpost.2016.12.011

Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., et al. (2009). "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (Amsterdam), 1–7. doi: 10.1109/ACII.2009.5349358

Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., and Epps, J. (2013). "Diagnosis of depression by behavioural signals: a multimodal approach," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge* (Barcelona), 11–20. doi: 10.1145/2512530.2512535

De Jong, S. (1993). Simpls: an alternative approach to partial least squares regression. *Chemometr. Intell. Labor. Syst.* 18, 251–263. doi: 10.1016/0169-7439(93)85002-X

De Melo, W. C., Granger, E., and Hadid, A. (2019). "Depression detection based on deep distribution learning," in *2019 IEEE International Conference on Image Processing (ICIP)* (Taipei), 4544–4548. doi: 10.1109/ICIP.2019.8803467

De Melo, W. C., Granger, E., and Lopez, M. B. (2020). "Encoding temporal information for automatic depression recognition from facial analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona), 1080–1084. IEEE. doi: 10.1109/ICASSP40776.2020.9054375

De Melo, W. C., Granger, E., and Lopez, M. B. (2021). MDN: a deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Trans. Affect. Comput.* 14, 578–590. doi: 10.1109/TAFFC.2021.3072579

Dhall, A., and Goecke, R. (2015). "A temporally piece-wise fisher vector approach for depression analysis," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (Xi'an), 255–259. doi: 10.1109/ACII.2015.7344580

Ellgring, H. (2007). *Non-Verbal Communication in Depression*. Cambridge University Press.

Guo, W., Yang, H., Liu, Z., Xu, Y., and Hu, B. (2021). Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. *Front. Neurosci.* 15:609760. doi: 10.3389/fnins.2021.609760

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

He, L., Chan, J. C.-W., and Wang, Z. (2021a). Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing* 422, 165–175. doi: 10.1016/j.neucom.2020.10.015

He, L., Guo, C., Tiwari, P., Pandey, H. M., and Dang, W. (2021b). Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *Int. J. Intell. Syst.* 37, 10140–10156. doi: 10.1002/int.22426

He, L., Guo, C., Tiwari, P., Su, R., Pandey, H. M., and Dang, W. (2022a). Depnet: an automated industrial intelligent system using deep learning for video-based depression analysis. *Int. J. Intell. Syst.* 37, 3815–3835. doi: 10.1002/int.22704

He, L., Jiang, D., and Sahli, H. (2018). Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding. *IEEE Trans. Multimedia* 21, 1476–1486. doi: 10.1109/TMM.2018.2877129

He, L., Tiwari, P., Lv, C., Wu, W., and Guo, L. (2022b). Reducing noisy annotations for depression estimation from facial images. *Neural Netw.* 153, 120–129. doi: 10.1016/j.neunet.2022.05.025

Huang, Q., Huang, C., Wang, X., and Jiang, F. (2021). Facial expression recognition with grid-wise attention and visual transformer. *Inform. Sci.* 580, 35–54. doi: 10.1016/j.ins.2021.08.043

Jan, A., Meng, H., Gaus, Y. F. B. A., and Zhang, F. (2017). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Trans. Cogn. Dev. Syst.* 10, 668–680. doi: 10.1109/TCDS.2017.2721552

Kang, Y., Jiang, X., Yin, Y., Shang, Y., and Zhou, X. (2017). "Deep transformation learning for depression diagnosis from facial images," *Chinese Conference on Biometric Recognition* (Shenzhen: Springer), 13–22. doi: 10.1007/978-3-319-69923-3_2

King, D. E. (2009). DLIB-Ml: a machine learning toolkit. *J. Mach. Learn. Res.* 10, 1755–1758. doi: 10.5555/1577069.1755843

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition* (Anchorage, AK), 1–8. doi: 10.1109/CVPR.2008.4587756

Liu, Z., Yuan, X., Li, Y., Shangguan, Z., Zhou, L., and Hu, B. (2023). PRA-Net: part-and-relation attention network for depression recognition from facial expression. *Comput. Biol. Med.* 2023:106589. doi: 10.1016/j.compbiomed.2023.106589

Loshchilov, I., and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. doi: 10.48550/arXiv.1711.05101

Maj, M., Stein, D. J., Parker, G., Zimmerman, M., Fava, G. A., De Hert, M., et al. (2020). The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry* 19, 269–293. doi: 10.1002/wps.20771

McPherson, A., and Martin, C. (2010). A narrative review of the beck depression inventory (BDI) and implications for its use in an alcohol-dependent population. *J. Psychiatr. Ment. Health Nursing* 17, 19–30. doi: 10.1111/j.1365-2850.2009.01469.x

Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., and Wang, Y. (2013). "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge* (Barcelona), 21–30. doi: 10.1145/2512530.2512532

Meng, H., and Pears, N. (2009). Descriptive temporal template features for visual motion recognition. *Pattern Recogn. Lett.* 30, 1049–1058. doi: 10.1016/j.patrec.2009.03.003

Michalak, J., Troje, N. F., Fischer, J., Vollmar, P., Heidenreich, T., and Schulte, D. (2009). Embodiment of sadness and depression—gait patterns associated with dysphoric mood. *Psychosom. Med.* 71, 580–587. doi: 10.1097/PSY.0b013e3181a2515c

Niu, M., He, L., Li, Y., and Liu, B. (2022). Depressioner: facial dynamic representation for automatic depression level prediction. *Expert Syst. Appl.* 2022:117512. doi: 10.1016/j.eswa.2022.117512

Niu, M., Tao, J., and Liu, B. (2019). "Local second-order gradient cross pattern for automatic depression detection," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (Cambridge), 128–132. doi: 10.1109/ACIIW.2019.8925158

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 971–987. doi: 10.1109/TPAMI.2002.1017623

Ojansivu, V., and Heikkilä, J. (2008). "Blur insensitive texture classification using local phase quantization," in *International Conference on Image and Signal Processing* (Cherbourg-Octeville: Springer), 236–243. doi: 10.1007/978-3-540-69905-7_27

Pampouchidou, A., Pediaditis, M., Kazantzaki, E., Sfakianakis, S., Apostolaki, I.-A., Argyraki, K., et al. (2020). Automated facial video-based recognition of depression and anxiety symptom severity: cross-corpus validation. *Mach. Vis. Appl.* 31, 1–19. doi: 10.1007/s00138-020-01080-7

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems, Vol. 32* (Vancouver).

Pérez Espinosa, H., Escalante, H. J., Villase nor-Pineda, L., Montes-y Gómez, M., Pinto-Aveda no, D., and Reyez-Meza, V. (2014). "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: inaoe-buap's participation at avec'14 challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (Orlando, FL), 49–55. doi: 10.1145/2661806.2661815

Scherer, S., Stratou, G., and Morency, L.-P. (2013). "Audiovisual behavior descriptors for depression assessment," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction* (Sydney), 135–140. doi: 10.1145/2522848.2522886

Schwartz, G. E., Fair, P. L., Salt, P., Mandel, M. R., and Klerman, G. L. (1976). Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science* 192, 489–491. doi: 10.1126/science.1257786

Song, S., Shen, L., and Valstar, M. (2018). "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an), 158–165. doi: 10.1109/FG.2018.00032

Uddin, M. A., Joolee, J. B., and Lee, Y.-K. (2020). Depression level prediction using deep spatiotemporal features and multilayer BI-LTSM. *IEEE Trans. Affect. Comput.* 13, 864–870. doi: 10.1109/TAFFC.2020.2970418

Valstar, M., Schuller, B., Smith, K., Almaev, T. R., Eyben, F., Krajewski, J., et al. (2014). "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (Orlando, FL), 3–10. doi: 10.1145/2661806.2661807

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., et al. (2013). "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge* (Barcelona), 3–10. doi: 10.1145/2512530.2512533

Wen, L., Li, X., Guo, G., and Zhu, Y. (2015). Automated depression diagnosis based on facial dynamic analysis and sparse coding. *IEEE Trans. Inform. Forens. Secur.* 10, 1432–1441. doi: 10.1109/TIFS.2015.2414392

World Health Organization (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*. doi: 10.48550/arXiv.1411.7923

Zhang, W., Shan, S., Gao, W., Chen, X., and Zhang, H. (2005). "Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Beijing), 786–791. doi: 10.1109/ICCV.2005.147

Zhou, X., Jin, K., Shang, Y., and Guo, G. (2020). Visually interpretable representation learning for depression recognition from facial images. *IEEE Trans. Affect. Comput.* 11, 542–552. doi: 10.1109/TAFFC.2018.2828819

Zhu, Y., Shang, Y., Shao, Z., and Guo, G. (2017). Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Trans. Affect. Comput.* 9, 578–584. doi: 10.1109/TAFFC.2017.2650899

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 02 June 2023
DOI 10.3389/fnins.2023.1202985



OPEN ACCESS

EDITED BY
Teng Li,
Anhui University, China

REVIEWED BY
Chongke Bi,
Tianjin University, China
Pengyi Hao,
Zhejiang University of Technology, China

*CORRESPONDENCE
Jing Ji
✉ jingji@xidian.edu.cn

RECEIVED 10 April 2023
ACCEPTED 03 May 2023
PUBLISHED 02 June 2023

CITATION
Chen Y, Zhao M, Xu Z, Li K and Ji J (2023) Wafer defect recognition method based on multi-scale feature fusion.
*Front. Neurosci.* 17:1202985.
doi: 10.3389/fnins.2023.1202985



# Wafer defect recognition method based on multi-scale feature fusion

Yu Chen[1], Meng Zhao[1,2], Zhenyu Xu[3], Kaiyue Li[1] and Jing Ji[1,2]*

[1]Research Center for Applied Mechanics, School of Electro-Mechanical Engineering, Xidian University, Xi'an, China, [2]Shaanxi Key Laboratory of Space Extreme Detection, Xi'an, China, [3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China

Wafer defect recognition is an important process of chip manufacturing. As different process flows can lead to different defect types, the correct identification of defect patterns is important for recognizing manufacturing problems and fixing them in good time. To achieve high precision identification of wafer defects and improve the quality and production yield of wafers, this paper proposes a Multi-Feature Fusion Perceptual Network (MFFP-Net) inspired by human visual perception mechanisms. The MFFP-Net can process information at various scales and then aggregate it so that the next stage can abstract features from the different scales simultaneously. The proposed feature fusion module can obtain higher fine-grained and richer features to capture key texture details and avoid important information loss. The final experiments show that MFFP-Net achieves good generalized ability and state-of-the-art results on real-world dataset WM-811K, with an accuracy of 96.71%, this provides an effective way for the chip manufacturing industry to improve the yield rate.

KEYWORDS

**wafer defect, deep learning, recognition, multi-scale feature, denoise**

## 1. Introduction

With the rapid development of technology and society, semiconductor manufacturing has become one of the most essential industries in the world (Chen et al., 2020) and wafer processing is the basis of it (Bengtsson, 1992). Due to the increasing complexity of semiconductor processes and an increase in the number of wafers produced (Chang et al., 2005), the amount of online and offline data required for diagnosis yield conditions has grown exponentially (Liao et al., 2013), with many of these wafers found to be defective on inspection. Wafer fabrication usually requires a series of processes such as photolithography, deposition, ion implantation, diffusion, machine handing, and chemical mechanical planarization (Cheon et al., 2019). Defects in wafer fabrication arise from variations in the manufacturing process, and defects in a single wafer can render the product in question completely ineffective or even discard the entire batch, so it is important to detect defects and improve the yield. But defects in wafer diagrams have a high tendency to derive necessary information about specific manufacturing process problems from different defect diagrams (Chen and Liu, 2000). Typical spatial patterns in Wafer Maps (WMs) consist of edge-ring, center, scratch, donut, and near-full, etc. (Wang et al., 2019) A center often arises due to problems in the thin film deposition, a ring is due to problems in the etching step, a scratch is a result of machine handing problems (Wang and Bensmial, 2006) and particle-type defects can be fixed by cleaning the surface with an air blower (Cheon et al., 2019). As the process issue happens, engineers can analyze the defective type of wafers to identify the root

causes of the problem and reduce the loss caused by excursion (Chien et al., 2007) as soon as possible. Since all tasks of improving yield require engineers to analyze and process large amounts of data, defect pattern recognition is usually performed through statistical data analysis (Chen and Liu, 2000). Cunningham and MacKinnon (2002) divided the common visual defect metrology into three types.

1. Quadrat statistics: the defect distribution on the wafer is analyzed to predict the yield model, such as by using the conventional Poisson model and Murphy model (Berglund, 1999). Many models (Collica, 1990; Weber et al., 1995; Nurani et al., 1998; Wang et al., 2002) have been based on this type of defect metrology statistics, but this type of defect metrological method has ignored spatial pattern and defect clustering phenomena (Chen and Liu, 2000), and when the data of the wafer does not meet the hypothetical assumptions, it does not work well.
2. Cluster statistics: wafer defects are often determined by defect coordinates, when one or more wafer defects are defined, they can be classified according to the characteristics of the coordinates. This type of method seeks clusters with high defect density and ignores information about the signatures of clusters, such as the shape and size, etc.
3. Spatial pattern recognition: besides defect clusters, the spatial pattern of the defects usually provides a good approach to wafer problem solving. Ken et al. (2002) outline that special shapes appearing on the defect map pattern may come from the machine or process, according to different map patterns, then can find out the root of problems.

Accurate and efficient wafer defect detection technology can identify production process problems and make adjustments to the production process in a timely manner, thus improving the quality and yield of production wafers. To address the problem of wafer defect detection and identification, operators have traditionally visually inspected defects and classified and identified them according to predetermined methods. However, this approach involves a great deal of effort and costs being invested in pre-training defect inspection and the classification of operators (Chen and Liu, 2000). Due to the influence of human factors, the results identified by different operators are different even for the same type of defect (Weber et al., 1995). Therefore, to save costs and improve accuracy, researchers have conducted a series of studies. In the classification of technology and automatic detection of semiconductor manufacturing, frequency domain filtering using optical methods, laser irradiation scanning, and various digital image processing techniques are applied to wafer surface image detection and mostly employed by charge coupled device cameras (Qu, 2002). Most automatic inspection systems scan the wafer surface to collect the coordinates of areas where defects may exist, then place a camera at the center of the coordinates to take pictures, before automatically performing defect detection. Due to the microcosmic nature of the scanning electron microscope sensing field, it is difficult to analyze and detect the surface characteristics of the whole wafer, and the classification accuracy is poor (Cheon et al., 2019), meaning manual detection is required to measure the physical parameters of the WMs like location, size, and color later (Lee et al., 2017). Moreover, Auto Detect Camera (ADC) based approaches apply machine learning and image recognition for wafer defect classification and are introduced to reduce labor and manufacturing costs. Knights Technology (Chen and Liu, 2000) proposed a software program named spatial pattern recognition, the core of the software is a signature classifier, which can be used to train models for different batches of wafer defects, but it takes a lot of time and has poor generalization in training new models. Lee and Inc (1996) propose a templates matching algorithm to detect wafer defects, which is based on the supervised learning method, and improves the detection accuracy; however, one weakness of this approach is that it requires a certain amount of the standard templates to be provided, and once the data volume is large, the effect is not so good. Due to the continuous reduction of wafer size (Qiang et al., 2010), the effect of traditional optical detection technology is gradually getting worse.

The rapid popularity of the Convolutional Neural Network (CNN) and its excellent effects have attracted people's attention. CNN consists of three types of layers including convolution layers, pooling layers, and fully connected layers (Saqlain et al., 2020). The convolution layer can automatically extract image features, the pooling layer can extract the main information required to create the image while reducing the number of parameters, and fully connected layers finally classify the input image using the extracted features (Krizhevsky et al., 2012). These three layers can be combined to extract the high-dimensional features of the images. In particular, the CNN models have performed well in classifying image data (Sengupta et al., 2018), and have been introduced into various industries due to their wide application. For example, to cracks in civil infrastructure (Cha et al., 2017) and classify surface defects in steel plates. The semiconductor industry has also tried to introduce CNN to improve the process for defect recognition of spring-wire sockets (Tao et al., 2018). Lee et al. (2017) designed a new CNN structure, which can identify global and invariant features in the sensor signal data, find the multivariable process fault and diagnose the fault source. Currently, deep learning methods have achieved good results in wafer detection, for example, Takeshi (Nakazawa and Kulkarni, 2018) et al. applied eight convolutional networks with activation functions to classify wafers and used simulated WMs to train a model and tested the performance on 1,191 real WMs. Cheon et al. (2019) proposed a CNN-based automatic defect classification method that can extract features from WMs and accurately classify known defect classes. The datasets used by all these studies were very small and cannot therefore fully represent the actual situation of production. CNN models can achieve higher training accuracy in the presence of bigger datasets (Najafabadi et al., 2015). Saqlain et al. (2020) proposed a deep layered CNN-based wafer defect identification (CNN-WDI) model, before training and testing the model on a real wafer dataset called WM-811K, a large dataset that consisted of eight different wafer defects and 811,457 wafer maps in total. Yu and Lu (2016) proposed a manifold learning-based wafer map defect detection and recognition system and their experimental results from WM-811K verified that the overall accuracy was 90.5%.

Noise is common in the wafer maps and can make an impact on the recognition effect, denoising can effectively preserve the defect type of the wafer and improve the accuracy. Thus, image denoising is a key step in the defect recognition procedure. Wang et al. (2006) used a spatial filter that compares the defect densities in each die of the wafer. On the other hand, noise is also a test of model robustness. When the robustness of the model is good, the impact of noise on the performance will be small. Multiscale analysis is a technique in pattern recognition and image processing that analyzes an image or pattern at

various scales (Li et al., 2016a,b,c). This benefits multiple applications, such as object identification, image categorization, and feature extraction, which can help understand phenomena or processes that occur over a range of scales and for extracting features (Ataky et al., 2022). Eseholi et al. (2020) decomposed the surface profile into three multiscale filtered image types: Low-pass, Band-pass, and High-pass filtered versions, respectively, by using a Gaussian Filter. Compared to conventional roughness descriptors, their method increased surface discrimination from 65 to 81%. The term "scale" has had many meanings in metrological studies. Scale can refer to the ratio of lengths on measurement renderings to the actual lengths on the actual surface (Brown et al., 2018). In this paper, multi-scale means that the image is processed by convolution to obtain feature maps with different channel numbers. We call these feature maps with different channel numbers "multi-scale." Through comprehensive utilization of these multi-scales, we call them "Multi-Scale Feature Fusion." To extract patterns from observable measurements we need to be able to define and identify stable features in observable measurements (Scott, 2004), convolution can extract stable abstract features of objects, so we use a convolution neural network to extract multi-scale information.

The contributions of this paper are as follows:

1. A Multi-Feature Fusion module (MFF) is proposed based on the attributes of wafers and can combine different fine-grained features, capturing the key information from local and global regions, which can improve the robustness of wafer defect recognition.
2. A Multi-Feature Fusion Perceptual Network (MFFP-Net) is designed to integrate information from different dimensions, and the next stage can abstract features from the different scales simultaneously. Therefore, the MFFP-Net can extract more information to achieve high precision wafer recognition. It also effectively resists the interference of noise.
3. Comprehensive experiments demonstrate that the proposed method can obtain good results for identifying wafer map defect patterns, which has a recognition accuracy of 96.71% and achieves state-of-the-art wafer recognition performance in WM-811K.

# 2. Methods

In this section, we first introduce the overall structure of MFFP-Net and then introduce the composition of MFF in detail.

## 2.1. Overview

We propose a Multi-Feature Fusion Perceptual Network (MFFP-Net) to address the recognition of wafer defects. As shown in Figure 1, MFFP-Net consists of four convolution layers and two branches. The network takes the original wafer defect map as input. The direction of the arrow represents the operation direction of the feature layer in turn. First, the Conv1 ~ Conv4 layer serves as the feature pre-extractor to output $28 \times 28 \times 128$ feature maps. Then, the feature maps are input into Multi-scale Branch and Global Branch to extract different perceptual field features. The Multi-scale Branch consists of three MFF modules. The Global Branch is composed of a Max Pooling layer, Conv5, and Conv6. Finally, we fused the feature maps with 256, 512, 512, and 1,024 channels to predict the wafer defect type.

## 2.2. Backbone

The Conv1 ~ Conv4 layers serve as the backbone. Then the feature maps are input into two branches to extract different perceptual field features. The Multi-scale Branch gets fine-grained features through MFFs, and the Global Branch gets features of higher dimension through further convolution operation. Finally, the recognition results are obtained by fusing the feature and decision level. GAP denotes global average pooling layers and is an element wise addition. We use the traditional convolution neural network, the most basic compositions of the neural network are convolution operation, Batch Normalization, Max pooling, and Global Average Pooling (GAP). The details of Conv1 ~ Conv4 are shown in Figure 2.

Conv1 ~ Conv4 are composed of $3 \times 3$ convolution operation, Batch Normalization (BN), and Swish activation function. However, the difference in this approach relates to the convolution operation parameters: including the stride operation, the input channel, and the output channel of each convolution. When the wafer image is input into the network, it will pass through Conv1 ~ Conv4 in turn. Finally, the shallow features are output by Conv4., and Conv5 and Conv6 both consist of convolution operation, BN, and Swish activation function. Conv5 uses $3 \times 3$ convolution and the input and output channels are 128. Conv6 uses $1 \times 1$ convolution to make the network deeper and the input and output channels are 128 and 256, respectively.

## 2.3. MFF module

By controlling the longest gradient path, the deeper network can learn and converge effectively (Wang et al., 2022). The MFF module aims to obtain higher fine-grained and richer features, it uses expand and merge channels to achieve the ability to continuously enhance the learning ability of the network. As shown in Figure 3, the MFF module is composed of three branches that are composed of one, two, and four convolutions, respectively. The outputs obtained from the three branches are joined together according to dimensions. The final output is obtained after the Max Pooling layer to reduce parameters. The module follows a philosophy that visual information should be processed at various scales and then aggregated so that the next stage can abstract features from the different scales simultaneously.

The MFF module can process information at various scales and then aggregate it so that the next stage can abstract features from the different scales simultaneously. $F$ denotes convolutional layers, and $y$ denotes output feature maps. The arrow points to the directions in which the feature map passes.

The MFF module can be formulated by Equations (1–4).

$$y_1 = F_1(x) \tag{1}$$

$$y_2 = F_3^1(y_1) \tag{2}$$

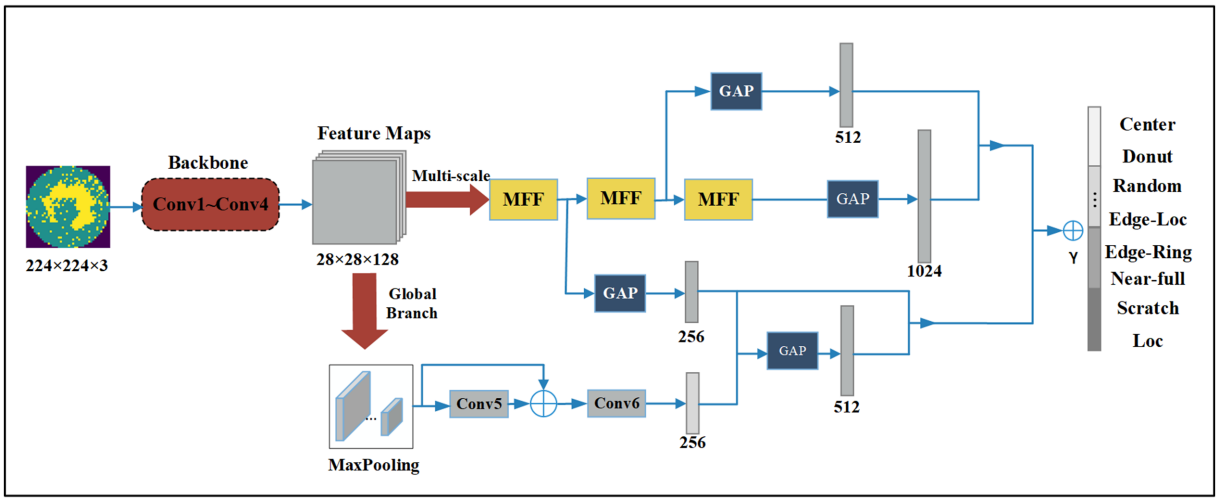



FIGURE 1
Structure of the proposed method.

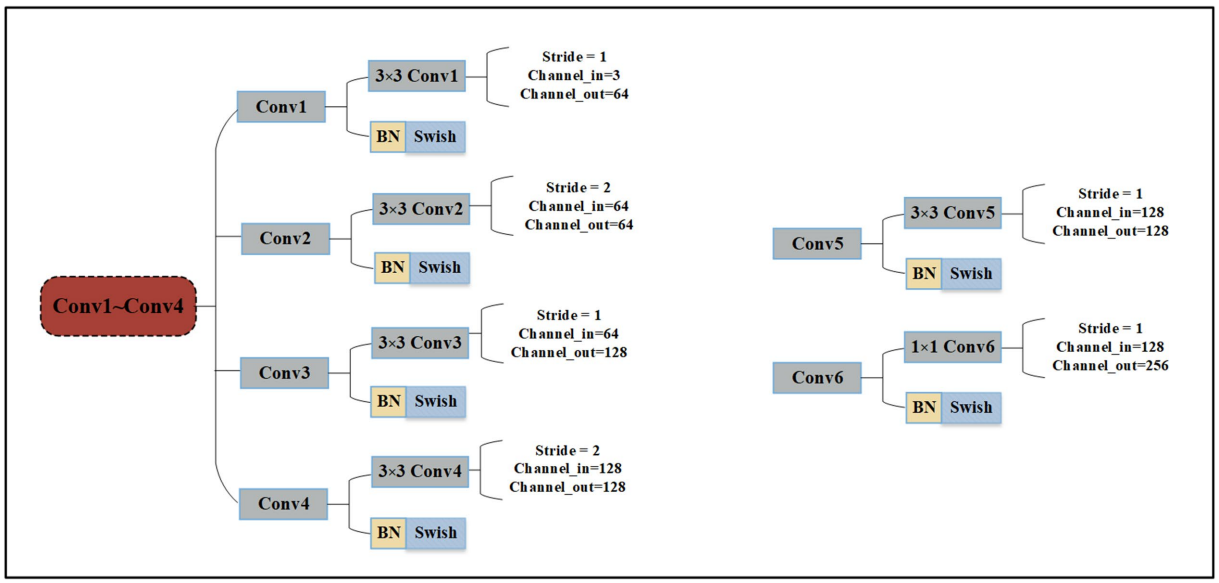



FIGURE 2
The detailed parameters of Conv1~Conv4, Conv5 and Conv6.

$$y_3 = F_3^3\left[F_3^2\left(y_2\right)\right] \tag{3}$$

$$F_{output} = MaxPool\left[Concat\left(y_1, y_2, y_3\right)\right] \tag{4}$$

Where $x$ is the input of the MMF module, $F_{output}$ is the output of the MFF module.

*Concat* is the tensor splicing function, the dimension of the tensor can be specified for splicing.

The details of the convolution layer are shown in Figure 4.

If the number of feature map channels input to MFF is $C$. The output channels of $F_{1x1}$ become $C/2$. In the convolution layer $F^1_{3x3}$ next to $F_{1x1}$ the input and output channels are both $C/2$. The input and output channels of $F^2_{3x3}$ are $C/2$ and $C$, respectively. The input and output channels of $F^3_{3x3}$ are $C$. Finally, the feature map with a channel number of $2C$ is obtained.

The MMF takes the feature map obtained through the convolution layers as the input. We assume the depth (the number of channels) of the feature map is $C$.

The model change in depth of the feature map through MMF is shown in Figure 5.

First, a $1 \times 1$ convolution layer is executed after the input to adjust the number of channel dimensions and make the depth $C/2$. The introduction of $1 \times 1$ convolution enables the combination of channels and the interaction of information between channels.

Second, after 1 × 1 convolution, the number of channels in the feature map halves, and then we use 3 × 3 convolution to further extract high-dimensional features. Third, based on the second step, after two convolution operations, the network becomes deeper and the number of channels becomes *C*. Fourth, the feature maps obtained in the first, second, and third steps are joined together according to the channel direction, the number of channels is 2*C*, and more fine-grained features are obtained. Last, feature maps with 2*C* channels pass through the convolution layer to achieve the final output.

The number of channels in the feature map is *C*. Through the first and second branches, the channels of input halves, then combine according to the channel direction, and the number of channels is still *C*. Through the third branch, the depth remains unchanged. Finally, splice feature maps with channel number *C* together and double the number of channels. *C* denotes the channels of the input. The arrow indicates the direction of channel number changes.

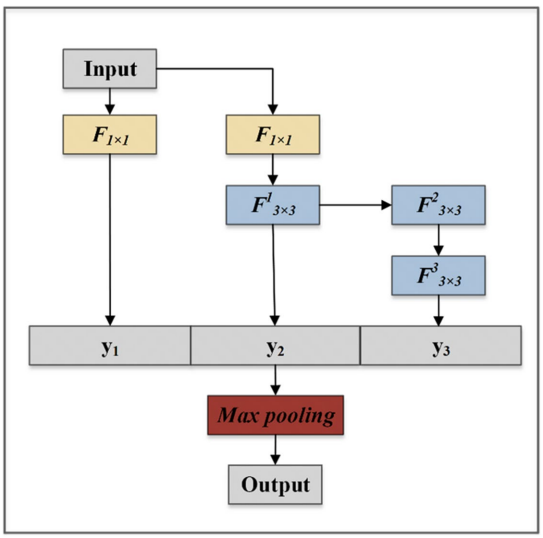



FIGURE 3
The structure of MFF module.

## 2.4. Auxiliary classifier and lead head

Deep supervision is a technique that is often used in training deep networks. We add auxiliary head in the middle layers of the network, auxiliary head is conducted and marked as A, B, C, and D, as shown in Figure 6. The shallow network weights with assistant loss as the guide. In this paper, we refer to the classification header responsible for the final output as lead head and the head used to assist training is called auxiliary classifier. Auxiliary classifiers located at different depth levels will learn different information, and the learning ability of an auxiliary classifier is not strong as a lead head. In order to avoid losing the information that needs to learn and combine useful information together, it is crucial to find out how to assign weights to auxiliary classifiers. We will discuss the details of assigning auxiliary classifier weights in the part of Ablation Experiments. As for the output of lead head, we filter the high precision results from the high recall as the final output.

# 3. Experiments

In this section, we first introduce two datasets and their characteristics. We then explain the details of the experimental implementation. Thirdly, we adjust the parameters of the experiment to obtain the best results and visualize the effect of model recognition. Finally, we analyze the error of the experimental results.

## 3.1. Dataset

To compare our results with previous studies and verify the effectiveness of the method outlined in the present study,

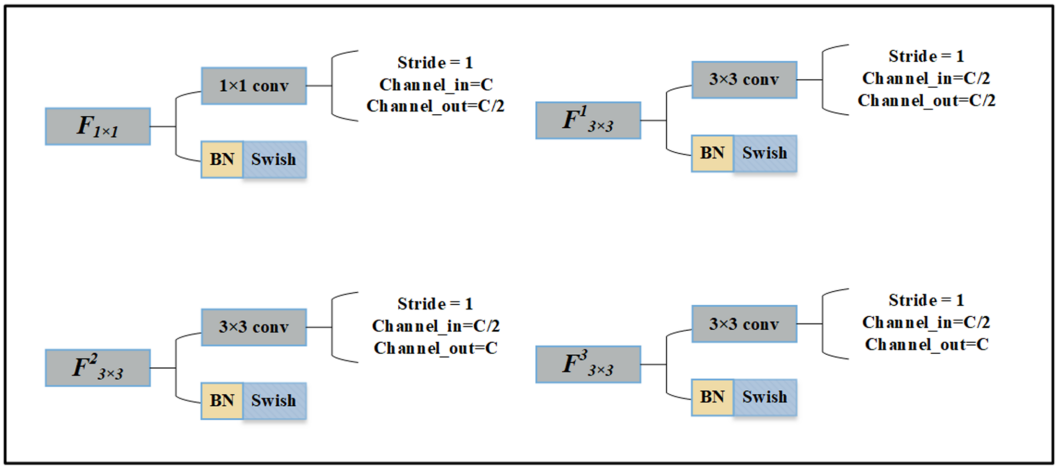



FIGURE 4
The details of 3×3 convolution layer and 1×1 convolution layer.

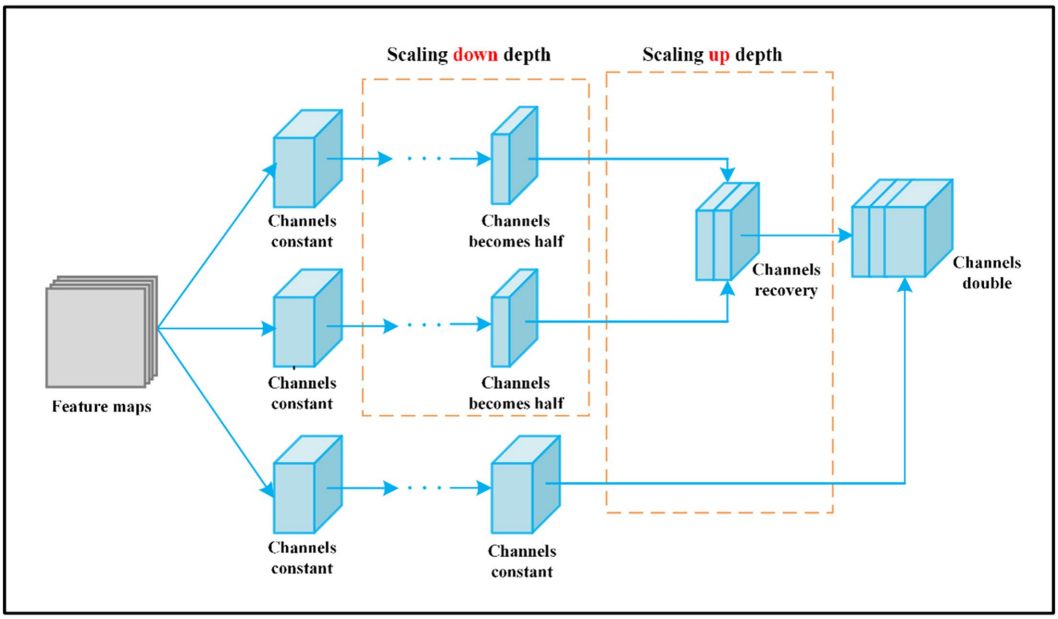



FIGURE 5
The change of channels number through MMF feature map.

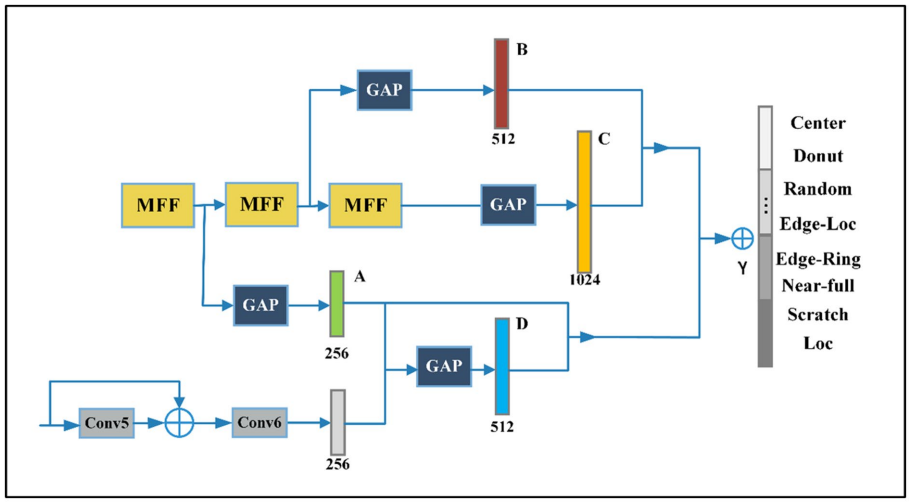



FIGURE 6
The proposed model contains four auxiliary classifiers.

we performed experiments on real-world wafer datasets WM-811K (MIR Corpora, 2015; Wu et al., 2015). The WM-811K dataset is the largest publicly available wafer data, consisting of 811,457 wafer maps collected from 46,293 different lots in real-world fabrication. This dataset contains eight different and labeled wafer failure patterns, a total of 24,653 wafer maps, the rest were unlabeled and defect-free wafer maps. Figure 7 shows the sample wafer maps from each defect type including Center, Donut, Edge-Ring, Scratch, Near-full, Loc, Edge-Ring, and Random. The yellow part represents the defect, and the green part represents the defect-free part. Domain experts were recruited to annotate the pattern type of the wafer maps in the WM-811K dataset. We also found a data set about wafers on (Karen and Andrew, 2015) (wafer-Kaggle), shown in Figure 8.

We used 25,519 wafer defect maps labeled in the WM-811K dataset to verify the performance of the model. The numbers of eight types are 4,294, 555, 5,189, 9,680, 3,593, 149, 866, and 1,193, respectively, and the proportion was 25:3:30:56:20:1:5:7. The eight wafer defect types in this data set were shown to be seriously imbalanced. The main problem in image resolution is noise and the wafer maps in the WM-811K dataset contain serious noise, as shown in Figure 9. If the robustness of the model is poor, the noise will greatly affect the performance of the model.

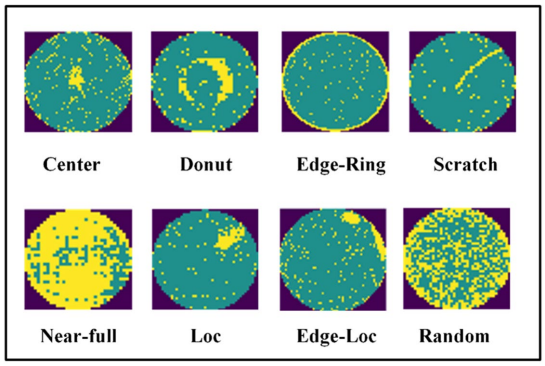



FIGURE 7
Eight different wafer defect types.

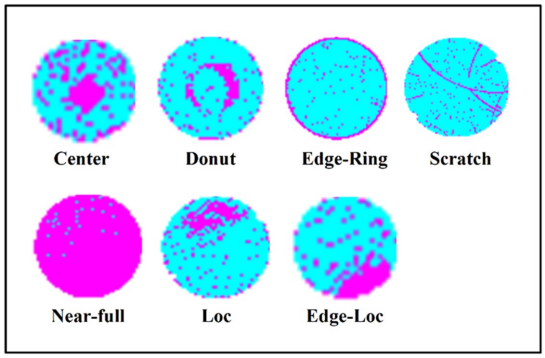



FIGURE 8
There are seven different wafer defect types in wafer-kaggle.

TABLE 1 The impact of four different auxiliary classifiers on wafer classification accuracy.

| (A, B) | (C, D) | Precision |
|---|---|---|
| (1, 0) | (0, 0) | 84.58% |
| (0, 1) | (0, 0) | 92.21% |
| (0, 0) | (1, 0) | 88.02% |
| (0, 0) | (0, 1) | 79.60% |
| (1, 1) | (1, 1) | 94.56% |

TABLE 2 The impact of auxiliary classifier D on wafer classification accuracy.

| (A, B) | (C, D) | Precision (%) |
|---|---|---|
| (1, 1) | (1, 1.3) | 87.40 |
| (1, 1) | (1, 0.7) | 88.49 |
| (1, 1) | (1, 0.5) | 94.56 |
| (1, 1) | (1, 0.3) | 95.73 |
| (1, 1) | (1, 0.1) | 95.65 |
| (1, 1) | (1, 0) | 95.07 |

TABLE 3 The impact of auxiliary classifier B on wafer classification accuracy.

| (A, B) | (C, D) | Precision (%) |
|---|---|---|
| (1, 1.2) | (1, 0.3) | 95.95 |
| (1, 1.4) | (1, 0.3) | 96.71 |
| (1, 1.6) | (1, 0.3) | 96.53 |
| (1, 1.8) | (1, 0.3) | 95.45 |

## 3.2. Implementation details

We divided WM-811K randomly into a training set, validation set, and test set in the ratio of 8:1:1. For the training set, we first used random clipping wafer maps, as part of which the pixel size became 224 × 224, then a random horizontal flip. For the test set, we changed the wafer map pixels to 256 × 256, then it became 224 × 224 through the center crop. The model was developed by using PyTorch. An NVIDIA 3080 GPU with 16 GB memory was engaged to accelerate the calculation. The learning rate was set to a constant of 0.0001, the weight decay coefficient was 0.05, and the minibatch size is set to 32. We train the model for a total of 100 epochs, during the training, we use Cosine Annealing with a period of 32. The number of parameters of the proposed net is 48.09 M.

## 3.3. Result and analysis

### 3.3.1. Ablation experiments

The features produced by the layers in the middle of the network are very discriminative, even low dimensional embeddings might contain a large amount of information. To study the impact of auxiliary classifiers connected to the middle layer on classification results, the experiment with only one auxiliary classifier is conducted and marked as A, B, C, and D, as shown in Figure 6.

The impact of four different auxiliary classifiers is shown in Table 1. When there is only one auxiliary classifier, auxiliary classifier B achieved the best accuracy of 92.21%, and auxiliary classifier D achieved the lowest accuracy of 79.60%. When we use four auxiliary classifiers at the same time and give them the same weight, the accuracy is higher than when using only one auxiliary classifier, at 94.56%. Combining the features from the different scales could improve recognition accuracy.

As shown in Table 2, when only auxiliary classifier D is used, the recognition accuracy of the model is far lower than that of other auxiliary classifiers. To study the influence of auxiliary classifier D on classification accuracy, we give different weights to D. When the weights of D are set as 1.3, 0.7, 0.5, 0.3, 0.1, and 0, respectively. The accuracy of the model is shown in Table 2, which indicates that when the A, B, C, and D ratios are 1:1:1:0.3, the model achieves the highest wafer recognition accuracy of 95.73%.

As shown in Table 3, when only the auxiliary classifier B is used, the recognition accuracy of the model is far higher than that of other auxiliary classifiers. We fixed the weight of the auxiliary classifier D to 0.3, then set different weights for B. When the weights of B are set as 1.2, 1.4, 1.6, and 1.8 respectively, the accuracy of the model is shown in Table 3. It is indicated that when the A, B, C, and D ratios are 1:1.4:1:0.3, the model achieves the best performance.

TABLE 4 Comparison to other methods tested in the WM-811K dataset.

| Model | Accuracy (%) |
|---|---|
| Ours | 96.71 |
| CNN-WDI (Saqlain et al., 2020) | 96.20 |
| SVE (Saqlain et al., 2019) | 95.86 |
| YOLOV4 (Shinde et al., 2022) | 95.70 |
| WMFPR (Wu et al., 2015) | 94.63 |
| YOLOV3 (Shinde et al., 2022) | 94.40 |
| CVAE (Ho et al., 2021) | 93.60 |
| SCSDAE (Yu et al., 2019) | 92.63 |
| Label reconstruction (Park and Jang, 2021) | 91.20 |
| DTE-FPR (Piao et al., 2018) | 90.50 |

TABLE 5 Comparison to other models tested in the WM-811K dataset.

| Model | Accuracy (%) |
|---|---|
| Ours | 96.71 |
| ResNet50 (He et al., 2015) | 95.23 |
| VGG16 (Karen and Andrew, 2015) | 95.20 |
| MobileNet (Andrew et al., 2017) | 93.20 |
| GoogleNet (Christian et al., 2015) | 93.82 |
| ResNet34 (He et al., 2015) | 92.64 |
| ResNet101 (He et al., 2015) | 91.04 |

TABLE 6 The performance of the proposed model in WM-811K.

| Defect type | Precision (%) | TPR (%) | TNR (%) |
|---|---|---|---|
| Loc | 95.3 | 89.4 | 99.3 |
| Center | 96.8 | 99.5 | 99.3 |
| Donut | 88.5 | 96.7 | 99.7 |
| Random | 100.0 | 88.5 | 100.0 |
| Scratch | 95.6 | 90.0 | 99.8 |
| Near-full | 75.0 | 100.0 | 99.8 |
| Edge-loc | 94.8 | 97.5 | 98.6 |
| Edge-Ring | 99.1 | 99.3 | 99.4 |

### 3.3.2. Metrics

The methods shown in Table 4 are the results of a test run on the WM-811K dataset. As shown in Table 4, the proposed method is the best in terms of performance. The proposed method is not only simple to process but can also achieve good results.

We used the same settings as the proposed model to test some common classified networks. As shown in Table 5, our model is 1.48% higher than that ranked second place, ResNet50.

We also used the True Positive Rate (TPR) and True Negative Rate (TNR) as metrics to measure the performance of the model. TPR is the proportion of positive examples predicted by the model to all real positive examples. TNR is the proportion of negative examples predicted by the model to all real negative examples. TPR and TNR are calculated by Equations (5, 6), respectively.

TABLE 7 The performance of the proposed model in wafer-Kaggle.

| Defect type | Precision (%) | Recall (%) | Specificity (%) |
|---|---|---|---|
| Loc | 88.5 | 85.8 | 98.8 |
| Center | 97.2 | 99.7 | 99.3 |
| Donut | 93.0 | 97.6 | 99.8 |
| Scratch | 87.5 | 98.0 | 99.6 |
| Near-full | 100.0 | 100.0 | 100.0 |
| Edge-loc | 93.0 | 87.2 | 98.9 |
| Edge-Ring | 99.1 | 99.4 | 99.1 |

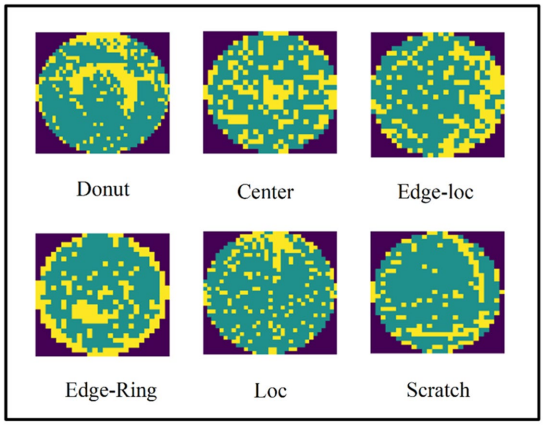



FIGURE 9
The wafer maps in WM-811K contains serious noise.

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$TNR = \frac{TN}{TN + FP} \tag{6}$$

TP is the number of positive examples correctly classified by the model.

FN is the number of positive examples incorrectly classified by the model.

FP is the number of negative examples incorrectly classified by the model.

TN is the number of negative examples correctly classified by the model.

As shown in Table 6, model performance in WM-811K, for other types except for Near-full, the recognition precision is above 88%. For Random and Edge-Ring, the precision is more than 99%. The reason for the low recognition accuracy of Near-full will be discussed in the error analysis. The specificity for all kinds of wafers exceeds 99%.

We also tested the proposed model in wafer-Kaggle, as shown in Table 7, the recognition accuracy of each type of wafer was more than 87%. The recall was more than 87% and the specificity exceeded 98%. The precision of Near-full was 100%.

The confusion matrix of the WM-811K and wafer-Kaggle dataset are shown in Figure 10.

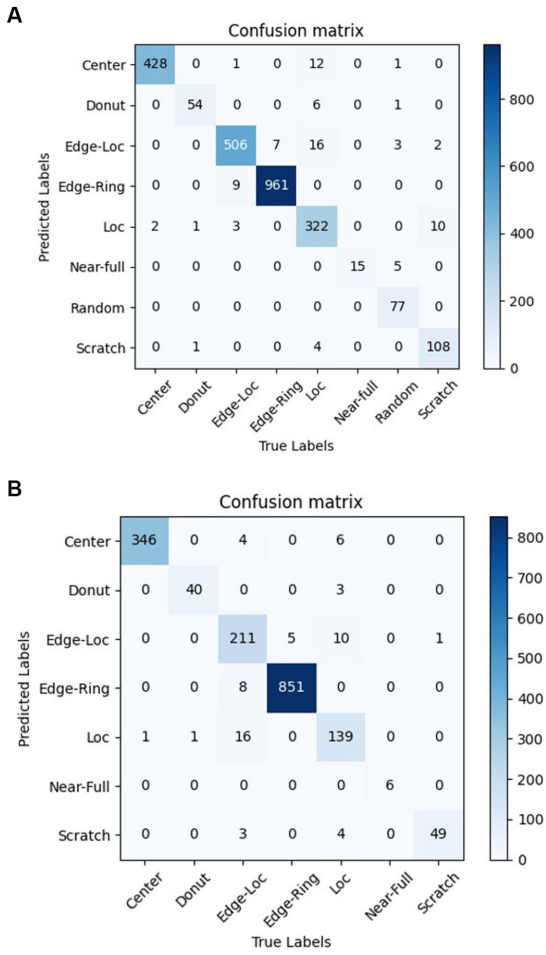



FIGURE 10
Confusion matrix of the WM-811K and wafer-Kaggle datatset. **(A)** Confusion matrix of the WM-811K. **(B)** Confusion matrix of the wafer-Kaggle datatset.

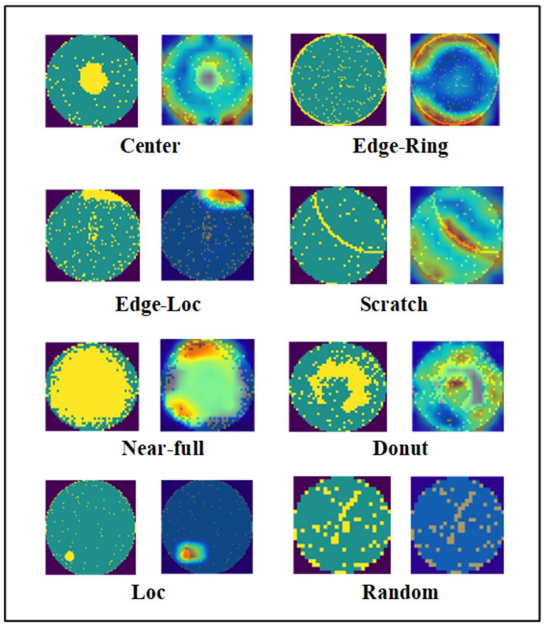



FIGURE 11
Attention maps of eight different wafers in WM-811K.

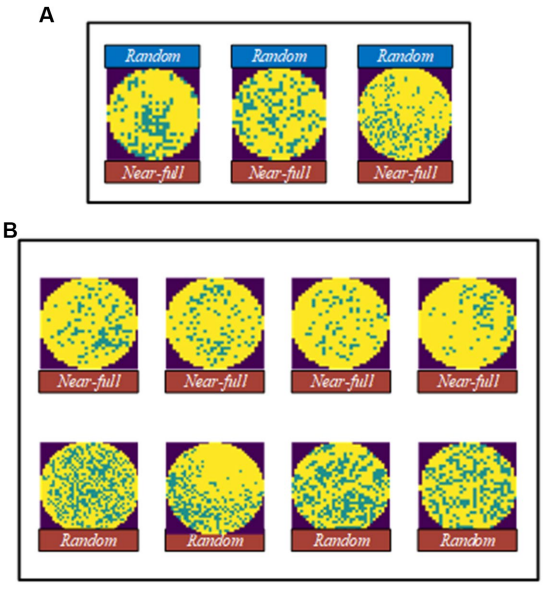



FIGURE 12
Error analysis. **(A)** Water images with random defect in the WM-811K dataset where MFFP-Net failed to predict the correct defect categories. **(B)** Comparison between near-full and random.

### 3.3.3. Visualization

To further investigate the performance of the proposed model in more detail, we use gradient weighted class activation mapping (Grad-CAM) (Du and Martinez, 2011) to visualize it. As shown in Figure 11, when the area more brightly colored, the model pays more attention to it. For different types of wafer defects, the proposed model can capture their unique features accurately and not be affected by noise.

Figure 12 shows wafer images with Random defects in the WM-811K dataset where MFFP-Net failed to predict the correct defect categories. Although MFFP-Net is robust to wafer maps with noise, great similarity between Random and Near-full leads to recognition errors. The solution to the problem is to supplement more information about these two defect types, such as using multiple data enhancement methods to increase differences.

## 4. Conclusion and discussion

This paper proposes a Multi-Feature Fusion Perceptual Network (MFFP-Net) inspired by the attributes of the wafers and human visual perception mechanism to recognize wafer defects. We designed a multi-feature fusion module through which information can be processed at various scales and then aggregated so that the next stage can abstract features from the different scales simultaneously. The final experiment and comparison with existing methods showed that the proposed method can effectively eliminate the influence of noise and achieve high precision recognition. DNA

computing is a novel intelligent method that can be applied to remote sensing image classification (Jiao et al., 2010) and sodar data classification (Ray and Mondal, 2011). Due to DNA computing having the characteristics of massive parallel computing, in future work, we plan to explore using it to classify wafers and compared it with the method based on neural networks in performance.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

YC: data curation and writing—original draft. ZX: formal analysis. MZ and JJ: project administration. KL and JJ: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

## Funding

This work was supported by the Natural Science Foundation of Shaanxi Province (2023-JC-YB-49).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Andrew, G. H., Menglong, Z., Bo, C., Dmitry, K., Weijun, W., Tobias, W., et al. (2017). MobileNets: Efficient convolutional neural networks for Mobile Vision applications. *arXiv*. doi: 10.48550/arXiv.1704.04861

Ataky, S., Saqui, D., and Koerich, A. (2022). Multiscale analysis for improving texture classification. *arXiv*. doi: 10.48550/arXiv.2204.09841

Bengtsson, S. (1992). Semiconductor wafer bonding: a review of interfacial properties and applications. *J. Electron. Mater.* 21, 841–862.

Berglund, N. (1999). A unified yield model incorporating both defect and parametric effects. *IEEE Trans. Semicond. Manuf.* 9, 447–454. doi: 10.1109/66.536115

Brown, A., Hansen, H., and Xiang, J. (2018). Multiscale analyses and characterizations of surface topographies. *CIRP Ann.* 67, 839–862. doi: 10.1016/j.cirp.2018.06.001

Cha, Y., Choi, W., and Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. *Comput. Aid. Civil Inf.* 32, 361–378. doi: 10.1111/mice.12263

Chang, C., Lin, S., and Jeng, M. (2005). *Two-layer competitive Hopfield neural network for wafer defect detection*. Proceedings 2005 IEEE Networking, Sensing and Control, AZ, 19–22.

Chen, X., Chen, J., and Han, X. (2020). A light-weighted CNN model for wafer structural defect detection. *IEEE Access* 8, 24006–24018. doi: 10.1109/ACCESS.2020.2970461

Chen, F., and Liu, S. (2000). A neural-network approach to recognize defect spatial pattern in semiconductor fabrication. *IEEE Trans. Semicond. Manuf.* 13, 366–373. doi: 10.1109/66.857947

Cheon, S., Lee, H., Kim, C., and Lee, S. (2019). Convolutional neural network for wafer surface defect classification and the detection of unknown defect class. *IEEE Trans. Semicond. Manuf.* 32, 163–170. doi: 10.1109/TSM.2019.2902657

Chien, C., Wang, W., and Cheng, J. (2007). Data Mining for Yield Enhancement in semiconductor manufacturing and an empirical study. *Expert Syst. Appl.* 33, 192–198. doi: 10.1016/j.eswa.2006.04.014

Christian, S., Wei, L., Jia, Y., and Pierre, S. (2015). *Going deeper with convolutions*. IEEE conference on computer Vision and pattern recognition, USA.

Collica, S. (1990). *The physical mechanisms of defect clustering and its correlation to yield model parameters for yield improvement*. IEEE/SEMI Conference on Advanced Semiconductor Manufacturing Workshop, USA, 11–12.

Cunningham, S., and MacKinnon, S. (2002). Statistical methods for visual defect metrology. *IEEE Trans. Semicond. Manuf.* 11, 48–53. doi: 10.1109/66.661284

Du, S., and Martinez, A. (2011). The resolution of facial expressions of emotion. *J. Vision* 11:24. doi: 10.1167/11.13.24

Eseholi, T., Coudoux, F., and Bigerelle, M. (2020). A multiscale topographical analysis based on morphological information: the HEVC multiscale decomposition. *Materials* 13:5582. doi: 10.3390/ma13235582

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv*. doi: 10.48550/arXiv.1512.03385

Ho, S., Erdenebileg, B., and Wan, C. (2021). Unsupervised pre-training of imbalanced data for identification of wafer map defect patterns. *IEEE Access.* 99:1. doi: 10.1109/ACCESS.2021.3068378

Jiao, H., Zhong, Y., and Zhang, J. (2010). Classification of hyperspectral remote sensing data based on DNA computing. *J. Remote Sens.* 14, 865–878. doi: 10.3724/SP.J.1011.2010.01138

Karen, S., and Andrew, Z. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv*. doi: 10.48550/arXiv.1409.1556

Ken, R., Brain, S., and Neil, H. (2002). *Using full wafer defect maps as process signatures to monitor and control yield*. 1991 Proceedings IEEE/SEMI International Semiconductor Manufacturing Science Symposium, USA.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). *ImageNet classification with deep convolutional neural networks*. Advances in neural information processing systems, USA.

Lee, F., and Inc, M. (1996). *Advanced yield enhancement: computer-based spatial pattern analysis*. IEEE/SEMI 1996 advanced semiconductor manufacturing conference and workshop, USA, 12–14.

Lee, H., Kim, Y., and Kim, K. (2017). A deep learning model for robust wafer fault monitoring with sensor measurement noise. *IEEE Trans. Semicond. Manuf.* 30, 23–31. doi: 10.1109/TSM.2016.2628865

Liao, C., Huang, Y., and Chen, C. (2013). Similarity searching for defective wafer bin maps in semiconductor manufacturing. *IEEE Trans. Autom. Sci Eng.* 11, 953–960. doi: 10.1109/TASE.2013.2277603

Li, T., Meng, Z., Ni, B., Shen, J., and Wang, M. (2016a). Robust geometric ℓp-norm feature pooling for image classification and action recognition. *Image and Vision Computing*. 55, 64–76.

Li, T., Cheng, B., Ni, B., Liu, G., and Yan, S. (2016b). Multitask low-rank affinity graph for image segmentation and image annotation. *ACM Transactions on Intelligent Systems and Technology*. 7, 1–18.

Li, T., Mei, T., Yan, S., Kweon, I. S., and Lee, C. (2016c). Contextual decomposition of multi-label images. *IEEE Conference on Computer Vision and Pattern Recognition*. 2270–2277.

MIR Corpora. (2015). MIR-WM811K: Dataset for wafer map failure pattern recognition. Available at: http://mirlab.org/dataset/public/ (Accessed January 24, 2023).

Najafabadi, M., Villanustre, F., Khoshgoftaar, T., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *J. Big Data* 2, 1–21. doi: 10.1186/s40537-014-0007-7

Nakazawa, T., and Kulkarni, D. (2018). Wafer map defect pattern classification and image retrieval using convolutional neural network. *IEEE Trans. Semicond. Manuf.* 31, 309–314. doi: 10.1109/TSM.2018.2795466

Nurani, R., Strojwas, A., Maly, W., Ouyang, C., Shindo, W., Akella, R., et al. (1998). In-line yield prediction methodologies using patterned wafer inspection information. *IEEE Trans. Semicond. Manuf.* 11, 40–47. doi: 10.1109/66.661283

Park, S., and Jang, J. (2021). Discriminative feature learning and cluster-based defect label reconstruction for reducing uncertainty in wafer bin map labels. *J. Intell. Manuf.* 32, 251–263. doi: 10.1007/s10845-020-01571-4

Piao, M., Jin, C., Lee, J., and Byun, J. (2018). Decision tree ensemble-based wafer map failure pattern recognition based on radon transform-based features. *IEEE Trans. Semicond. Manuf.* 31, 250–257. doi: 10.1109/TSM.2018.2806931

Qiang, Z., Li, Z., and Zhou, S. (2010). Statistical detection of defect patterns using Hough transform. *IEEE Trans. Semicond. Manuf.* 23, 370–380. doi: 10.1109/TSM.2010.2048959

Qu, Y. (2002). Wafer defect detection using directional. *Eurasip J. Adv. Sig. Pr.* 9, 47–54. doi: 10.1155/S1687617202204035

Ray, K., and Mondal, M. (2011). Classification of Sodar data by DNA computing. *New Math. Nat.* 07, 413–432. doi: 10.1142/S1793005711002074

Saqlain, M., Abbas, Q., and Lee, J. (2020). A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes. *IEEE Trans. Semicond. Manuf.* 33, 436–444. doi: 10.1109/TSM.2020.2994357

Saqlain, M., Jargalsaikhan, B., and Lee, J. (2019). A voting ensemble classifier for wafer map defect patterns identification in semiconductor manufacturing. *IEEE Trans. Semicond. Manuf.* 32, 171–182. doi: 10.1109/TSM.2019.2904306

Scott, P. (2004). Pattern analysis and metrology: the extraction of stable features from observable measurements. *Proc. Royal Soc. Lond. A* 460, 2845–2864. doi: 10.1098/rspa.2004.1291

Sengupta, A., Ye, Y., and Wang, R. (2018). Going deeper in spiking neural networks: VGG and residual architectures. *arXiv*. doi: 10.48550/arXiv.1802.02627

Shinde, P., Pai, P., and Adiga, S. (2022). Wafer defect localization and classification using deep learning techniques. *IEEE Access.* 10, 39969–39974. doi: 10.1109/ACCESS.2022.3166512

Tao, X., Wang, Z., and Zhang, Z. (2018). Wire defect recognition of spring-wire socket using multitask convolutional neural networks. *IEEE Trans. Comp. Pack. Man.* 8, 689–698. doi: 10.1109/TCPMT.2018.2794540

Wang, C., and Bensmial, H. (2006). Detection and classification of defects patterns on semiconductor wafers. *IIE Trans.* 38, 1059–1068. doi: 10.1080/07408170600733236

Wang, P., Chan, P., Goodner, R., Lee, F., and Ceton, R. (2002). *Development of the yield enhancement system of high-volume 8-inch wafer fab*. Proceedings of International Symposium on Semiconductor Manufacturing, Austin, USA.

Wang, C., Kuo, W., and Bensmail, H. (2006). Detection and classification of defect patterns on semiconductor wafers. *IIE Trans.* 38, 1059–1068. doi: 10.1080/07408170600733236

Wang, Z., Yang, Z., and Zhang, J. (2019). AdaBalGAN: an improved generative adversarial network with imbalanced learning for wafer defective pattern recognition. *IEEE Trans. Semicond. Manuf.* 32, 310–319. doi: 10.1109/TSM.2019.2925361

Wang, C., Yuan, H., Liao, M., and Yeh, I. (2022). Designing network design strategies through gradient path analysis. *arXiv*. doi: 10.48550/arXiv.2211.04800

Weber, C., Moslehi, B., and Dutta, M. (1995). An integrated framework for yield management and defect fault reduction. *IEEE Trans. Semicond. Manuf.* 8, 110–120. doi: 10.1109/66.382274

Wu, M., Jang, J., and Chen, J. (2015). Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Trans. Semicond. Manuf.* 28, 1–12. doi: 10.1109/TSM.2014.2364237

Yu, J., and Lu, X. (2016). Wafer map defect detection and recognition using joint local and nonlocal linear discriminant analysis. *IEEE Trans. Semicond. Manuf.* 29, 33–43. doi: 10.1109/TSM.2015.2497264

Yu, J., Zheng, X., and Liu, J. (2019). Stacked convolutional sparse Denoising auto-encoder for identification of defect patterns in semiconductor wafer map. *Comput. Ind.* 109, 121–133. doi: 10.1016/j.compind.2019.04.015

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 05 July 2023
DOI 10.3389/fnins.2023.1225312



OPEN ACCESS

EDITED BY
Fudong Nian,
Hefei University, China

REVIEWED BY
Jianjun Sun,
Northwestern Polytechnical University, China
Shimeng Yang,
Anhui University, China

*CORRESPONDENCE
Ziliang Ren
✉ renzl@dgut.edu.cn

RECEIVED 19 May 2023
ACCEPTED 12 June 2023
PUBLISHED 05 July 2023

CITATION
Yang H, Ren Z, Yuan H, Xu Z and Zhou J (2023) Contrastive self-supervised representation learning without negative samples for multimodal human action recognition. *Front. Neurosci.* 17:1225312. doi: 10.3389/fnins.2023.1225312



# Contrastive self-supervised representation learning without negative samples for multimodal human action recognition

Huaigang Yang[1], Ziliang Ren[1,2]*, Huaqiang Yuan[1], Zhenyu Xu[2] and Jun Zhou[1]

[1]School of Computer Science and Technology, Dongguan University of Technology, Dongguan, China, [2]CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Action recognition is an important component of human-computer interaction, and multimodal feature representation and learning methods can be used to improve recognition performance due to the interrelation and complementarity between different modalities. However, due to the lack of large-scale labeled samples, the performance of existing ConvNets-based methods are severely constrained. In this paper, a novel and effective multi-modal feature representation and contrastive self-supervised learning framework is proposed to improve the action recognition performance of models and the generalization ability of application scenarios. The proposed recognition framework employs weight sharing between two branches and does not require negative samples, which could effectively learn useful feature representations by using multimodal unlabeled data, e.g., skeleton sequence and inertial measurement unit signal (IMU). The extensive experiments are conducted on two benchmarks: UTD-MHAD and MMAct, and the results show that our proposed recognition framework outperforms both unimodal and multimodal baselines in action retrieval, semi-supervised learning, and zero-shot learning scenarios.

KEYWORDS

**human action recognition, multimodal representation, feature encoder, contrastive self-supervised learning, Transformer**

## 1. Introduction

Automatic recognition framework is a research field that aims to develop systems capable of identifying and classifying human actions or behaviors, which is to enable machines to understand and interpret human behavior, with applications in areas including video surveillance, healthcare, sports analysis, and human-computer interaction (Li et al., 2016a,b; He et al., 2023). Different techniques in real life adopt different types of data inputs, but each modality has its own advantages and limitations ( Sun et al., 2023). To achieve more robust and accurate feature extraction, some approaches improve the performance of models by aggregating the advantages of various modalities in a reasonable manner. Due to the success of deep learning in the past decades, a large number of ConvNets-based frameworks have made impressive achievements in the field of multimodal visual tasks (Grillini et al., 2021; Mughal et al., 2022; Li et al., 2023). However, most of them require many large amounts of labeled data, especially for multimodal data (Zhang et al., 2019, 2020), and labeling the data requires exponentially more time and effort (Li et al., 2009).

Recently, self-supervised representation learning has made significant progress on visual tasks, which is mainly divided into the pre-training and fine-tuning stages (Chen et al., 2020; Grill et al., 2020). In the pre-training stage, it focuses on constructing feature representations of different views by unlabeled samples. In the fine-tuning stage, these representations are used as inputs and fed into a small-scale linear classifier, which requires only a small amount of labeled data. Moreover, contrastive learning is one of the self-supervised learning, where the core concept is to pull the representation distance between positive samples closer and push the distance away from other negative samples. For example, the CMC framework (Tian et al., 2020) is mainly to form positive samples between different data modalities, and consider other different samples as negative sample pairs. Due to the problem of relying too much on negative sample pairs, it is necessary to set a large batch size or a queue for storing negative samples in the learning process, therefore leads to a complex model and is vulnerable to information collapse.

In order to overcome the above shortcomings, inspired by Barlow Twins and VICReg (Zbontar et al., 2021; Bardes et al., 2022), we propose a contrastive self-supervised learning framework for unimodal and multimodal without relying on negative samples. Our proposed method employs multimodal samples as input data, e.g., skeleton sequence and inertial measurement unit signal (IMU). The main contributions of this paper are as follows:

- A unimodal contrastive self-supervised framework is proposed to encode and learn feature representations for multimodal action recognition with skeleton sequence and IMU data.
- The proposed recognition framework is extended to multimodal contrastive self-supervised learning. The model is designed to obtain simple and efficient feature representations without negative samples.

The remainder of this paper is organized as follows. Section 2 presents an overview of related works. In Section 3, we provided a detailed introduction to the proposed method. Section 4 provides experimental results for benchmark datasets and comparisons with state-of-the-art. Section 5 concludes this paper and look forward to future work.

## 2. Related works

In this section, we discuss unimodal, multimodal, and contrastive learning methods for human action recognition from the perspective of input data modality.

### 2.1. Unimodal human action recognition

Unimodal human action recognition primarily focuses on classifying and recognizing actions by using a single modality, including RGB videos, depth and skeleton sequences, IMU data, etc. This field encompasses tasks such as feature extraction, feature representation, and the construction of deep learning models, including convolution neural networks (CNNs) (Andrade-Ambriz et al., 2022; Islam et al., 2022; Xu et al., 2022), recurrent neural networks (RNNs) (Shu et al., 2021; Shen and Ding, 2022; Wang et al., 2022), graph convolution networks (GCNs) (Cheng et al., 2020; Chi et al., 2022; Feng et al., 2022; Tu et al., 2022) and Transformer models (Chen and Ho, 2022; Mazzia et al., 2022; Ahn et al., 2023).

Since the skeleton sequence would not be sensitive to viewpoint variation and circumstance disturbance, there are numerous skeleton-based methods is developed for human action recognition. In CNN-based methods, Li et al. (2018) proposed an end-to-end convolutional co-occurrence feature learning framework from the perspectives of intra-frame representation and inter-frame representation of skeleton temporal evolutions, which introduced a global spatial aggregation method and discarded the local aggregation approach. In RNN-based methods, Xie et al. (2018) aimed to address the issue of skeleton variations in 3D spatiotemporal space, which proposed a spatiotemporal memory attention network based on RNN and CNN to perform frame recalibration of skeleton data in the temporal domain. Regarding GNN-based methods, Yan et al. (2018) emerged as a classic approach based on spatial-temporal graph convolution networks. The core idea was to model human body joints as graph nodes and the connections between joints as graph edges, and the multiple graph convolutional layers were stacked to extract high-level spatial-temporal features. In Transformer-based methods, Plizzari et al. (2021) model employed a spatial self-attention module to capture intra-frame interactions among different body parts and a temporal self-attention module to model inter-frame correlations.

For IMU data, due to its ability to provide good complementary features and better privacy protection, it is gradually being used for human action recognition tasks. Through convolutional layers and pooling layers, CNN (Yi et al., 2023) were able to capture local and global features in IMU data, extract relationships between skeleton body parts, and achieve accurate classification of different actions. In IMU-based human action recognition, RNN (Al-qaness et al., 2022) utilized their memory units (e.g., Long Short-Term Memory Units or Gated Recurrent Units) to capture the temporal evolution of skeleton sequence, extracting crucial motion patterns and action features from it. Additionally, there have been research efforts that combined the strengths of CNNs and RNNs to comprehensively utilize the spatiotemporal information in IMU data for human activity recognition (Challa et al., 2022; Dua et al., 2023). It is worth noting that, with the progress of research, other IMU-based human action recognition methods have emerged, such as those based on Transformers (Shavit and Klein, 2021; Suh et al., 2023).

### 2.2. Multimodel human action recognition

Due to the limitation of single modal, it is difficult to further improve the performance of recognition model.

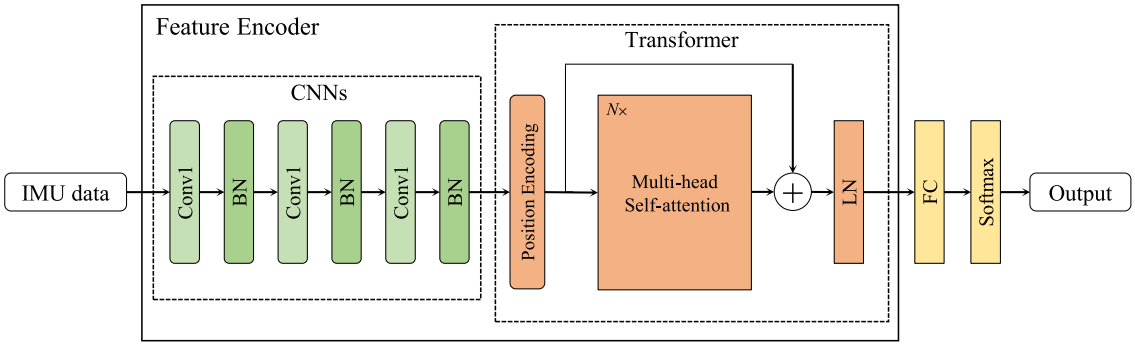



FIGURE 1
The feature encoder for IMU data. "BN" denotes batch normalization, "LN" indicates layer normalization, and $N\times$ represents that there are multiple multi-head self-attention modules.

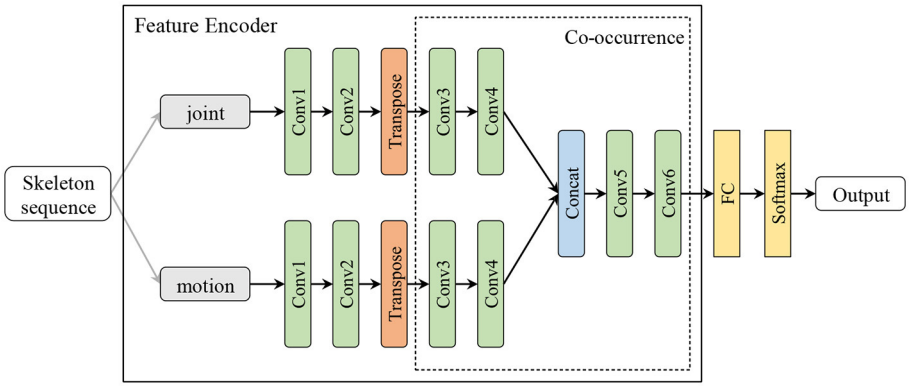



FIGURE 2
The feature encoder for skeleton sequence. The output channels of the 6 blocks 2D convolution layer are [64, 32, 32, 64, 128, 256]. The transpose layer transposes the dimensions of the input tensor according to the sequential parameters.

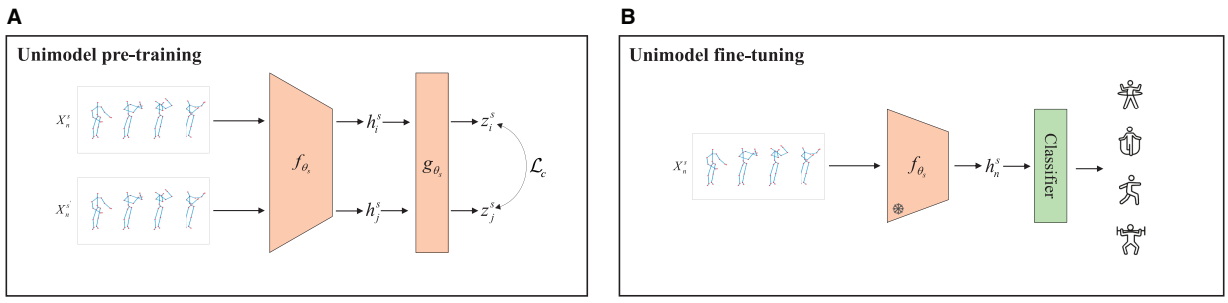



FIGURE 3
Contrastive learning framework for unimodal recognition. **(A)** Pre-training stage: for a skeleton sequence, the embedding representation $z$ is generated by the same encoder $f$ and projection head $g$ after data augmentation using contrast loss $L_c$, respectively. **(B)** Fine-tuning stage: the labeled skeleton sequence is passed through the frozen encoder $f$, and then processed through the classifier to obtain the action recognition label.

Since the complementary information provided by different modalities, researchers have become interested in combining multimodal features to improve recognition performance, such as skeleton and IMU data (Das et al., 2020; Khaertdinov and Asteriadis, 2022). There are many excellent recognition models are developed to leverage the strengths of different modalities and achieve more robust and accurate action recognition. However, the main challenge in executing multimodal recognition

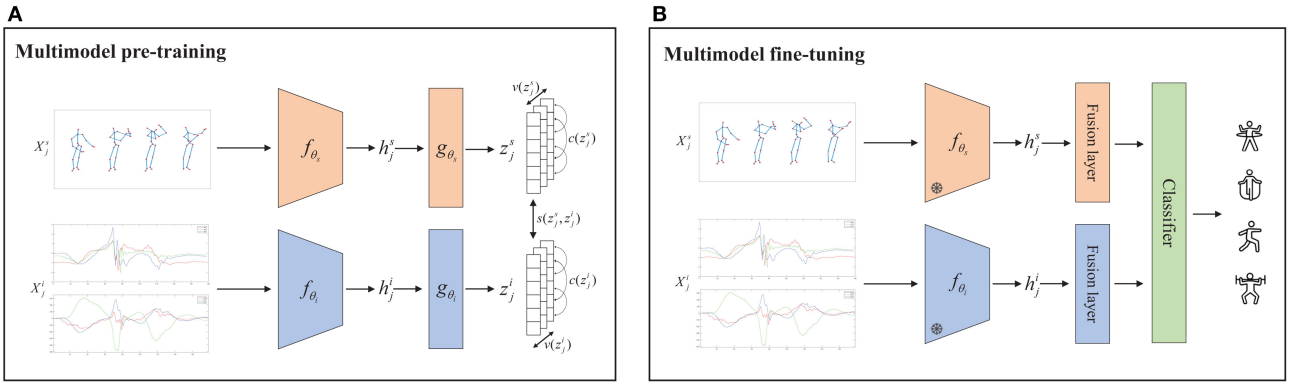



FIGURE 4
Contrastive learning framework for multimodal recognition. **(A)** Pre-training stage: the non-labeled skeleton sequences and IMU data were passed through modality-specific encoders *f* and projection heads *g* to generate embeddings representing *z* using contrasting loss. **(B)** Fine-tuning phase: the skeleton sequences with labels and the IMU data with labels were passed through frozen modality-specific encoders *f*, respectively, and then obtained action recognition label via fusion layer and classifier.

lies in effectively fuse the feature information from different modalities. Based on the above statement, the related work in multi-modal human action recognition can be roughly categorized into modality fusion and feature fusion, and we focus on the fusion method of skeleton sequence and IMU signal features.

Skeleton data provides precise positional information of human joints, while IMU data provides measurements from sensors such as accelerometers and gyroscopes (Das et al., 2020). By fusing skeleton and IMU data, more comprehensive and rich action features can be obtained. From the perspective of modality fusion, Fusion-GCN (Duhme et al., 2022) directly integrates IMU data into existing skeletons in the channel dimension during data preprocessing. Furthermore, RGB modality is processed to extract high-level semantic features, which are then fed into the GCNs as new nodes for fusion with other modalities. From the perspective of feature fusion (Khaertdinov and Asteriadis, 2022), features from different modalities are combined and integrated to achieve more representative and discriminative representations. In addition, cross-modal contrastive learning networks through knowledge distillation are also an effective identification method. Liu et al. (2021) proposed a Semantics-aware Adaptive Knowledge Distillation Network (SAKDN) that utilizes IMU data and RGB videos as inputs for the teacher and student model, respectively. The SAKDN adaptively fuses knowledge from different teacher networks and transfers the trained knowledge from the teacher network to the student network. The CMC (Tian et al., 2020) framework proposed a multi-modal learning architecture based on contrastive representation learning, which extended the representation learning to multiple modalities for improving the quality of the learned features with the number of modalities increased. It demonstrated the subtle relationship between mutual information across multiple modalities and multiple viewpoints. Similarly, CMC-CMKM (Brinzea et al., 2022) employed cross-modal knowledge distillation to perform feature-level fusion of IMU data and Skeleton information, which has achieved good recognition performance.

## 2.3. Contrastive learning for human action recognition

Recently, several advanced self-supervised learning methods have been proposed with excellent results in image and video tasks. Self-supervised contrast learning focuses on the variation between different views of the same or different samples, and better robust and transferable feature representations can be learned through contrast loss. SimCLR (Chen et al., 2020) incorporated a new contrastive loss function called Normalized Temperature-Scaled Cross-Entropy Loss (NT-Xent) into the network, which is a simple and effective contrastive learning framework. In contrast, BYOL (Grill et al., 2020) designed a more scalable and easily trainable self-supervised learning approach by contrasting the hidden representations in the network. Furthermore, to obtain more distinctive representations without requiring negative samples, Barlow Twins (Zbontar et al., 2021) minimized the correlation between features by employing the Barlow Twins loss. In addition, the biggest advantage of VICReg (Bardes et al., 2022) is its simplicity and effectiveness, which only necessary to compare along the batch dimension by invariance, variance and covariance, and does not require the weights of two branches to be shared.

In the case of action recognition tasks, most of the self-supervised contrastive learning is mainly applied to individual modalities, such as sensor data, skeleton sequence, or RGB video. To date, there has been a large number of works on fully supervised learning for multimodal human action recognition, and the disadvantage of these methods is that they require a large number of labeled samples for training. In contrary, to our knowledge, self-supervised contrastive learning frameworks

```
# f: encoder network
# lambda, mu, nu: coefficients of the
invariance, variance and covariance
 losses
# N: batch size
# D: dimension of the representations
# mse_loss: Mean square error loss function
# off_diagonal: off-diagonal elements
 of a matrix
# relu: ReLU activation function
for x_j^s, x_j^i in loader: # load a batch with
N samples
  # obtain augmented skeleton and
  # IMU samples
  x̃_j^s = 𝒯(x_j^s)
  x̃_j^i = 𝒯(x_j^i)
  # compute representations
  h_j^s = (f_θs(x̃_j^s)) # hidden layer feature
  h_j^i = (f_θs(x̃_j^i)) # hidden layer feature
  z_j^s = g_θs(h_θi)
  # embeddings for skeleton [N × D]
  z_j^i = g_θi(h_θi) # embeddings for IMU [N × D]
  # variance loss
  z_j^s = z_j^s − z_j^s.mean(dim = 0)
  z_j^i = z_j^i − z_j^i.mean(dim = 0)
  std_z_j^s = torch.sqrt(z_j^s.var(dim = 0) + 1e − 04)
  std_z_j^i = torch.sqrt(z_j^i.var(dim = 0) + 1e − 04)
  std_loss = torch.mean(relu(1 − std_z_j^s))
  +torch.mean(relu(1 − std_z_j^i))
  # invariance loss
  sim_loss = mse_loss(z_j^s, z_j^i)
  # covariance loss
  cov_z_j^s = (z_j^s.T @ z_j^s)/(N − 1)
  cov_z_j^i = (z_j^i.T @ z_j^i)/(N − 1)
  cov_loss = off_diagonal(cov_z_j^s).pow_(2).sum()/D
  +off_diagonal(cov_z_j^i).pow_(2).sum()/D
  # total loss
  loss = lambda ∗ sim_loss + mu ∗ std_loss
  +nu ∗ cov_loss
  # optimization step
  loss.backward()
  optimizer.step()
```

Algorithm 1. Multimodal pre-training pytorch pseudocode.

are rarely used in the field of multimodal human action recognition. Akbari et al. (2021) adopted a convolution-free Transformer architecture to train unlabeled video, audio, and text data end-to-end, and evaluated the model performance through downstream tasks such as video action recognition, audio event classification, image classification, and text-to-video retrieval. Inspired by VicReg (Bardes et al., 2022) and multimodal framework CMC, we propose a simple and effective self-supervised contrastive learning framework based on VICReg to address the multimodal human action recognition problem of IMU and skeleton data.

# 3. Methodology

## 3.1. Problem definition

Multimodal-based action recognition is defined as the fusion of different data modalities to obtain more comprehensive human pose and more precise action information. Specifically, for a given input $\{X^m | m \in M\}$ from a multimodal set $M$, the goal is to predict the label $y \in Y$ with the associated input $X$. In our work, we focus on IMU signal data and Skeleton sequences. IMUs could be used to measure the pose and acceleration of the human body with multivariate time series on the x, y and z axes for human motion recognition and analysis. Specifically, for S wearable sensors with S signal channels acquired at any t time stamp, we can define the input signal as $\boldsymbol{x_t} = \left[x_t^1, x_t^2, \ldots, x_t^S\right] \in \mathbb{R}^S$. Therefore, the IMU modal inputs are represented in matrix form as $X^i = [x_1, x_2, ..., x_T] \in \mathbb{R}^{T \times S}$ for any $T$ time stamp. Furthermore, skeleton sequences can be collected by a pose estimation algorithm or a depth camera, which contain several joints of a human body, and each joint has multiple position coordinates. For a given skeleton sequence $X^s \in \mathbb{R}^{C \times T \times V}$, as 2D coordinates are used, the input channel $C = 2$, $T$ denotes the number of frames in a sequence, and V means that the number of joints with respect to the dataset collection method.

## 3.2. Feature encoder

In order to obtain more effective features, we designed two feature encoders to handle IMU data and skeleton sequence, respectively, as shown in Figures 1, 2. In IMU data feature encoder, inspired by CSSHAR (Khaertdinov et al., 2021), we first employ a 1D convolution layer with 3 blocks for modeling in the temporal dimension, which includes a convolution kernel size of 3 and a feature map with channels of [32,64,128]. Furthermore, we employ a Transformer with a Multi-head self attention (heads $N = 2$) as the backbone to capture long-range dependencies from IMU data. Besides, inspired by hierarchical co-occurrence feature learning strategy, a two-stream framework is designed to learn and fuse the "joint" and "motion" features of skeleton sequences. Specifically, a skeleton sequence is divide into spatial joints and temporal motions. Then, they are fed into each of the four 2D CNN modules and assembled into semantic representations in both spatial and temporal domains, and point-level information of each joint is encoded independently.

## 3.3. Contrastive learning for unimodal recognition

As shown in Figure 3, given a skeleton sample in the pre-training, a positive sample pair $X_n^s$ and $X_n^{s'}$ could be obtained in a small batch by normal data augmentation. Then, they are fed into an encoder $f_{\theta_s}$ with HCN to yield the hidden layer features as

$$h_i^s = f_{\theta_s}(X_n^s) \tag{1}$$

$$h_j^s = f_{\theta_s}(X_n^{s'}) \tag{2}$$

TABLE 1 Pre-training hyperparameter settings.

| Modality | UTD-MHAD | | | MMAct | | |
|---|---|---|---|---|---|---|
| | Learning rate | Training scale | Batch size | Learning rate | Training scale | Batch size |
| IMU | 1e-2 | 100 epochs | 128 | 1e-3 | 100 epochs | 96 |
| Skeleton | 1e-2 | 100 epochs | 128 | 1e-3 | 100 epochs | 96 |
| IMU+Skeleton | 1e-3 | 200 epochs | 256 | 1e-4 | 200 epochs | 128 |

TABLE 2 The performance of action recognition for accuracy (%) and F1 score (%) is compared with the baseline methods.

| Method | Modality | UTD-MHAD | | MMAct cross-subject | | MMAct cross-scene | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score |
| Supervised_transformer | IMU | 79.77 | 79.59 | 62.15 | 62.32 | 78.27 | 71.86 |
| Supervised_cooccurrence | Skeleton | 93.49 | 93.43 | 80.53 | 81.93 | 78.61 | 74.30 |
| SimCLR | IMU | 64.65 | 64.64 | 52.32 | 51.94 | 66.16 | 60.28 |
| SimCLR | Skeleton | 92.09 | 91.87 | 75.97 | 76.75 | 72.62 | 62.04 |
| Barlow Twins | IMU | 58.60 | 57.69 | 45.17 | 44.11 | 59.96 | 51.77 |
| Barlow Twins | Skeleton | 88.84 | 88.24 | 67.86 | 69.24 | 60.68 | 52.34 |
| Barlow Twins | IMU+Skeleton | 91.63 | 91.72 | 82.17 | 81.98 | 82.70 | 80.05 |
| CMC | IMU+Skeleton | 95.12 | 95.08 | 82.05 | <u>83.06</u> | 84.01 | 82.41 |
| CMC-CMKM§ | IMU+Skeleton | 95.81 | 95.74 | <u>82.34</u> | 82.69 | 85.24 | 83.60 |
| **Ours** | IMU | 75.58 | 75.93 | 49.04 | 47.08 | 60.81 | 53.80 |
| **Ours** | Skeleton | 86.05 | 86.23 | 73.78 | 75.66 | 74.94 | 73.29 |
| **Ours** | IMU+Skeleton | <u>96.06</u> | **96.96** | **82.95** | **83.62** | <u>87.06</u> | <u>85.78</u> |
| Supervised | IMU+Skeleton | **96.51** | <u>96.36</u> | 81.78 | 82.86 | **89.47** | **87.94** |

Bolded data indicate the best results, underlined data the second best. § represents the reproduced results.

Inspired by the Barlow Twins, the feature representations $z_i^s$ and $z_j^s$ are obtained by an MLP projection layer, which are denoted as

$$z_i^s = g_{\theta_s}(h_i^s) \tag{3}$$

$$z_j^s = g_{\theta_s}(h_j^s) \tag{4}$$

Finally, to explore the relationship between the two views $X_n^s$ and $X_n^{s'}$, the cross-correlation matrix $\mathscr{C}$ between embedding $z_i^s$ and $z_j^s$ can be computed as follows

$$\mathscr{C}_{ij} = \frac{\sum_b z_{b,i} z'_{b,j}}{\sqrt{\sum_b \left(z_{b,i}\right)^2} \sqrt{\sum_b \left(z'_{b,j}\right)^2}}, \tag{5}$$

where $b$ denotes the batch dimension, $i$ and $j$ represent the embedding dimension. Finally, by enforcing the empirical cross-correlation matrix between the embeddings $Z^s$ of variations to be an identity matrix, the encoder could be used to capture the relationship between the two-stream siamese networks. The contrastive loss function is formulated as follows

$$L_c(Z^s) = \sum_i \left(1 - \mathscr{C}_{ii}\right)^2 + \beta \sum_i \sum_{j \neq i} \mathscr{C}^2{}_{ij} \tag{6}$$

Intuitively, the first term encourages the diagonal elements of $\mathscr{C}$ to converge to 1, so that the embedding is not subject to variation. The second term is intended to drive the different embedding components to be independent of each other, minimizing the redundancy of the output units and avoiding becoming a constant. $\beta$ is a positive constant used to weigh the first term and against the second term.

## 3.4. Contrastive learning for multimodal recognition

Our proposed VICReg-based multimodal recognition framework focuses on generating and contrasting embeddings from the IMU data and skeleton sequence branches, which eventually form a joint embedding architecture with variance, invariance and covariance regularization. It is a self-supervised learning method that incorporates two different modality training architectures based on the principle of preserving the content of the embedding information.

As shown in Figure 4, given a multimodal training sample $\{x_j^s, x_j^i\}$, where $s$ and $i$ refer to skeleton and IMU data modalities respectively. The augmented inputs are

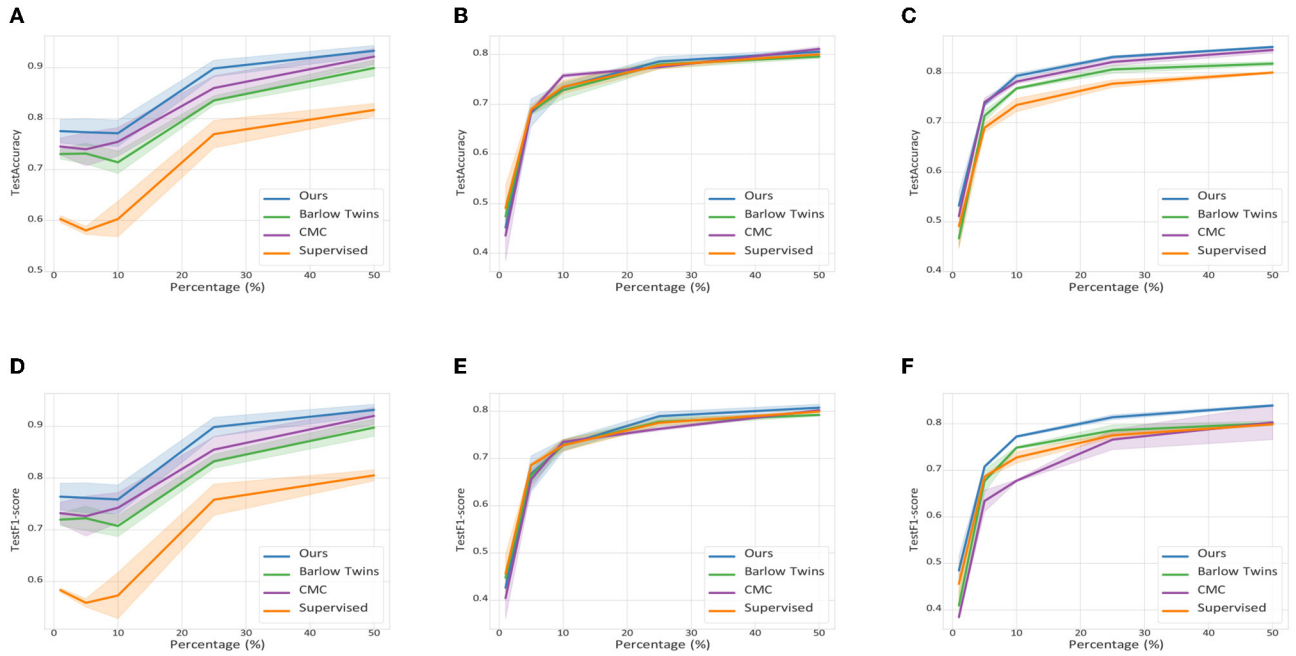



FIGURE 5
Average accuracy and F1 score with 95% confidence intervals for the semi-supervised learning scenario. **(A)** UTD-MHAD. **(B)** MMAct (cross-subject). **(C)** MMAct (cross-scene). **(D)** UTD-MHAD. **(E)** MMAct (cross-subject). **(F)** MMAct (cross-scene).

generated by modality-specific data augmentation in accordance with

$$x_j^s = \mathscr{T}(x_j^s) \tag{7}$$

$$\tilde{x}_j^i = \mathscr{T}(x_j^i) \tag{8}$$

In details, for the skeleton sequence augmentation methods are jittering, scaling, rotation, shearing, cropping and resizing, whereas the IMU data augmentation methods are jittering, scaling, rotation, permutation, shuffle of channel. Then, the feature representation of the two modalities are computed. Specifically, two modality-specific encoders $f_{\theta_s}$ and $f_{\theta_i}$ perform feature extraction to obtain the high-dimensional hidden layer features.

$$h_j^s = \left(f_{\theta_s}\left(\tilde{x}_j^s\right)\right) \tag{9}$$

$$h_j^i = \left(f_{\theta_s}\left(\tilde{x}_j^i\right)\right) \tag{10}$$

Both of these are passed through projection heads $g_{\theta_s}$ and $g_{\theta_i}$, implemented by a multilayer perceptron, and finally generate mode-specific embeddings representations of the two modalities which are $z_j^s = g_{\theta_s}\left(h_{\theta_i}\right)$ and $z_j^i = g_{\theta_i}\left(h_{\theta_i}\right)$. The loss function is calculated at the embedding level with respect to $z_j^s$ and $z_j^i$. We describe the three components of variance, invariance and covariance that constitute our loss function in the pre-training process.

Firstly, we define the variance regularization term $v$ to adopt the form of a hinge function that represents the standard deviation of the embeddings along the batch dimension.

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0, \gamma - Std\left(z^j, \epsilon\right)\right), \tag{11}$$

where $Std$ denotes the regularization standard deviation formula as:

$$Std(x, \epsilon) = \sqrt{\mathrm{Var}(x) + \epsilon}, \tag{12}$$

where we define $Z = [z_1, ..., z_n]$ consisting of $n$ vectors of dimension $d$ with embeddings $z_j$ from the feature encoding network of two modalities. $z^j$ is represented as the value of each vectors in $Z$ in dimension $j$, $\gamma$ denotes a fixed value of the standard deviation and defaults to 1 in our experiments. $\epsilon$ is a small scalar to guarantee data stability, which is set to 0.0001. The objective of this regularization term $v(Z)$ is to ensure that the variance of all embeddings $Z^s$ and $Z^i$ are close to $\gamma$ in the current batch ($s$ indicates the skeleton modality and i indicates the IMU modality), preventing all inputs from mapping on the same vector.

Secondly, we define the invariance regularization term $s$ by using the mean square Euclidean distance between two positive sample pairs $Z^s$ and $Z^i$. The formulation is as follows:

$$s(Z^s, Z^i) = \frac{1}{N}\sum_{j}^{N} \left\| z_j^s - z_j^i \right\|_2^2, \tag{13}$$

where $N$ denotes the batch size, both embeddings $Z^s$ and $Z^i$ come from the siamese architecture of the two branches.

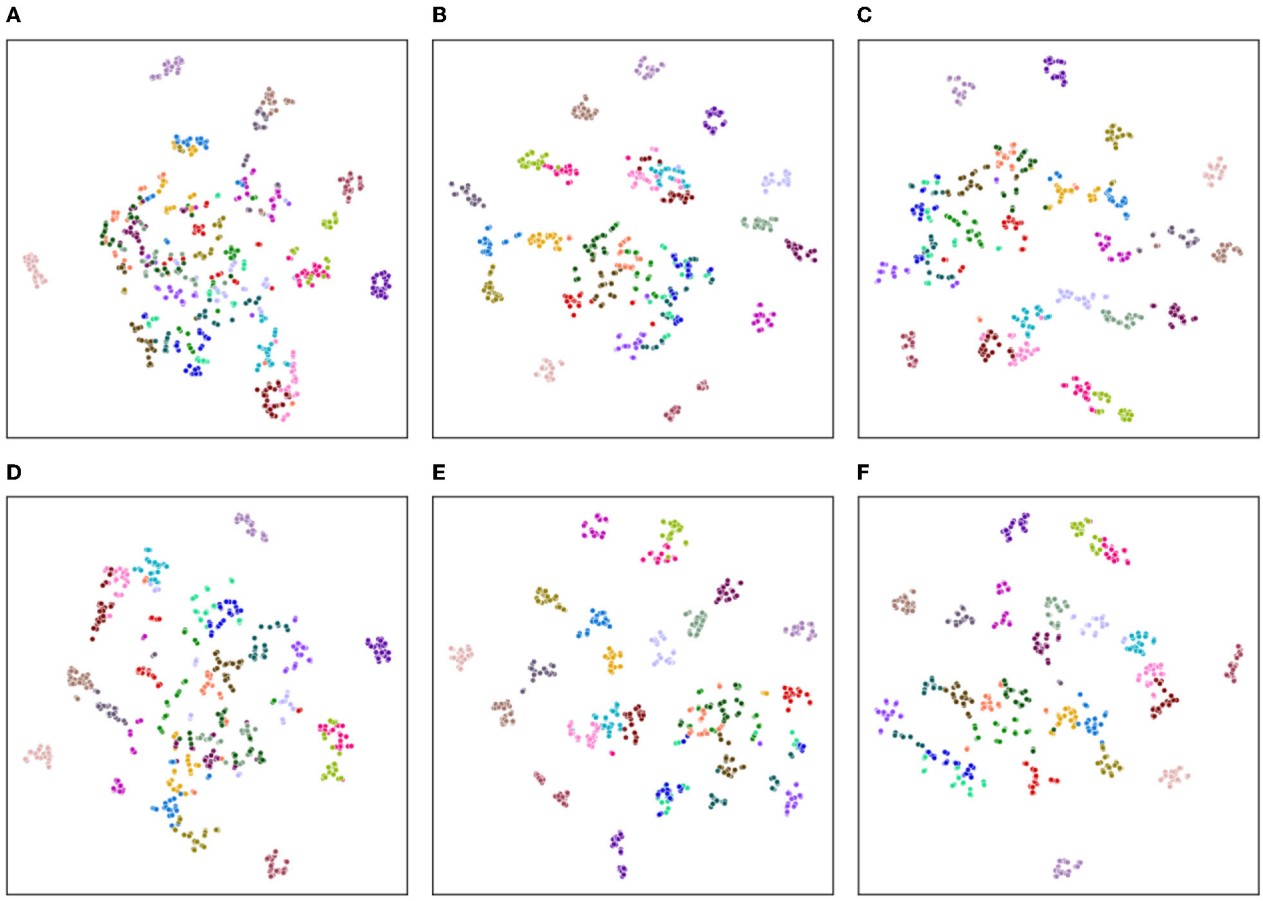



FIGURE 6
Visualization of representations learned using t-SNE for the UTD-MHAD benchmark. **(A)** Barlow Twins (IMU). **(B)** Barlow Twins (Skeleton). **(C)** Barlow Twins (IMU+Skeleton). **(D)** Ours (IMU). **(E)** Ours (Skeleton). **(F)** Ours (IMU+Skeleton).

Finally, the most critical component of the loss function, this term approximates the covariance between each pair of embedding variables to zero. Generally, it is the embeddings of the model that are decorrelated to each embedding variable to ensure the independence of the variables and prevent the model from learning similar or identical feature information. Inspired by Barlow Twins, we define the variance regularization term $c$ as:

$$c(Z) = \frac{1}{d}\sum_{i\neq j}\left[C(Z)\right]_{i,j}^{2}, \tag{14}$$

where the $1/d$ scales this function at the dimensional level and $C(Z)$ denotes the covariance matrix of the embeddings $Z$. The formula is expressed as follows:

$$C(Z) = \frac{1}{N-1}\sum_{j=1}^{n}\left(z_j - \bar{z}\right)\left(z_j - \bar{z}\right)^{T}, \bar{z} = \frac{1}{N}\sum_{j}^{N} z_j. \tag{15}$$

Therefore, the overall loss function with weighted average of the invariance, variance and covariance terms could be expressed as follows:

$$L(Z^s, Z^i) = \lambda * s(Z^s, Z^i) + \mu * \left[v(Z^s) + v(Z^i)\right] + \varphi * \left[c(Z^s) + c(Z^i)\right], \tag{16}$$

where $\lambda$, $\mu$, and $\varphi$ are hyperparameters that measure the importance of each loss component. In our experiment, $\varphi$ is set to 1 and a grid search is performed for the values of $\lambda$ and $\varphi$ with the basic condition $\lambda = \varphi > 1$.

The pseudo-code algorithm implementation is illustrated in Algorithm 1.

# 4. Experiments

## 4.1. Datasets

**UTD-MHAD** (Chen et al., 2015). The dataset is a multimodal dataset widely used for human action recognition, which includes RGB video, depth sequences, skeleton and IMU data. During the capturing process, 8 subjects perform 27 categories of actions, each individual repeating each action 4 times, for a total of

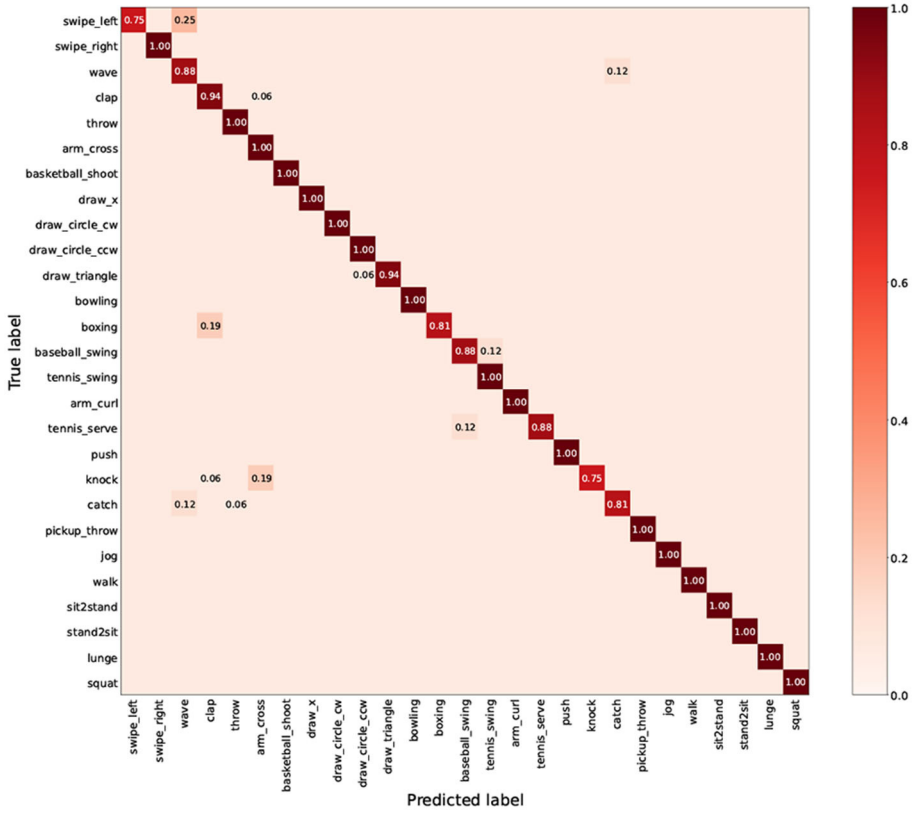



FIGURE 7
The normalized confusion matrix for UTD-MHAD.

861 samples. For the skeleton sequences, the Kinect camera would capture information regarding the subject's posture and movements. For the IMU data, the subjects were required to wear gloves, shoes and belts with IMU sensors attached, which recorded motion information on the subject's body parts, including accelerations, angular velocities and gyroscope data. Similar to the evaluation protocol in the original paper, we use data from odd-numbered subjects: 1, 3, 5, 7 as the training and validation sets, and data from even-numbered subjects: 2, 4, 6, 8 as the testing set, and report the accuracy and F1 score on the testing set.

**MMACT** (Kong et al., 2019). The dataset is a multimodal dataset consisting of 20 subjects performing 36 classes of actions, including skeleton sequences and IMU data. In this work, a challenge version of the dataset with 2D keypoints is adopted for the skeleton data. The IMU data is derived from smartphones including accelerometers, gyroscopes and orientation sensors. We verify our proposed recognition framework against the evaluation protocol from the previous study: cross-subject and cross-scene. For the cross-subject setting, the first 16 subject samples are used for training and validation, while the remaining ones are used for testing. For the cross-scene setting, the numbered 2 samples from the occlusion scene were used for testing and the rest for training, numbered 1, 3, 4. We report the accuracy and F1 score on the testing set.

## 4.2. Implementations details

Our experimental environment is implemented on the A5000 GPU platform using the Pytorch framework. Subsequently, we detailed three aspects: data pre-processing, pre-training and fine-tuning.

**Data pre-processing**. In order to normalize the IMU data and skeleton sequences, we employed a resampling method to uniformly represent all sequences with 50 time steps. Furthermore, to ensure consistency and comparability, we applied a standard normalization procedure to normalize the joints in all skeleton sequences. This normalization process involved scaling the joint positions based on the reference frame established by the first frame of each sequence. For data augmentation of skeleton sequences, we employ {jittering, random resized crops, scaling, rotation, shearing} for two benchmarks. For data augmentation of IMU data, we employ {jittering, scaling, permutation, rotation, channel shuffle}.

**Pre-training**. For the UTD-MHAD dataset, in unimodal pretraining, our proposed method uses a batch size of 100 and sets the random seed for both skeleton and IMU modalities to 28. The training is performed for 100 epochs with a learning rate of 1e-2 and Adam optimizer. In the case of multimodal pretraining, our proposed method increases the batch size to 200 epochs, adjusts the learning rate to 1e-3, and sets the training scale to 200 epochs. The optimizer remains Adam. For the MMAct dataset, we maintain

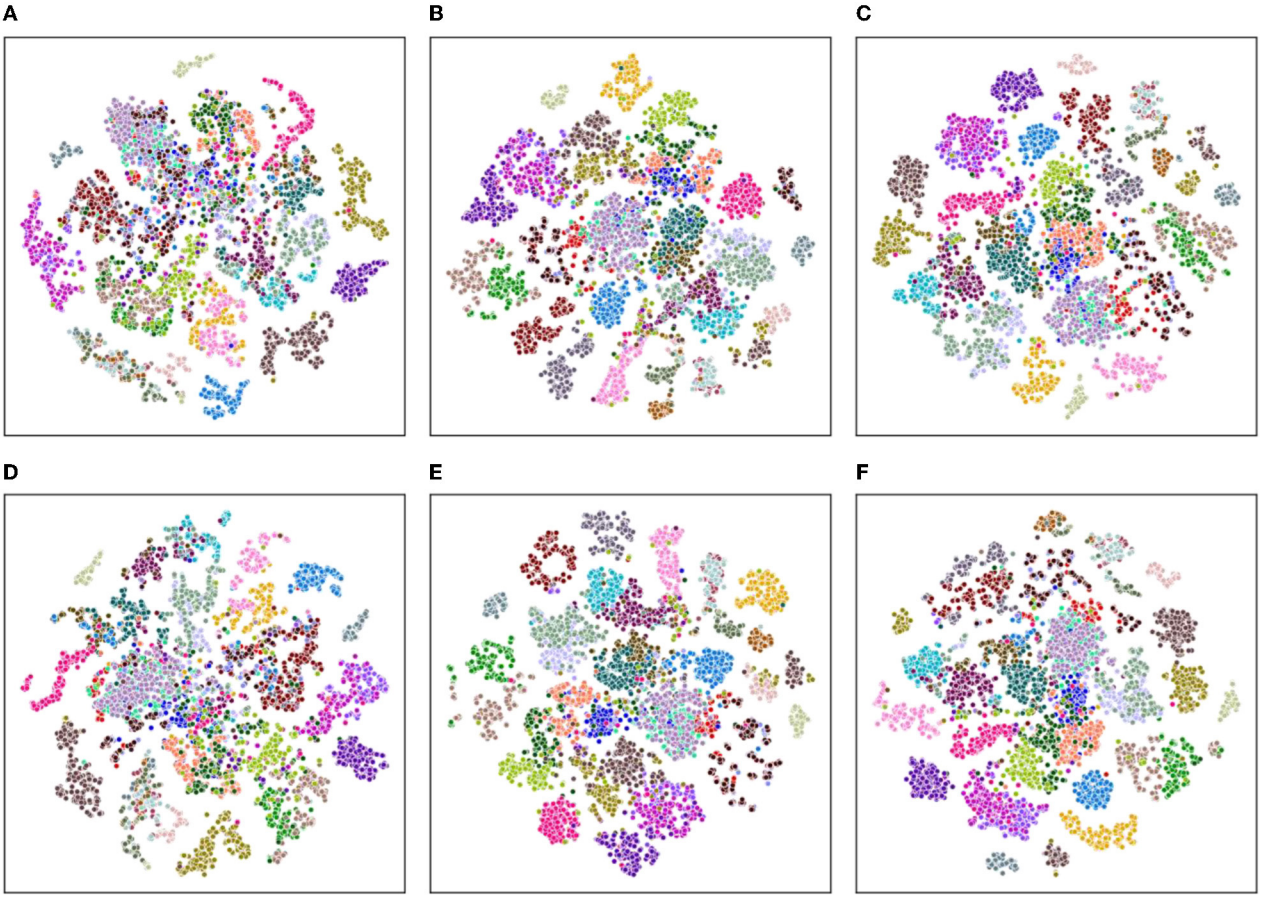



FIGURE 8
Visualization of representations learned using t-SNE for the MMAct benchmark. **(A)** Barlow Twins (IMU). **(B)** Barlow Twins (Skeleton). **(C)** Barlow Twins (IMU+Skeleton). **(D)** Ours (IMU). **(E)** Ours (Skeleton). **(F)** Ours (IMU+Skeleton).

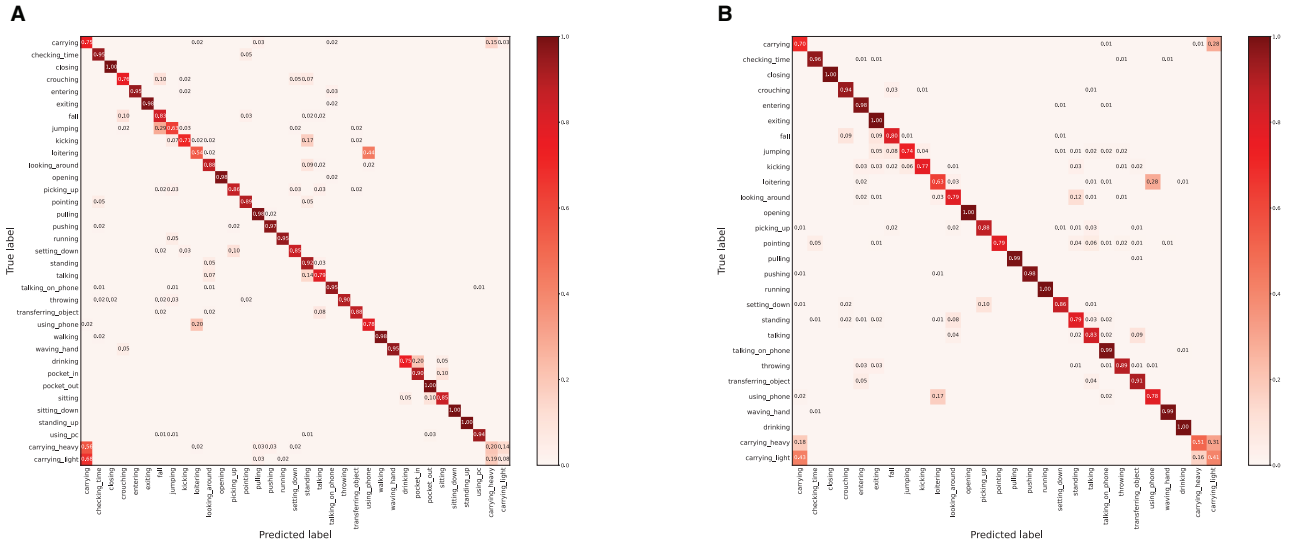



FIGURE 9
The normalized confusion matrix for MMAct benchmarks. **(A)** MMAct (cross-subject). **(B)** MMAct (cross-scene).

TABLE 3 Zero shot performance (%) on UTD-MHAD benchmark.

| Modality | num_classes=1 | | num_classes=2 | | num_classes=5 | |
|---|---|---|---|---|---|---|
| | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score |
| IMU | 73.95 | 73.95 | 73.49 | 73.79 | 75.35 | 75.54 |
| Skeleton | 88.84 | 88.43 | 87.91 | 87.84 | 89.77 | 89.51 |
| IMU+Skeleton | 95.58 | 95.59 | 93.95 | 93.85 | 96.05 | 96.00 |

TABLE 4 Zero shot performance (%) on MMAct benchmark.

| Modality | num_classes=1 | | num_classes=2 | | num_classes=5 | |
|---|---|---|---|---|---|---|
| | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score |
| IMU | 48.39 | 48.11 | 48.31 | 47.34 | 48.81 | 48.63 |
| Skeleton | 73.67 | 75.74 | 72.35 | 73.97 | 73.68 | 75.83 |
| IMU+Skeleton | 81.39 | 81.19 | 81.73 | 82.35 | 82.45 | 83.02 |

the same training settings as before, regardless of single or multi-modality. In unimodal pretraining, the learning rate is set to 1e-3, and the batch size is 96. In multimodal pretraining, we increase the batch size to 128 and adjust the learning rate to 1e-4. Similarly, the parameter initialization random seed is set to 28. All settings are shown in Table 1.

**Fine-tuning**. Following prior fine-tuning routines, we implemented modality-specific feature fusion layers for the multimodal fine-tuning process, including batch normalization and non-linear ReLU, mapping the embeddings of IMU data and skeleton sequence to the same size of 256. And then concatenated them up by a linear classifier with Softmax function. We train the samples with labels by fine-tuning the model both to 100 epochs either unimodal or multimodal for our action recognition task.

## 4.3. Evaluations

### 4.3.1. Learning feature representation

To evaluate the multimodal learned feature representation, we perform linear evaluation of the features extracted from a specific encoder and then input the labeled samples into the fine-tuned training encoder and linear classifier. The performance of our model is compared with existing state of the art methods, and the results as shown in Table 2.

From the accuracy and F1 score terms obtained from the linear evaluation, our method significantly outperforms unimodal (more than 20% for IMU and almost 10% for Skeleon) for two benchmarks when multimodal contrastive learning is implemented. When comparing the self-supervised learning baseline models, our method is superior to other contrastive learning methods in terms of the multimodal learning approach. However, for the unimodal learning approach, our method has relatively no advantage. It is possible that our method undergoes a certain degree of embeddings collapse when calculating the standard deviation and variance. Meanwhile, the accuracy and F1 score of our method are also slightly lower when comparing fully supervised learning, which may be due to the fact that the supervised learning approach can perform end-to-end feature extraction for specific modalities. It is worth noting that our proposed method achieves 82.95% accuracy and 83.62% F1 score for MMAct (cross-subject), which exceeds the supervised learning method by 1.17 and 0.76%, indicating that our method has a better learned feature representation for multimodal training.

### 4.3.2. Semi-supervised learning

In the experiments, we adopt proportional unlabeled IMU and Skeleton data to perform contrastive learning in the pre-training phase. In particular, we set a random percentage $p \in \{1\%, 5\%, 10\%, 25\%, 50\%\}$ to conduct the experiment. To obtain a reasonable fine-tuning result, we calculate the average accuracy under the evaluation protocol corresponding to that presented in the colored interval by repeating the training 10 times on each $p$. In addition, we train a supervised learning multimodal model using the same encoders (Transformer for IMU and Co-occurrence for Skeleton). Similarly, fine-tuning the two-stream siamese networks and performing feature fusion, the final recognition results are obtained by a linear classifier, especially noting that the weights of the encoders are randomly initialized.

As shown in Figure 5, despite training only a small number of labeled samples, the contrastive learning methods all exhibit excellent robustness and performance. Specifically, the contrastive learning based approach outperforms the supervised learning based approach when the labeled samples are less than 25%, regardless of the dataset. Besides, our proposed method is superior to both Barlow Twins and CMC contrastive learning based multimodal methods with arbitrary $p$ values, which further validate the effectiveness and generalization ability of our proposed method.

### 4.3.3. Qualitative analysis

In order to evaluate the clustering effect of the model from a qualitative perspective, we employ a t-Distributed Stochastic

Neighbor Embedding (t-SNE, van der Maaten and Hinton, 2008) method to visualize the high-dimensional embeddings into a two-dimensional plane.

As shown in Figures 6, 7, we explore the IMU-based, Skeleton-based and multimodal approaches on the UTD-MHAD and MMAct datasets, respectively. Compared to the Barlow Twins, from an intuitive point of view, our proposed method is obviously effective in separating action class. Moreover, it is discovered that the multimodal data clustering is better than the unimodal clustering by fusing the features of IMU and Skeleton modalities. Furthermore, to measure the classification performance of our proposed method after fine-tuning, we performed accuracy evaluation by normalizing the confusion matrix. As shown in Figures 8, 9, we plot the normalized confusion matrices on UTD-MHAD, MMAct (cross-subject) and MMAct (cross-scene) to intuitively evaluate the performance of the classifier.

### 4.4. Zero shot setting

In the zero shot setting, we further explore the proposed method on the IMU and skeleton modalities through hiding certain action groups during the pre-training process. Specifically, we ensured that the action categories index [1, 2, 5] were not leaked during the training process by masking them.

As shown in Tables 3, 4, the performance of our model is compared with existing state of the art methods. Regarding UTD-MHAD benchmark for the unimodal evaluation, we could observe that the difference of the model is not significant after fine-tuning, but the skeleton sequence-based is much higher 15% than the IMU-based method. This is probably due to the fact that the skeleton sequences are modeled in both spatial and temporal dimensions, whereas IMU is only considered in the temporal dimension. For the multimodal evaluation, the model achieved 96.05% for accuracy and 96.00% for F1 score with class_id = $<$ 5 $>$ hidden, which is very close to the results achieved without the zero shot approach. Furthermore, regardless of the action class hidden, it is noted that the multimodal-based achieves much higher accuracy than the unimodal-based approach, exceeding the IMU-based approach by approximately 20% and the skeleton-based approach by approximately 6%. This validates that our proposed method achieves superior results with multimodal data inputs, which demonstrate the ability of the proposed method to learn complementary information.

## 5. Conclusion

In this paper, we propose a simple and effective contrastive self-supervised learning framework for human action recognition. Specifically, we construct a multimodal dataset by combining skeleton sequences and IMU signal data, and feed them into pretrained modality-specific two-stream networks for feature encoding. During the fine-tuning stage, labeled data is fed into the frozen encoders with weight initialization, and a linear classifier is applied to predict actions. Extensive experiments demonstrate that our proposed method outperforms unimodal approaches. It is worth noting that our model achieves comparable performance to pure supervised multimodal learning in certain metrics. In the future, we plan to further investigate other modalities, such as depth maps and RGB videos, to enhance multimodal human action recognition methods. Additionally, by incorporating knowledge distillation and unsupervised learning techniques, we aim to explore different ways of feature fusion between modalities to improve its performance in complex scenarios.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

ZR and HYu: conceptualization and writing—review and editing. HYa: methodology and validation. ZR and ZX: software. JZ: formal analysis. HYu: resources. ZX and JZ: data curation and visualization. HYa and ZR: writing—original draft preparation. ZR: supervision and funding acquisition. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the National Natural Science Foundation of Guangdong Province (Nos. 2022A1515140119 and 2023A1515011307), Dongguan Science and Technology Special Commissioner Project (No. 20221800500362), Dongguan Science and Technology of Social Development Program (No. 20231800936242), and the National Natural Science Foundation of China (Nos. 61972090, U21A20487, and U1913202).

## Acknowledgments

The authors thank everyone who contributed to this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

# References


Ahn, D., Kim, S., Hong, H., and Ko, B. C. (2023). "Star-transformer: a spatio-temporal cross attention transformer for human action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3330–3339. doi: 10.1109/WACV56688.2023.00333

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., et al. (2021). "VATT: transformers for multimodal self-supervised learning from raw video, audio and text," in *Advances in Neural Information Processing Systems, Vol. 34,* eds M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Curran Associates, Inc), 24206–24221.

Al-qaness, M. A., Dahou, A., Abd Elaziz, M., and Helmi, A. (2022). Multi-resAtt: multilevel residual network with attention for human activity recognition using wearable sensors. *IEEE Trans. Indus. Inform.* 19, 144–152. doi: 10.1109/TII.2022.3165875

Andrade-Ambriz, Y. A., Ledesma, S., Ibarra-Manzano, M.-A., Oros-Flores, M. I., and Almanza-Ojeda, D.-L. (2022). Human activity recognition using temporal convolutional neural network architecture. *Expert Syst. Appl.* 191, 116287. doi: 10.1016/j.eswa.2021.116287

Bardes, A., Ponce, J., and Lecun, Y. (2022). "VICReg: variance-invariance-covariance regularization for self-supervised learning," in *ICLR 2022 - International Conference on Learning Representations*.

Brinzea, R., Khaertdinov, B., and Asteriadis, S. (2022). "Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition," in *2022 International Joint Conference on Neural Networks (IJCNN)* (Padua: IEEE), 1–8. doi: 10.1109/IJCNN55064.2022.9892522

Challa, S. K., Kumar, A., and Semwal, V. B. (2022). A multibranch cnn-bilstm model for human activity recognition using wearable sensor data. *Visual Comput.* 38, 4095–4109. doi: 10.1007/s00371-021-02283-3

Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). "UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International Conference on Image Processing (ICIP)* (Quebec City, QC), 168–172.

Chen, J., and Ho, C. M. (2022). "MM-VIT: multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI), 1910–1921. doi: 10.1109/WACV51458.2022.00086

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (PMLR)*, 1597–1607.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., and Lu, H. (2020). "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA), 183–192. doi: 10.1109/CVPR42600.2020.00026

Chi, H.-G., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., and Ramani, K. (2022). "InfoGCN: representation learning for human skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA), 20186–20196. doi: 10.1109/CVPR52688.2022.01955

Das, A., Sil, P., Singh, P. K., Bhateja, V., and Sarkar, R. (2020). MMHAR-ensemNet: a multi-modal human activity recognition model. *IEEE Sens. J.* 21, 11569–11576. doi: 10.1109/JSEN.2020.3034614

Dua, N., Singh, S. N., Semwal, V. B., and Challa, S. K. (2023). Inception inspired CNN-GRU hybrid network for human activity recognition. *Multimedia Tools Appl.* 82, 5369–5403. doi: 10.1007/s11042-021-11885-x

Duhme, M., Memmesheimer, R., and Paulus, D. (2022). "Fusion-GCN: multimodal action recognition using graph convolutional networks," in *Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021* (Bonn: Springer), 265–281. doi: 10.1007/978-3-030-92659-5_17

Feng, L., Zhao, Y., Zhao, W., and Tang, J. (2022). A comparative review of graph convolutional networks for human skeleton-based action recognition. *Artif. Intell. Rev.* 55, 4275–4305. doi: 10.1007/s10462-021-10107-y

Grill, J.-B., Strub, F., Altche, F., Tallec, C., Richemond, P., Buchatskaya, E., et al. (2020). "Bootstrap your own latent-a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems Vol. 33*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc), 21271–21284.

Grillini, A., Hernández-García, A., Renken, R. J., Demaria, G., and Cornelissen, F. W. (2021). Computational methods for continuous eye-tracking perimetry based on spatio-temporal integration and a deep recurrent neural network. *Front. Neurosci.* 15, 650540. doi: 10.3389/fnins.2021.650540

He, M., Hou, X., Ge, E., Wang, Z., Kang, Z., Qiang, N., et al. (2023). Multi-head attention-based masked sequence model for mapping functional brain networks. *Front. Neurosci.* 17, 1183145. doi: 10.3389/fnins.2023.1183145

Islam, M. M., Nooruddin, S., Karray, F., and Muhammad, G. (2022). Human activity recognition using tools of convolutional neural networks: a state of the art review, data sets, challenges, and future prospects. *Comput. Biol. Med.* 2022, 106060. doi: 10.1016/j.compbiomed.2022.106060

Khaertdinov, B., and Asteriadis, S. (2022). "Temporal feature alignment in contrastive self-supervised learning for human activity recognition," in *2022 IEEE International Joint Conference on Biometrics (IJCB)* (Abu Dhabi), 1–9. doi: 10.1109/IJCB54206.2022.10007984

Khaertdinov, B., Ghaleb, E., and Asteriadis, S. (2021). "Contrastive self-supervised learning for sensor based human activity recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, (Shenzhen: IEEE), 1–8. doi: 10.1109/IJCB52358.2021.9484410

Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., and Murakami, T. (2019). "MMACT: a large-scale dataset for cross modal human action understanding," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul), 8657–8666. doi: 10.1109/ICCV.2019.00875

Li, C., Zhong, Q., Xie, D., and Pu, S. (2018). Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*. doi: 10.24963/ijcai.2018/109

Li, H., Liu, M., Yu, X., Zhu, J., Wang, C., Chen, X., et al. (2023). Coherence based graph convolution network for motor imagery-induced EEG after spinal cord injury. *Front. Neurosci.* 16, 1097660. doi: 10.3389/fnins.2022.1097660

Li, T., Cheng, B., Ni, B., Liu, G., and Yan, S. (2016a). Multitask low-rank affinity graph for image segmentation and image annotation. *ACM Trans. Intell. Syst. Technol.* 7, 1–18. doi: 10.1145/2856058

Li, T., Mei, T., Yan, S., Kweon, I.-S., and Lee, C. (2009). "Contextual decomposition of multi-label images," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA), 2270–2277. doi: 10.1109/CVPR.2009.5206706

Li, T., Meng, Z., Ni, B., Shen, J., and Wang, M. (2016b). Robust geometric p-norm feature pooling for image classification and action recognition. *Image Vision Comput.* 55, 64–76. doi: 10.1016/j.imavis.2016.04.002

Liu, Y., Wang, K., Li, G., and Lin, L. (2021). Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Trans. Image Process.* 30, 5573–5588. doi: 10.1109/TIP.2021.3086590

Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., and Chiaberge, M. (2022). Action transformer: a self-attention model for short-time pose-based human action recognition. *Pattern Recogn.* 124, 108487. doi: 10.1016/j.patcog.2021.108487

Mughal, N. E., Khan, M. J., Khalil, K., Javed, K., Sajid, H., Naseer, N., et al. (2022). EEG-fNIRS based hybrid image construction and classification using CNN-LSTM. *Front. Neurorobot.* 16, 873239. doi: 10.3389/fnbot.2022.873239

Plizzari, C., Cannici, M., and Matteucci, M. (2021). "Spatial temporal transformer network for skeleton based action recognition," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event* (Springer), 694–701. doi: 10.1007/978-3-030-68796-0_50

Shavit, Y., and Klein, I. (2021). Boosting inertial-based human activity recognition with transformers. *IEEE Access* 9, 53540–53547. doi: 10.1109/ACCESS.2021.3070646

Shen, X., and Ding, Y. (2022). Human skeleton representation for 3d action recognition based on complex network coding and LSTM. *J. Vis. Commun. Image Represent.* 82, 103386. doi: 10.1016/j.jvcir.2021.103386

Shu, X., Zhang, L., Sun, Y., and Tang, J. (2021). Host-parasite: graph LSTM-in-LSTM for group activity recognition. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 663–674. doi: 10.1109/TNNLS.2020.2978942

Suh, S., Rey, V. F., and Lukowicz, P. (2023). Tasked: transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation. *Knowledge Based Syst.* 260, 110143. doi: 10.1016/j.knosys.2022.110143

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2023). Human action recognition from various data modalities: a review. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3200–3225. doi: 10.1109/TPAMI.2022.3183112

Tian, Y., Krishnan, D., and Isola, P. (2020). "Contrastive multiview coding," in *Computer Vision-ECCV 2020: 16th European Conference* (Glasgow), 776–794. doi: 10.1007/978-3-030-58621-8_45

Tu, Z., Zhang, J., Li, H., Chen, Y., and Yuan, J. (2022). Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Trans. Multimedia* 25, 1819–1831. doi: 10.1109/TMM.2022.3168137

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, T., Li, 479 J., Wu, H.-N., Li, C., Snoussi, H., and Wu, Y. (2022). Reslnet: deep residual lstm network with longer input for action recognition. *Front. Comput. Sci.* 16, 166334. doi: 10.1007/s11704-021-0236-9

Xie, C., Li, C., Zhang, B., Chen, C., Han, J., Zou, C., et al. (2018). "Memory attention networks for skeleton-based action recognition," in *International Joint Conference on Artificial Intelligence* (Stockholm), 1639–1645. doi: 10.24963/ijcai.2018/227

Xu, K., Ye, F., Zhong, Q., and Xie, D. (2022). "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2866–2874. doi: 10.1609/aaai.v36i3.20191

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence* (Phoenix, Arizona). doi: 10.1609/aaai.v32i1.12328

Yi, M.-K., Lee, W.-K., and Hwang, S. O. (2023). A human activity recognition method based on lightweight feature extraction combined with pruned and quantized CNN for wearable device. *IEEE Trans. Cons. Electron.* 1. doi: 10.1109/TCE.2023.3266506

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). "Barlow twins: self-supervised learning via redundancy reduction," in *International Conference on Machine Learning (PMLR)*, 12310–12320.

Zhang, J., Wu, F., Hu, W., Zhang, Q., Xu, W., and Cheng, J. (2019). "Wienhance: towards data augmentation in human activity recognition using wifi signal," in *MSN* (Shenzhen), 309–314. doi: 10.1109/MSN48538.2019.00065

Zhang, J., Wu, F., Wei, B., Zhang, Q., Huang, H., Shah, S. W., et al. (2020). Data augmentation and dense-LSTM for human activity recognition using wifi signal. *IEEE Internet Things J.* 8, 4628–4641. doi: 10.1109/JIOT.2020.3026732

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 02 August 2023
DOI 10.3389/fnins.2023.1235480



OPEN ACCESS

EDITED BY
Teng Li,
Anhui University, China

REVIEWED BY
Dawei Wang,
Northwestern Polytechnical University, China
Jinghui Sun,
Medical University of South Carolina,
United States

*CORRESPONDENCE
Zhengjun Hou
houzj@cqupt.edu.cn
Yin Tian
tianyin@cqupt.edu.cn

RECEIVED 06 June 2023
ACCEPTED 17 July 2023
PUBLISHED 02 August 2023

CITATION
Jiang Y, Qiao R, Shi Y, Tang Y, Hou Z and Tian Y (2023) The effects of attention in auditory–visual integration revealed by time-varying networks.
*Front. Neurosci.* 17:1235480.
doi: 10.3389/fnins.2023.1235480



# The effects of attention in auditory–visual integration revealed by time-varying networks

Yuhao Jiang[1,2,3], Rui Qiao[1,2], Yupan Shi[1,2], Yi Tang[1,2], Zhengjun Hou[1,2]* and Yin Tian[1,2]*

[1]Institute for Advanced Sciences, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]Guangyang Bay Laboratory, Chongqing Institute for Brain and Intelligence, Chongqing, China, [3]Central Nervous System Drug Key Laboratory of Sichuan Province, Luzhou, China

Attention and audiovisual integration are crucial subjects in the field of brain information processing. A large number of previous studies have sought to determine the relationship between them through specific experiments, but failed to reach a unified conclusion. The reported studies explored the relationship through the frameworks of early, late, and parallel integration, though network analysis has been employed sparingly. In this study, we employed time-varying network analysis, which offers a comprehensive and dynamic insight into cognitive processing, to explore the relationship between attention and auditory-visual integration. The combination of high spatial resolution functional magnetic resonance imaging (fMRI) and high temporal resolution electroencephalography (EEG) was used. Firstly, a generalized linear model (GLM) was employed to find the task-related fMRI activations, which was selected as regions of interesting (ROIs) for nodes of time-varying network. Then the electrical activity of the auditory-visual cortex was estimated via the normalized minimum norm estimation (MNE) source localization method. Finally, the time-varying network was constructed using the adaptive directed transfer function (ADTF) technology. Notably, Task-related fMRI activations were mainly observed in the bilateral temporoparietal junction (TPJ), superior temporal gyrus (STG), primary visual and auditory areas. And the time-varying network analysis revealed that V1/A1↔STG occurred before TPJ↔STG. Therefore, the results supported the theory that auditory-visual integration occurred before attention, aligning with the early integration framework.

KEYWORDS

**auditory–visual integration, attention, time-varying network connectivity, fMRI, EEG**

## 1. Introduction

Individuals are constantly exposed to a plethora of sensory information that they unconsciously integrate in order to comprehend their environment. Visual and auditory information constitutes the majority (over 90%) of the information that is perceived (Treichler, 1967; Ristic and Capozzi, 2023). Auditory–visual integration occurs when auditory and visual stimuli coincide temporally and spatially, and when two stimuli are presented within a close time interval and similar spatial arrangement (Stein and Meredith, 1990; Frassinetti et al., 2002; Stevenson et al., 2012; Spence, 2013; Tang et al., 2016; Ľuboš et al., 2021). Attention plays a crucial role in selectively processing external information and improving information processing performance through focusing on target locations (Posner and Rothbart, 2006; Zhang T. et al., 2022). Attention is instrumental in processing dynamic stimuli efficiently and enhancing

perception, as it directs limited cognitive resources toward information relevant to the current task (Tian et al., 2014; Li et al., 2015). In addition, the researches on the attention mechanism may help to improve deep neural networks for visual processing tasks (Zhang et al., 2019; Wang et al., 2020).

There is ongoing debate regarding the role of attention in multisensory integration, particularly in the case of auditory–visual integration. Three mainstream theories about the relationship between auditory–visual integration and attention were proposed in previous studies (Koelewijn et al., 2010; Xu et al., 2020). The first, the early integration framework, asserts that integration occurs prior to attention and can even drive it (Vroomen et al., 2001; Rachel et al., 2022). Evidence for this is seen in the "pip-pop effect," where the addition of auditory stimulation to a visual search task led to faster results (Erik et al., 2008). Then non-spatial auditory stimulation was added to the spatial visual experiment. The second theory, the late integration framework, demonstrates that multisensory integration appears behind attention. In other words, two unimodal (i.e., auditory and visual) events are attended to separately before they are integrated. This model indicates that attention is necessary for multisensory integration (Laura et al., 2005; Sébastien et al., 2022). A later study used a cross-modal attention preference task to prove that cross-modal interactions are influenced by attention (Romei et al., 2013; Wen et al., 2021). Furthermore, late integration suggests that late cross-modal effects are mediated by attentional mechanisms. The third theory is the parallel integration framework; here, the stage at which multisensory integration takes place is uncertain. Multisensory integration can be early or late, and it depends on experimental or external conditions (Calvert and Thesen, 2004; Sébastien et al., 2022). Some studies extended the seminal methods of the parallel integration framework (Talsma et al., 2010; Stoep et al., 2015). This may produce different results as a result of several factors, including task type (detection or identification), stimulus properties (simple or complex), and attention resources (exogenous or endogenous).

In the study of the relationship between attention and auditory–visual integration, various methods have been employed. Early research utilized behavioral data and discovered that an auditory stimulus influences the reaction time (RT) of a synchronous or nearly synchronous visual stimulus (Mcdonald et al., 2000; Shams et al., 2000; Laura et al., 2005; Zhang X. et al., 2022) and the reverse is also true (Platt and Warren, 1972; Bertelson, 1999). These results indicate that a simultaneous or near-simultaneous bimodal stimulus reduces stimulation uncertainty (Calvert et al., 2000), potentially supporting the early integration framework or enhancing stimulation response for the late framework (Stein et al., 1989; Zhang et al., 2021). However, external factors, such as the state of the experimental subjects, may be overlooked.

With the advancement of brain imaging technology, increasing numbers of researchers have turned to brain imaging to investigate the relationship between attention and auditory–visual integration. By utilizing an event-related potential component (ERP) of an auditory–visual streaming design and a rapid serial visual presentation paradigm, they explored the interactions between multisensory integration and attention (Durk and Woldorff, 2005; Kang-jia and Xu, 2022). The results indicated that activity associated with multisensory integration processes is heightened when they are attended to, suggesting that attention plays a critical role in auditory–visual integration and aligning with the late integration criteria. The improvement of the spatial resolution of scalp EEG has long been a subject of interest for researchers.

Studies using functional magnetic resonance imaging (fMRI) with high spatial resolution have reported the accurate location of many areas involved in auditory–visual integration and attention; these mainly include the prefrontal, parietal, and temporal cortices (Calvert et al., 2001; Macaluso et al., 2004; Tedersälejärvi et al., 2005; Noesselt et al., 2007; Cappe et al., 2010; Chen et al., 2015).The superior temporal gyrus (STG) and sulcus (STS) both participate in speech auditory–visual integration (Klemen and Chambers, 2012; Rupp et al., 2022) and non-speech auditory–visual stimuli (Yan et al., 2015). In the past, STG was considered an area of pure sound input (Mesgarani et al., 2014). The temporoparietal junction (TPJ), which is close to the STG, is an important area of the ventral attention network (VAN) that is located mostly in the right hemisphere, and is recruited at the moment a behaviorally relevant stimulus is detected (Corbetta et al., 2008; Tian et al., 2014; Branden et al., 2022). The TPJ is activated during detection of salient stimuli in a sensory environment for a visual (Corbetta et al., 2002, 2008), auditory (Alho et al., 2015), and auditory–visual task (Mastroberardino et al., 2015). However, as many studies have mentioned, it is difficult to determine accurately the timing characteristics when using fMRI with poor temporal resolution.

For the reason that EEG and fMRI are two prominent noninvasive functional neuroimaging modalities, and they demonstrate highly complementary attributes, there has been a considerable drive toward integrating these modalities in a multimodal manner (Abreu et al., 2018). The combination of scalp EEG's exceptional temporal resolution and fMRI's remarkable spatial resolution enables a more comprehensive exploration of brain activity, surpassing the limitations inherent to individual techniques (Bullock et al., 2021). Previous investigations have examined the functional aspects of the brain in various pathological conditions, such as schizophrenia (Baenninger et al., 2016; Ford et al., 2016). Multiple researchers have employed combination of EEG and fMRI to explore cognitive mechanisms (Jorge et al., 2014; Shams et al., 2015; Wang et al., 2018). Some other studies have investigated brain dynamics in relation to complex cognitive processes like decision-making and the onset of sleep (Bagshaw et al., 2017; Pisauro et al., 2017; Hsiao et al., 2018; Muraskin et al., 2018). In this study, we used these two neuroimaging technologies to investigate the appearance order of auditory–visual integration and attention. Previous studies have tended to apply a specific experimental paradigm to investigate this relationship, but few have used network analyses to resolve this conundrum. We employed time-varying network analysis based on the adaptive directed transfer function (ADTF) method to uncover dynamic information processing. This method can uncover the dynamic information processing with a multivariate adaptive autoregressive mode (Li et al., 2016; Tian et al., 2018b; Nazir et al., 2020). This approach may offer new insights into the temporal order of multisensory integration and attention in a stimulated EEG network.

## 2. Materials and methods

### 2.1. Participants

The data for this study was obtained through separate EEG and fMRI recordings, conducted on 15 right-handed, healthy adult males

(mean ± standard deviation (SD) = 21.4 ± 2.8 years). Participants provided informed consent and were free from visual or auditory impairments and any mental health conditions. Upon completion of the experiments, participants were compensated for their time. The study was approved by the Ethics Committee of the University of Electronic Science and Technology of China.

## 2.2. Experimental design

Throughout the experiment, a white fixation cross of dimensions (0.5° × 0.5°) was presented at the center of a black monitor. The visual stimuli consisted of rectangular boxes that randomly appeared in either the left or right visual field (LVF or RVF, respectively). The box was 2° × 2° and its width was 0.2Y. The boxes remained on the screen for 50 ms and were followed by an auditory stimulus, a 1,000 Hz pure tone that also randomly appeared in the left or right auditory field (LAF and RAF, respectively) after a 50 or 750 ms interval. Participants were instructed to respond by pressing the 'Z' key with their left hand if the tone appeared in the LAF, and the '/' key with their right hand if it appeared in the RAF. Participants were required to react as soon as they heard the pure tone, which lasted for 200 ms. The fixation cross remained on the monitor for an additional 800 ms to ensure participants had sufficient time to respond correctly. The experimental procedure is illustrated in Figure 1.

## 2.3. Behavioral data and analysis

The behavioral data was obtained via EEG and fMRI. We analyzed RT using repeated measures analysis of variance (ANOVA) with the following factors: stimulus visual field (LVF vs. RVF), cue validity (valid vs. invalid), stimulus-onset asynchrony (SOA), and the interval between the cue and target stimulus (long vs. short). Data consistency was ensured by excluding RTs greater than 900 ms and less than 200 ms, as well as any instances of missed or incorrect key presses.

## 2.4. EEG and fMRI data recording

In the study, EEG and fMRI data were collected separately. We used a Geodesic Sensor Net (GSN) with 129-scalp electrodes located according to the International 10–20 system (Tucker, 1993) to record the EEG at a rate of 250 Hz. The Oz, Pz, CPz, Cz, FCz, and Fz electrodes were placed in the middle of the skull, and the remaining electrodes were distributed along both sides of the midline. The central top electrode (Cz) was used as the reference electrode and all electrodes had impedances lower than 40 kΩ (Tucker, 1993).

fMRI data was collected using the fast T2*-weighted gradient echo EPI sequence on a 3-T GE MRI scanner (TR = 2000 ms, TE = 30 ms, FOV = 24 cm × 24 cm, flip angle = 90°, matrix = 64 × 64, 30 slices) at the University of Electronic Science and Technology of

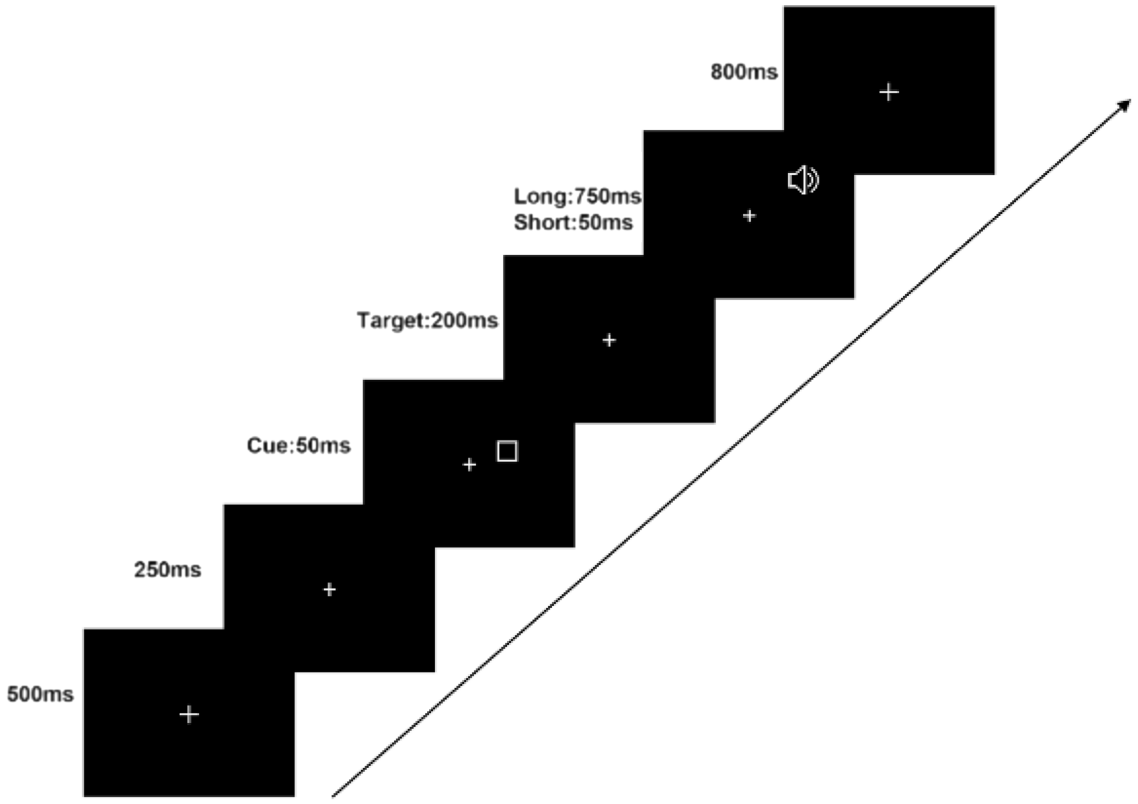



FIGURE 1
Illustration of stimulus sequence in the experiment.

China. This method obtained 198 volumes for each session. Because the machine at was unstable at the beginning of the data collection, we discarded the first five image volumes of each run for preprocessing.

## 2.5. The processing framework for time-varying networks

The processing framework for calculating time-varying networks consisted of three stages, as illustrated in Figure 2.

1. ROI selection based on task-related fMRI activations, as shown in Figure 2A.

The fMRI data was preprocessed and constructed by a generalized linear model (GLM). The results of the GLM were then subjected to a statistical test. Reply on the statistical results, 4 activations for the left cue and 4 activations for the right cue in the fMRI experiment were selected as ROIs (nodes) in the cerebral cortex, providing relatively accurate MNI coordinates for the construction of the time-varying network in the following steps.

2. Source wave extraction (Figure 2B).

The EEG data was preprocessed, and the scalp electrical signals are mapped to the cerebral cortex by MNE source localization method. Then, the MNI coordinates provided by fMRI were converted to the corresponding positions of the head model and the corresponding time series of the cortical electrical signals are extracted.

3. Time-varying network construction (Figure 2C).

In the third stage, the time-varying network was constructed using the ADTF technology, based on the results from steps 1 and 2.

## 2.6. fMRI data processing

The remaining volumes underwent preprocessing using Statistical Parametric Mapping version 8 (SPM8) software. Four preprocessing pipelines were applied in this study. Firstly, slice timing correction was implemented to address temporal differences among the slices. Secondly, spatial realignment was performed to eliminate head movement, whereby all volumes were aligned with the first volume. Participants whose head movement exceeded 2 mm or 2 degrees were excluded (Bonte et al., 2014). Thirdly, normalization was carried out to standardize each participant's original fMRI image to the standard Montreal Neurological Institute (MNI) space using EPI templates.

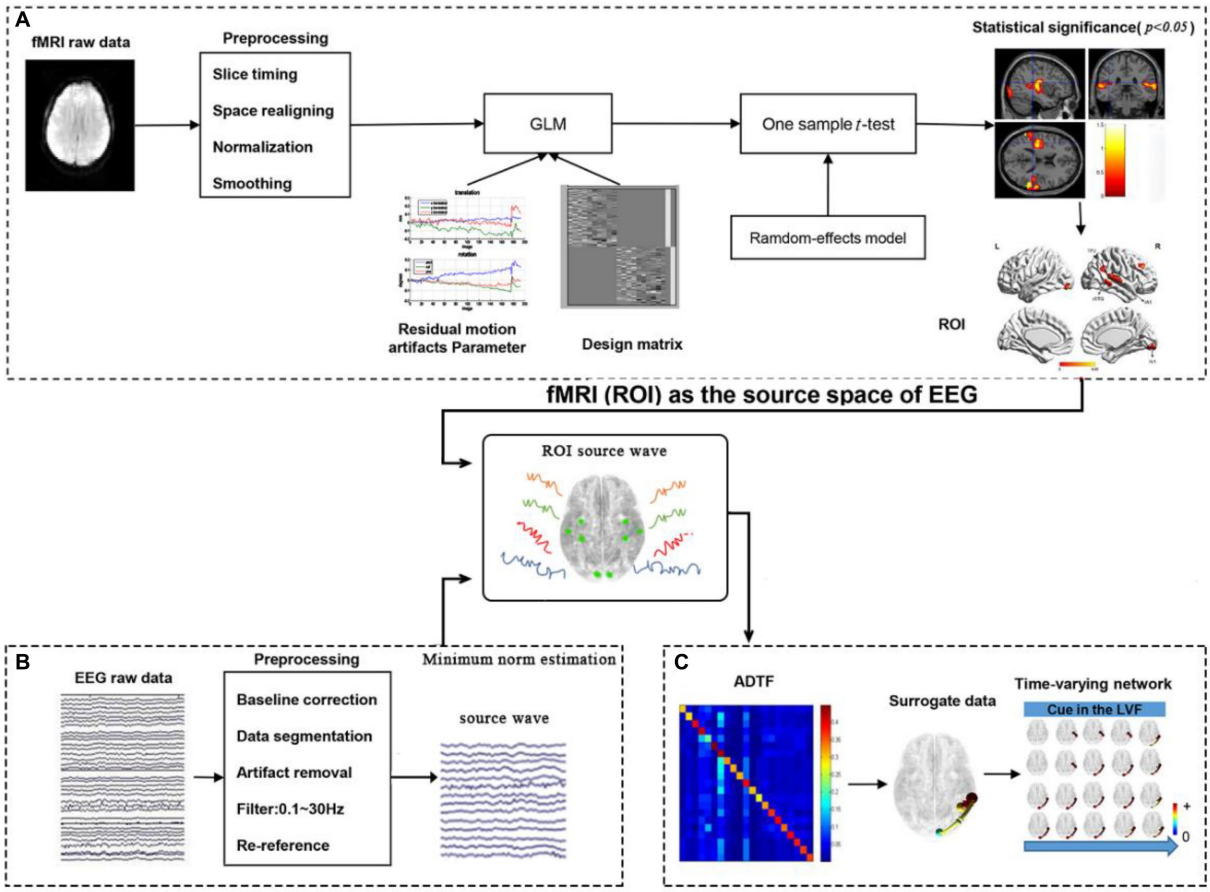



FIGURE 2
The processing framework for calculating time-varying networks. **(A)** fMRI processing; **(B)** EEG processing; **(C)** Time-varying network constructing.

Voxel resampling to $3\times3\times3\,mm^3$ was performed to overcome head size inconsistencies. Lastly, spatial smoothing was implemented to ensure high signal-to-noise ratio (SNR) by smoothing the functional images with a Gaussian kernel of full width half maximum (FWHM) of $6\times6\times6\,mm^3$.

After data preprocessing, the time series of all voxels underwent a high-pass filter at 1/128 Hz and were then analyzed with a general linear model (GLM; Friston et al., 1995) using SPM8 software. Temporal autocorrelation was modeled using a first-order autoregressive process. At the individual level, a multiple regression design matrix was constructed using the GLM, that included two experimental events based on the cue location (left visual field or right visual field). The two events were time-locked to the target of each trial by a canonical synthetic hemodynamic response function (HRF) and its temporal and dispersion derivatives. By including dispersion derivatives, the analysis accounted for variations in the duration of neural processes induced by the cue location. Nuisance covariates, such as realignment parameters, were included to account for residual motion artifacts. Parameter estimates were obtained for each voxel using weighted least-squares, which provided maximum likelihood estimators based on the temporal autocorrelation of the data (Wang et al., 2013).

In this study, to compute simple main effects for each participant, baseline contrasts were applied to the experimental conditions. Subsequently, the resulting individual contrast images were entered into a second-level one sample *t*-test using a random-effects model. In order to identify areas of significant activation, a threshold of $p<0.05$ (false discovery rate [FDR] corrected) and a minimum cluster size of 10 voxels were utilized. These stringent criteria were employed to ensure robust and reliable identification of neural activation patterns.

## 2.7. EEG data preprocessing

The EEG data underwent five preprocessing steps. Firstly, the EEG epochs were set to a time range of −200 to 1,000 ms. Secondly, we used the average of 200 ms pre-stimulus data as a baseline to correct the epochs. Thirdly, we performed artifact rejection, excluding epochs contaminated by eye blinks, eye movements, amplifier clipping or muscle potentials that exceeded ±75 μv. Fourthly, we filtered the EEG recordings using a band-pass filter of 0.1-30 Hz. Finally, we re-referenced the data using the reference electrode standardization technique (REST) (Yao et al., 2005; Tian and Yao, 2013; Tian et al., 2018a). We excluded trials with incorrect behavioral responses and bad channel replacements, and averaged the ERPs from the stimulus onset time point based on the validity of the cue, visual field, and SOA length.

## 2.8. Minimum norm estimation

The volume conductor effect may lead to the generation of pseudo-connections during brain network construction using scalp brain electricity. And invasive methods for directly collecting brain electricity in the cerebral cortex are challenging to use. To overcome this problem, we employed source localization technology to transfer scalp brain electrical signal to the cortex, enabling estimation of cortical electrical signals (Tian et al., 2018a; Tian and Ma, 2020), and we converted 129 scalp electrodes into 19 electrodes covering the whole brain.

In this study, we used the normalized minimum norm estimation (MNE) source localization method to estimate the electrical activity of the auditory–visual cortex. Compared to other methods, the normalized MNE offers higher dipole positioning accuracy, especially in depth source analysis. Our head model consisted of a three-layer realistic representation of the cortex, skull, and scalp. The formula for MNE calculation is expressed as follows:

$$\varphi(t)=\omega x(t) \tag{1}$$

Where $x(t)$ is the EEG collected by the scalp, $\varphi(t)$ is the corresponding cortical EEG, and $\omega$ is the field matrix, which can be obtained from the following formula:

$$\omega=C_sA^T\left(AC_sA^T+\mu^2C_n\right)^{-1} \tag{2}$$

where $C_s$ is the signal covariance, $C_n$ is the noisy covariance, and $A$ is the transfer matrix obtained by the boundary element theory. μ is a regularization parameter and is obtained by the following formula:

$$\mu=\frac{trace\left(AC_sA^T\right)}{trace\left(C_n\right)\times snr^2} \tag{3}$$

where *snr* is the signal to noise ratio.

## 2.9. Cortical time-varying network

The MNE source localization method was employed to transfer scalp electrical signals to the cerebral cortex. Next, MNI coordinates obtained from fMRI were mapped to the corresponding positions on the head model, and the cortical electrical signal time series at these positions were extracted. Subsequently, we designated these positions as nodes of the network and constructed the time-varying network using the relationship between these time series as the network edges.

To calculate the ADTF, we computed the multivariate adaptive autoregressive (MVAAR) model for all conditions. The model was normalized and expressed by following equation:

$$X(t)=\sum_{j}^{p}\omega(j,t)X(t-j)+\eta(t) \tag{4}$$

where $X(t)$ is the EEG data vector over the entire time window, $\omega(j,t)$ is the coefficient matrix of the time-varying model, which can be calculated by the Kalman filter algorithm, and $\eta(t)$ represents the multivariate independent white noise. The symbol $p$ denotes the MVAAR model order selected by Schwarz Bayesian Criterion (Schwarz, 1978; Wilke et al., 2008; Tian et al., 2018b).

After obtaining the coefficients of the MVAAR model, we calculated the ADTF by applying Equation (5) to convert the model coefficient $\omega(j,t)$ to the frequency domain. The $H_{ij}$ element of

$H(f,t)$ describes the directional information flow between the *j*th and the *i*th element at each time point *t* as:

$$\omega(f,t) * X(f,t) = \eta(f,t) \tag{5}$$

$$X(f,t) = \omega^{-1}(f,t) * \eta(f,t) = H(f,t) * \eta(f,t) \tag{6}$$

where $\omega(f,t) = \sum_{k=0}^{p} \omega_k(t) e^{-j2\pi f \cdot tk} \omega_k$ is the matrix of the time-varying coefficients. $\acute{E}(f,t)$ and $\eta(f,t)$ are transformed into the frequency domain as $X(t)$ and $\eta(f,t)$, respectively.

Defining the directed causal interrelation from the *j*th to the *i*th element, the normalized ADTF is described between (0,1) as follows:

$$\iota_{ij}^2(f,t) = \frac{|H_{ij}(f,t)|^2}{\sum_k^n |H_{ik}(f,t)|^2} \tag{7}$$

To obtain total information flow from a single node, the integrated ADTF is calculated as the ratio of summed ADTF values divided by the interested frequency bands ($f_1$, $f_2$):

$$\upsilon_{ij}^2(t) = \frac{\sum_{f1}^{f2} \acute{e}_{ij}^2(k,t)}{f2 - f1} \tag{8}$$

Surrogate data were used to establish the empirical ADTF value distribution under the connectionless zero assumption since the ADTF function has a highly non-linear correlation with the time series it derives, making it impossible to determine the distribution of the ADTF estimator under zero assumption without causality. The shuffling procedure independently and randomly iterated Fourier coefficient phases to produce new surrogate data while preserving the spectral structure of the time series (Wilke et al., 2008). To establish a statistical network, the nonparametric signed rank test was used to select statistically significant edges. The shuffling procedure was repeated 200 times for each model-derived time series from each participant to obtain the significance threshold of $p < 0.05$ with Bonferroni correction (Tian et al., 2018b).

## 2.10. Correlation analysis

The relationship between the information flow and the corresponding average response time (RT) was calculated using Pearson correlation based on the results of time-varying network analysis.

# 3. Results

## 3.1. Behavioral data analysis

Significant effects were observed for SOA ($F[1,14] = 9.85$, $p < 0.01$) and validity ($F[1,14] = 8.74$, $p < 0.05$), as well as their interaction ($F[1,14] = 27.54$, $p < 0.001$). However, no significant visual field effect ($F[1,14] = 3.60$, $p > 0.05$) or interactions between visual field and SOA or validity were found.

Because SOA, validity, and their interaction were significant, we conducted paired t-tests for the effects of SOA and validity (Figure 3). The results showed that participants reacted significantly faster in long SOA-invalid trials (268.94 ± 19.33 ms) than in long SOA-valid trials (277.79 ± 17.91 ms). In short SOA-invalid trials (291.91 ± 20.76 ms), participants took significantly more time to react than in short SOA-valid trials (273.80 ± 20.87 ms). There were also significant differences between long and short SOA-invalid trials. Although the RTs of long SOA-valid trials were slower than those of short SOA-valid trials, the difference was not significant.

## 3.2. fMRI results

A single sample t-test was performed to analyze fMRI data, revealing areas related to visual (V1), auditory (A1), multisensory integration (STG), and attention (angular, middle frontal cortex [MFG]) in both the left and right visual field (LVF and RVF). In the LVF, the main activated areas ($p < 0.05$, FDR correction) included the right angular gyrus (BA39), which is part of the right temporoparietal junction (rTPJ), right STG (BA21), right Heschl's gyrus as A1 (BA48), right lingual gyrus as rV1 (BA18), and right MFG (BA46), as shown in Figure 4A. In the RVF, more activated areas ($p < 0.05$, FDR correction) included left STG (BA21), left A1 (BA48), left V1 (BA17), bilateral MFG (BAs44/45/46), right TPJ (BA39), and so on, as shown in Figure 4B.

We selected four ROIs based on the task-related fMRI activations depicted in Figure 4 for both LVF and RVF cues. Specifically, when the cue appeared in either the LVF or RVF, the right TPJ (rTPJ) was selected for both LVF and RVF cues. When the cue appeared in the LVF, we chose the right STG (rSTG), right A1 (rA1), and right V1 (rV1). When the cue appeared in the RVF, we chose the contralateral STG, A1, and V1. The coordinates and sizes of the ROIs are presented in Table 1.

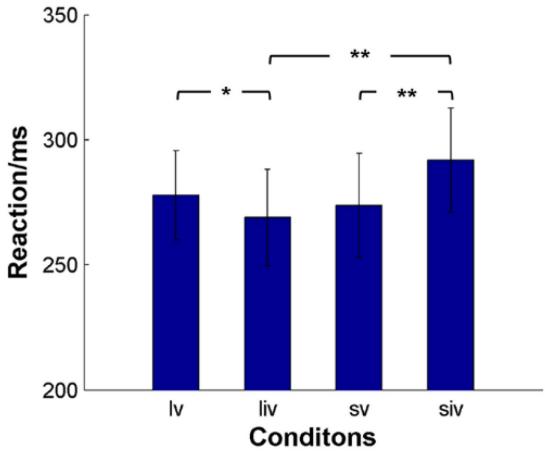



FIGURE 3
The average response time (RT) of subjects for the four conditions. lv denotes long SOA-valid condition, liv denotes long SOA-invalid condition, sv denotes short SOA-valid condition, siv denotes short SOA-valid condition. $^{**}p < 0.001$, $^{*}p < 0.05$.

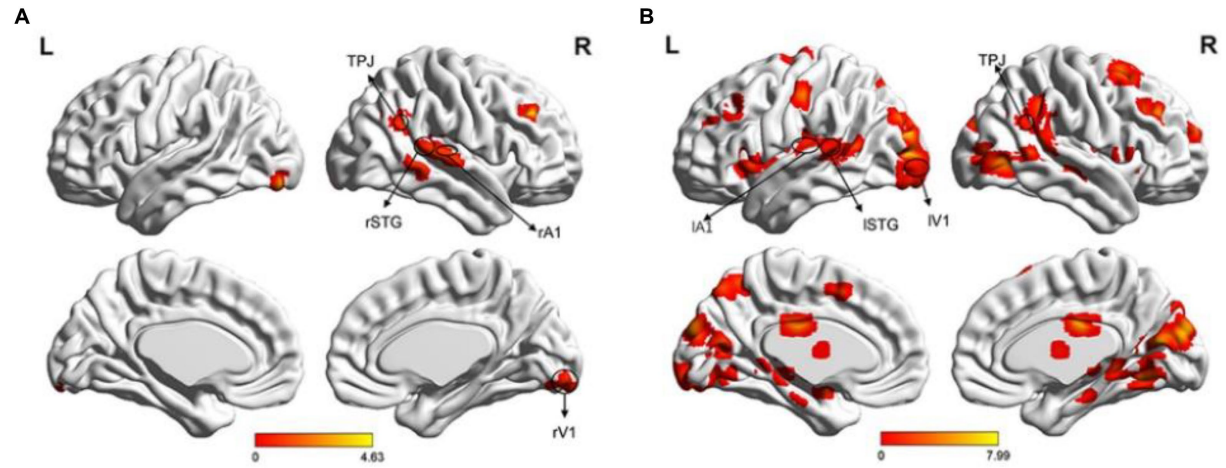



FIGURE 4
Brain activation maps. **(A)** Illustration of the activation map when the cue stimulus appeared in the LVF. The main activated areas ($p < 0.05$, FDR) included the right TPJ, right STG, right A1, right V1, and right MFG; **(B)** Illustration of the activation map when the cue stimulus appeared in the RVF ($p < 0.05$, FDR correction). The activated areas included the left STG, left A1, left V1, bilateral MFG, right TPJ, and right MT.

TABLE 1 The four selected regions of interest (ROIs) in each visual field.

| ROI | MNI coordinates (mm) | | | The size of ROI (mm) |
|---|---|---|---|---|
| | *x* | *y* | *z* | |
| Cue in the LVF | | | | |
| rTPJ | 45 | −54 | 30 | 6 |
| rSTG | 63 | −39 | 18 | 6 |
| rA1 | 45 | −21 | 12 | 10 |
| rV1 | 9 | 96 | −3 | 10 |
| Cue in the RVF | | | | |
| rTPJ | 45 | −54 | 33 | 6 |
| lSTG | −64 | −46 | 18 | 6 |
| lA1 | −40 | −26 | 14 | 10 |
| lV1 | −9 | −96 | −9 | 10 |

## 3.3. Time-varying network

We computed the time-varying network at time points ranging from 200 ms to 900 ms and displayed the connection time points only when it changed in the four conditions. When the cue appeared in the LVF or RVF, the changes in cue conditions were illustrated in Figure 5A and Figure 5B, respectively. Figure 5C summarizes the results of the time-varying network analysis. The first step for the long SOA condition was A1↔STG, whereas for the short SOA condition, it was V1↔STG. The last step for both long and short SOA conditions was V1↔STG and STG↔TPJ. Notably, in the long SOA-valid condition, V1↔STG was the middle step.

## 3.4. Correlation analysis

Our analysis revealed significant correlations between reaction and information flow (such as STG → TPJ and TPJ → STG) for all conditions, as shown in Figure 6. For the long SOA, distinct differences were observed for each condition. Negative correlations were evident when the cue was invalid and appeared in the LVF (STG → TPJ: $r = -0.54$, $p < 0.05$; TPJ → STG: $r = -0.52$, $p < 0.05$) or RVF (STG → TPJ: $r = -0.51$, $p < 0.05$; TPJ → STG: $r = -0.58$, $p < 0.05$). Conversely, positive correlations were evident when the cue was valid and appeared in either the LVF (STG → TPJ:r = 0.65, $p < 0.01$; TPJ → STG: r = 0.52, p < 0.05) or RVF (STG → TPJ:r = 0.53, p < 0.05; TPJ → STG: r = 0.57, p < 0.05). Similar trends were noted for all conditions for the short SOA, as shown in Figure 6. Positive correlations between mean RT and information flow were observed when the cue was invalid and appeared in the LVF (STG → TPJ: $r = 0.59$, $p < 0.05$; TPJ → STG: $r = 0.56$, $p < 0.05$) or RVF (STG → TPJ: $r = 0.59$, $p < 0.05$; TPJ → STG: $r = 0.56$, $p < 0.05$). Similarly, positive correlations were observed when the cue was valid, regardless of whether it appeared in the LVF (STG → TPJ:r = 0.72, $p < 0.005$ TPJ → STG: r = 0.55, p < 0.05) or RVF (STG → TPJ: r = 0.52, p < 0.05; TPJ → STG: r = 0.53, p < 0.05)

# 4. Discussions

In this study, the behavioral results showed that the RT for a valid cue was significantly shorter than an invalid cue in the short SOA condition, while the opposite was opposite for the long SOA, which was similar to the unimodal task. In both long and short SOA conditions, we observed STG activation, a critical auditory–visual integration region (Klemen and Chambers, 2012). Additionally, we observed activation in TPJ and MFG, which are important VAN areas (Corbetta et al., 2008), indicating that attention plays a role in auditory–visual integration. Our time-varying network analysis revealed that V1/A1↔STG occurred before TPJ↔STG, as shown in Figure 5, indicating that pre-attention in auditory–visual integration.

## 4.1. Similar results observed between the bimodal and unimodal cue-target paradigms

Previous researches have reported that there is a significant cue effect for short SOAs in the visual cue-target paradigm. On the

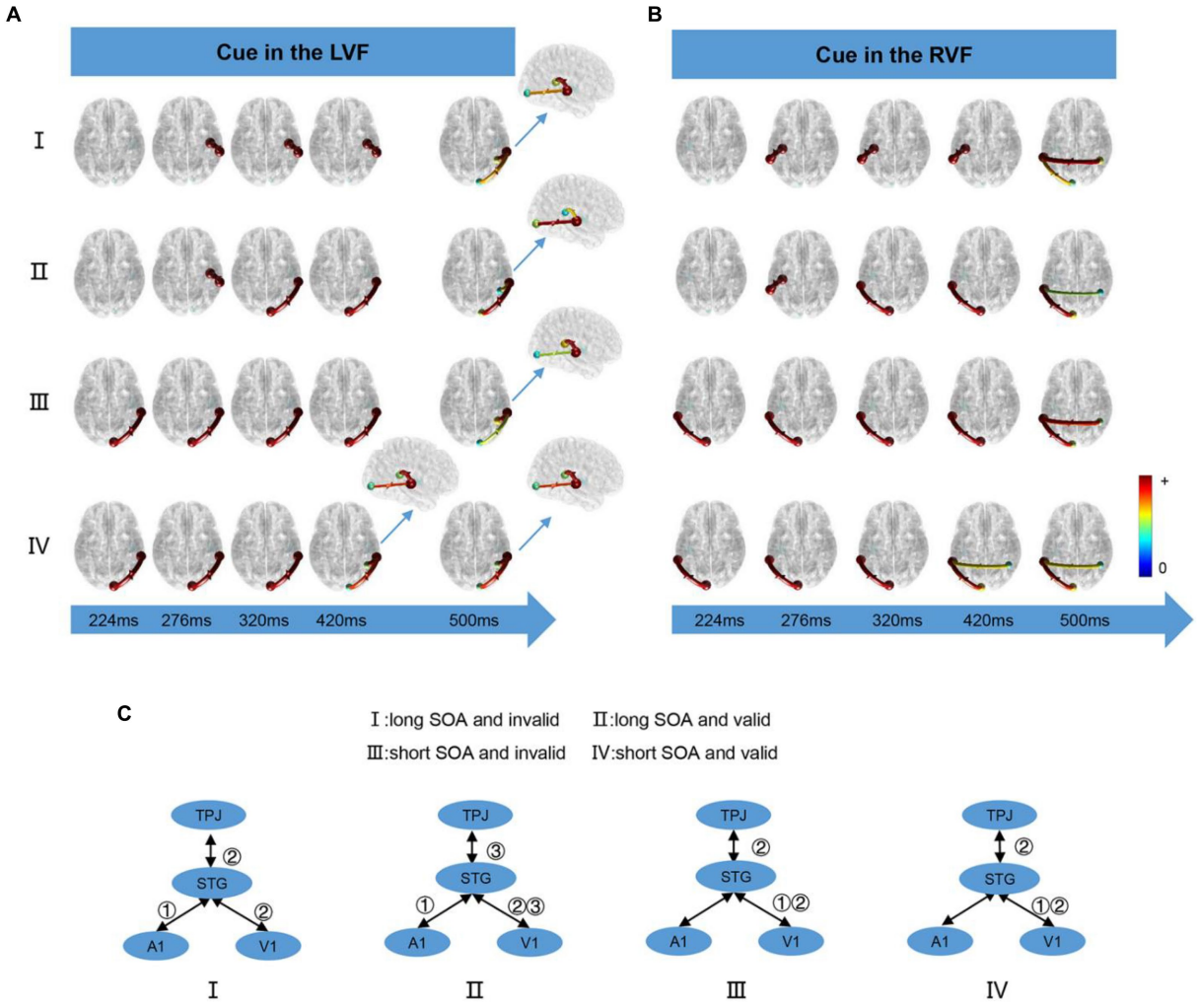



FIGURE 5
The time-varying networks when the cue stimulus appeared in the **(A)** LVF and **(B)** RVF. **(C)** Summary of time-varying networks. ①②③ denote the order in which the connections appear.

condition of the time interval of the cue and target stimulus is shorter than 300 ms, the subjects exhibited faster responses when the cue was valid as compared to when it was invalid. However, the subjects showed slower responses when the cue was valid rather than invalid for long SOA (more than 300 ms). These findings were consistent with previous studies (Lepsien and Pollmann, 2002; Mayer et al., 2004a,b; Tian and Yao, 2008; Tian et al., 2011) and suggested that stimulus-driven attention effects are faster and more transient than goal-directed attention effects (Jonides and Irwin, 1981; Shepherd and Müller, 1989; Corbetta et al., 2002; Busse et al., 2008; Macaluso et al., 2016; Tang et al., 2016). Similar outcomes have been observed in the auditory paradigm (Alho et al., 2015; Hanlon et al., 2017). Our behavioral analysis aligns with previous research on the unimodal paradigm and suggests that there is no difference between unimodal and bimodal paradigms in the cue-target paradigm.

## 4.2. Integration and attention exist in the bimodal cue-target paradigm

Previous studies have emphasized that auditory–visual integration in the cue-target paradigm occurs when the cue with one modal stimulus and the target with a different modal stimulus are presented from around the same spatial position (Stein and Meredith, 1990; Spence, 2013; Wu et al., 2020) and at approximately the same time (Stein and Meredith, 1990; Frassinetti et al., 2002; Bolognini et al., 2005; Spence and Santangelo, 2010; Stevenson et al., 2012; Tang et al., 2016). However, it will not appear if the cue precedes the target by more than 300 ms (Spence, 2010). In our paradigm, the time intervals between cue and target stimulus were divided into 100 and 800 ms, which cannot be directly compared to previous studies. Our fMRI results, where the STG appeared in all conditions, indicate that auditory and visual integration occurs even when these two stimuli are not aligned in space or time (i.e., more than 300 ms interval).

The role of attention in multisensory integration is still under debate. Some studies proposed that multisensory integration is an automatic process (Vroomen et al., 2001; José et al., 2020), while others suggested that attention played an important role in multisensory integration (Talsma et al., 2007, 2010; Fairhall and Macaluso, 2009; Tang et al., 2016). Since rTPJ and MFG important parts of the ventral attention network (Corbetta et al., 2008; Klein et al., 2021), the experimental activation of these rTPJ and MFG suggested that attention may also be involved in integration. We used time-varying networks to determine the temporal order between

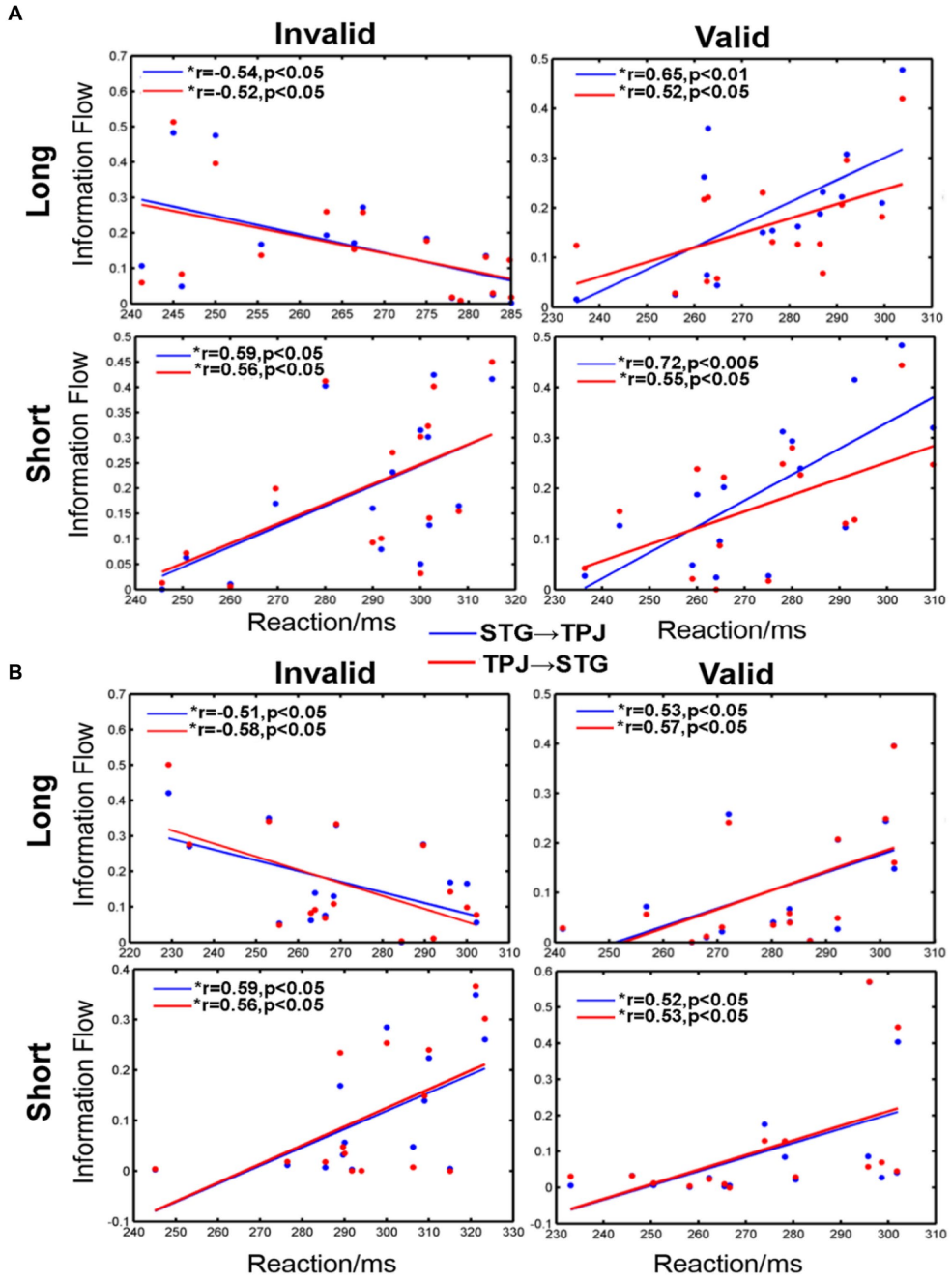



FIGURE 6
The correlation between information flow and average reaction time (RT) when the cue stimulus appeared in the **(A)** LVF and **(B)** RVF. *a significant difference between the two regions. STG → TPJ denotes the causal flow from STG to TPJ. TPJ → STG denotes the causal flow from TPJ to STG.

multisensory integration and attention using combination of fMRI and EEG data, which allowed for greater precision than EEG data alone. Additionally, the fMRI data provided a more precise spatial resolution for the time-varying networks.

## 4.3. Auditory–visual integration prior to attention

In this research, we constructed a time-varying network using task-related fMRI activations as nodes, including TPJ as the core of the VAN (Corbetta et al., 2008), and STG as an important area for integration (Yan et al., 2015). Our aim was to investigate the relationship between multisensory integration and attention. As depicted in Figure 5C, the V1/A1↔STG connection was always the first order, followed by STG↔TPJ, regardless of the conditions. This finding supports the notion that pre-attention is involved in auditory and visual integration, which is consistent with previous studies (Erik et al., 2008). However, we observed some differences under different conditions, such as the SOA length. For short SOA, the first connection was V1↔STG, as visual stimuli are dominant in processing spatial characteristics, while auditory events dominate temporal characteristic processing (Bertelson et al., 2000; Stekelenburg et al., 2004; Bonath et al., 2007; Navarra et al., 2010). Conversely, for long SOA,

information flowed from the A1 to STG. This result might be due to the auditory–visual stimuli being temporally unsynchronized in our data collection. As the SOA increased, the dominant role of the visual stimulus diminished, and the auditory effect became stronger, leading to a significant A1↔STG flow. Interestingly, TPJ↔STG and STG↔V1 were the last step in all conditions, indicating that the TPJ modulates the primary cortex by using integration areas as a transfer node in all cases.

### 4.4. Relationship between information flow and RT

Numerous studies have investigated how attention affects a subject's reaction time, but there is disagreement on whether attention boosts or limits the reflection (Senkowski et al., 2005; Karns and Knight, 2009; Macaluso et al., 2016). Some studies have suggested that attention accelerates reaction speed (Mcdonald et al., 2005, 2009; Van der Stoep et al., 2017), while others have proposed that attention may actually inhibit reaction (Tian and Yao, 2008). A recent study has demonstrated that both stimulus-driven attention and multisensory integration can accelerate responses (Van der Stoep et al., 2017; Motomura and Amimoto, 2022).

In this study, we compared the correlation between mean RT and the information flow of STG↔TPJ under different circumstances. Our findings suggest that attention has a direct influence on multisensory integration, as the extent of information flow reflects the mutual influence of the two brain regions. Specifically, we observed a negative correlation between the two regions in the long SOA-invalid condition, indicating that larger information flow led to faster reflection times. We inferred that this phenomenon is due to bottom-up attention, where increased information flow leads to greater information exchange between the STG and TPJ and, thus, faster reactions. However, in other conditions, we observed positive correlations, which we attribute to the modulation of attention. Specifically, greater attention modulation results in inhibited reactions.

## 5. Conclusion

In this paper, our analysis of the behavioral data showed no discernible difference between the multisensory and unisensory cue-target paradigms. We also employed fMRI data analysis to demonstrate the existence of auditory–visual integration in the long SOA condition and the necessity of attention for such integration. The constructed time-varying networks based on fMRI coordinates revealed that multisensory integration occurs prior to attention and pre-attention is involved in auditory–visual integration. Furthermore, our findings suggest that attention can impact the subject's reaction time, but the effect depends on the situation, and greater attention modulation results in inhibited reactions.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the University of Electronic Science and Technology of China. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

YJ performed experiments and data analysis. RQ and YS contributed to data analysis. YTa wrote the draft of the manuscript. ZH and YTi contributed to experiments design and conception. All the authors edited the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was sponsored in part by the National Natural Science Foundation of China (Grant No. 62171074), China Postdoctoral Science Foundation (No. 2021MD703941), special support for Chongqing postdoctoral research project (2021XM2051), Natural Science Foundation of Chongqing (cstc2019jcyj-msxmX0275), Project of Central Nervous System Drug Key Laboratory of Sichuan Province (200028-01SZ), the Doctoral Foundation of Chongqing University of Posts and Telecommunications (A2022-11), and in part by Postdoctoral Science Foundation of Chongqing (cstc2021jcyj-bshX0181), in part by Chongqing Municipal Education Commission (21SKGH068).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Abreu, R., Leal, A., and Figueiredo, P. (2018). EEG-informed fMRI: a review of data analysis methods. *Front. Hum. Neurosci.* 12:29. doi: 10.3389/fnhum.2018.00029

Alho, K., Salmi, J., Koistinen, S., Salonen, O., and Rinne, T. (2015). Top-down controlled and bottom-up triggered orienting of auditory attention to pitch activate overlapping brain networks. *Brain Res.* 1626, 136–145. doi: 10.1016/j.brainres.2014.12.050

Baenninger, A., Diaz Hernandez, L., Rieger, K., Ford, J. M., Kottlow, M., and Koenig, T. (2016). Inefficient preparatory fMRI-BOLD network activations predict working

memory dysfunctions in patients with schizophrenia. *Front. Psych.* 7:29. doi: 10.3389/fpsyt.2016.00029

Bagshaw, A. P., Hale, J. R., Campos, B. M., Rollings, D. T., Wilson, R. S., Alvim, M. K. M., et al. (2017). Sleep onset uncovers thalamic abnormalities in patients with idiopathic generalised epilepsy. *Neuroimage Clin.* 16, 52–57. doi: 10.1016/j.nicl.2017.07.008

Bertelson, P. (1999). Ventriloquism: a case of crossmodal perceptual grouping. *Adv. Psychol.* 129, 347–362. doi: 10.1016/S0166-4115(99)80034-X

Bertelson, P., Vroomen, J., De, G. B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept. Psychophys.* 62, 321–332. doi: 10.3758/BF03205552

Bolognini, N., Frassinetti, F., Serino, A., and Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs. *Exp. Brain Res.* 160, 273–282. doi: 10.1007/s00221-004-2005-z

Bonath, B., Noesselt, T., Martinez, A., Mishra, J., Schwiecker, K., Heinze, H. J., et al. (2007). Neural basis of the ventriloquist illusion. *Curr. Biol.* 17, 1697–1703. doi: 10.1016/j.cub.2007.08.050

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., and Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *J. Neurosci.* 34, 4548–4557. doi: 10.1523/JNEUROSCI.4339-13.2014

Branden, J. B., Arvid, G., Mark, P., and Andrew, I. W. (2022). Michael SAG right temporoparietal junction encodes inferred visual knowledge of others. *Neuropsychologia* 171:108243. doi: 10.1016/j.neuropsychologia.2022.108243

Bullock, M., Jackson, G. D., and Abbott, D. F. (2021). Artifact reduction in simultaneous EEG-fMRI: a systematic review of methods and contemporary usage. *Front. Neurol.* 12:622719. doi: 10.3389/fneur.2021.622719

Busse, L., Katzner, S., and Treue, S. (2008). Temporal dynamics of neuronal modulation during exogenous and endogenous shifts of visual attention in macaque area MT. *Proc. Natl. Acad. Sci.* 105, 16380–16385. doi: 10.1073/pnas.0707369105

Calvert, G. A., Campbell, R., and Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Curr. Biol.* 10, 649–657. doi: 10.1016/s0960-9822(00)00513-3

Calvert, G. A., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *NeuroImage* 14, 427–438. doi: 10.1006/nimg.2001.0812

Calvert, G. A., and Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *J. Physiol. Paris* 98, 191–205. doi: 10.1016/j.jphysparis.2004.03.018

Cappe, C., Thut, G., Romei, V., and Murray, M. M. (2010). Auditory-visual multisensory interactions in humans: timing, topography, directionality, and sources. *J. Neurosci.* 30, 12572–12580. doi: 10.1523/JNEUROSCI.1099-10.2010

Chen, T., Michels, L., Supekar, K., Kochalka, J., Ryali, S., and Menon, V. (2015). Role of the anterior insular cortex in integrative causal signaling during multisensory auditory-visual attention. *Eur. J. Neurosci.* 41, 264–274. doi: 10.1111/ejn.12764

Corbetta, M., Kincade, J. M., and Shulman, G. L. (2002). Neural systems for visual orienting and their relationships to spatial working memory. *J. Cogn. Neurosci.* 14, 508–523. doi: 10.1162/089892902317362029

Corbetta, M., Patel, G., and Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron* 58, 306–324. doi: 10.1016/j.neuron.2008.04.017

Durk, T., and Woldorff, M. G. (2005). Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114. doi: 10.1162/0898929054475172

Erik, V. D. B., Olivers, C. N. L., Bronkhorst, A. W., and Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1053–1065. doi: 10.1037/0096-1523.34.5.1053

Fairhall, S. L., and Macaluso, E. (2009). Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29, 1247–1257. doi: 10.1111/j.1460-9568.2009.06688.x

Ford, J. M., Roach, B. J., Palzes, V. A., and Mathalon, D. H. (2016). Using concurrent EEG and fMRI to probe the state of the brain in schizophrenia. *NeuroImage Clin.* 12, 429–441. doi: 10.1016/j.nicl.2016.08.009

Frassinetti, F., Bolognini, N., and Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp. Brain Res.* 147, 332–343. doi: 10.1007/s00221-002-1262-y

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., et al. (1995). Analysis of fMRI time-series revisited. *NeuroImage* 2, 45–53. doi: 10.1006/nimg.1995.1007

Hanlon, F. M., Dodd, A. B., Ling, J. M., Bustillo, J. R., Abbott, C. C., and Mayer, A. R. (2017). From behavioral facilitation to inhibition: the neuronal correlates of the orienting and reorienting of auditory attention. *Front. Hum. Neurosci.* 11:293. doi: 10.3389/fnhum.2017.00293

Hsiao, F. C., Tsai, P. J., Wu, C. W., Yang, C. M., Lane, T. J., Lee, H. C., et al. (2018). The neurophysiological basis of the discrepancy between objective and subjective sleep during the sleep onset period: an EEG-fMRI study. *Sleep* 41:zsy056. doi: 10.1093/sleep/zsy056

Jonides, J., and Irwin, D. E. (1981). Capturing attention. *Cognition* 10, 145–150. doi: 10.1016/0010-0277(81)90038-X

Jorge, J., Zwaag, W. V. D., and Figueiredo, P. (2014). EEG–fMRI integration for the study of human brain function. *NeuroImage* 102, 24–34. doi: 10.1016/j.neuroimage.2013.05.114

José, P., Ossandón, K. P., and Heed, T. (2020). No evidence for a role of spatially modulated alpha-band activity in tactile remapping and short-latency, overt orienting behavior. *J. Neurosci.* 40, 9088–9102. doi: 10.1523/JNEUROSCI.0581-19.2020

Kang-jia, J., and Xu, W. (2022). Research and Prospect of visual event-related potential in traumatic brain injury and visual function evaluation. *J. Forensic Med.* 40, 9088–9102. doi: 10.1523/JNEUROSCI.0581-19.2020

Karns, C. M., and Knight, R. T. (2009). Intermodal auditory, visual, and tactile attention modulates early stages of neural processing. *J. Cogn. Neurosci.* 21, 669–683. doi: 10.1162/jocn.2009.21037

Klein, H. S., Vanneste, S., and Pinkham, A. E. (2021). The limited effect of neural stimulation on visual attention and social cognition in individuals with schizophrenia. *Neuropsychologia* 157:107880. doi: 10.1016/j.neuropsychologia.2021.107880

Klemen, J., and Chambers, C. D. (2012). Current perspectives and methods in studying neural mechanisms of multisensory interactions. *Neurosci. Biobehav. Rev.* 36, 111–133. doi: 10.1016/j.neubiorev.2011.04.015

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol.* 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Laura, B., Roberts, K. C., Crist, R. E., Weissman, D. H., and Woldorff, M. G. (2005). The spread of attention across modalities and space in a multisensory object. *Proc. Natl. Acad. Sci.* 102, 18751–18756. doi: 10.1073/pnas.0507704102

Lepsien, J., and Pollmann, S. (2002). Covert reorienting and inhibition of return: an event-related fMRI study. *J. Cogn. Neurosci.* 14, 127–144. doi: 10.1162/089892902317236795

Li, F., Chen, B., Li, H., Zhang, T., Wang, F., Jiang, Y., et al. (2016). The time-varying networks in P300: a task-evoked EEG study. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 725–733. doi: 10.1109/TNSRE.2016.2523678

Li, F., Liu, T., Wang, F., Li, H., Gong, D., Zhang, R., et al. (2015). Relationships between the resting-state network and the P3: evidence from a scalp EEG study. *Sci. Rep.* 5:15129. doi: 10.1038/srep15129

Ľuboš, H., Aaron, R., Seitz, , and Norbert, K. (2021). Auditory-visual interactions in egocentric distance perception: ventriloquism effect and aftereffecta. *J. Acoust. Soc. Am.* 150, 3593–3607. doi: 10.1121/10.0007066

Macaluso, E., George, N., Dolan, R., Spence, C., and Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage* 21, 725–732. doi: 10.1016/j.neuroimage.2003.09.049

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., and Adam, R. (2016). The curious incident of attention in multisensory integration: bottom-up vs Top-down. *Multisens. Res.* 29, 557–583. doi: 10.1163/22134808-00002528

Mastroberardino, S., Santangelo, V., and Macaluso, E. (2015). Crossmodal semantic congruence can affect visuo-spatial processing and activity of the fronto-parietal attention networks. *Front. Integr. Neurosci.* 9:45. doi: 10.3389/fnint.2015.00045

Mayer, A. R., Dorflinger, J. M., Rao, S. M., and Seidenberg, M. (2004a). Neural networks underlying endogenous and exogenous visual-spatial orienting. *NeuroImage* 23, 534–541. doi: 10.1016/j.neuroimage.2004.06.027

Mayer, A., Seidenberg, M., Dorflinger, J., and Rao, S. (2004b). An event-related fMRI study of exogenous orienting: supporting evidence for the cortical basis of inhibition of return? *J. Cogn. Neurosci.* 16, 1262–1271. doi: 10.1162/0898929041920531

Mcdonald, J. J., Hickey, C., Green, J. J., and Whitman, J. C. (2009). Inhibition of return in the covert deployment of attention: evidence from human electrophysiology. *J. Cogn. Neurosci.* 21, 725–733. doi: 10.1162/jocn.2009.21042

Mcdonald, J. J., Teder-Sälejärvi, W. A., Di, R. F., and Hillyard, S. A. (2005). Neural basis of auditory-induced shifts in visual time-order perception. *Nat. Neurosci.* 8, 1197–1202. doi: 10.1038/nn1512

Mcdonald, J. J., Teder-sälejärvi, W. A., and Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature* 407, 906–908. doi: 10.1038/35038085

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Motomura, K., and Amimoto, K. (2022). Development of stimulus-driven attention test for unilateral spatial neglect – accuracy, reliability, and validity. *Neurosci. Lett.* 772:136461. doi: 10.1016/j.neulet.2022.136461

Muraskin, J., Brown, T. R., Walz, J. M., Tu, T., Conroy, B., Goldman, R. I., et al. (2018). A multimodal encoding model applied to imaging decision-related neural cascades in the human brain. *NeuroImage* 180, 211–222. doi: 10.1016/j.neuroimage.2017.06.059

Navarra, J., Soto-Faraco, S., and Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Inf. Fusion* 11, 4–11. doi: 10.1016/j.inffus.2009.04.001

Nazir, H. M., Hussain, I., Faisal, M., Mohamd Shoukry, A., Abdel Wahab Sharkawy, M., Fawzi Al-Deek, F., et al. (2020). Dependence structure analysis of multisite river inflow

data using vine copula-CEEMDAN based hybrid model. *PeerJ* 8:e10285. doi: 10.7717/peerj.10285

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27, 11431–11441. doi: 10.1523/JNEUROSCI.2252-07.2007

Pisauro, M. A., Fouragnan, E., Retzler, C., and Philiastides, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI. *Nat. Commun.* 8:15808. doi: 10.1038/ncomms15808

Platt, B. B., and Warren, D. H. (1972). Auditory localization: the importance of eye movements and a textured visual environment. *Percept. Psychophys.* 12, 245–248. doi: 10.3758/BF03212884

Posner, M. I., and Rothbart, M. K. (2006). Research on attention networks as a model for the integration of psychological science. *Annu. Rev. Psychol.* 58, 1–23. doi: 10.1146/annurev.psych.58.110405.085516

Rachel, W., Kathleen, B., and Mark, A. (2022). Evaluating the co-design of an age-friendly, rural, multidisciplinary primary care model: a study protocol. *Methods Protoc.* 5:23. doi: 10.3390/mps5020023

Ristic, J., and Capozzi, F. (2023). The role of visual and auditory information in social event segmentation. *Q. J. Exp. Psychol.* 1:174702182311764. doi: 10.1177/17470218231176471

Romei, V., Murray, M. M., Cappe, C., and Thut, G. (2013). The contributions of sensory dominance and attentional Bias to cross-modal enhancement of visual cortex excitability. *J. Cogn. Neurosci.* 25, 1122–1135. doi: 10.1162/jocn_a_00367

Rupp, K., Hect, J. L., Remick, M., Ghuman, A., Chandrasekaran, B., Holt, L. L., et al. (2022). Neural responses in human superior temporal cortex support coding of voice representations. *PLoS Biol.* 20:e3001675. doi: 10.1371/journal.pbio.3001675

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sébastien, A. L., Arin, E. A., Kristina, C., Blake, E. B., and Ryan, A. S. (2022). The relationship between multisensory associative learning and multisensory integration. *Neuropsychologia* 174:108336. doi: 10.1016/j.neuropsychologia.2022.108336

Senkowski, D., Talsma, D., Herrmann, C. S., and Woldorff, M. G. (2005). Multisensory processing and oscillatory gamma responses: effects of spatial selective attention. *Exp. Brain Res.* 166, 411–426. doi: 10.1007/s00221-005-2381-z

Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. *Nature* 408, 788–2671. doi: 10.1038/35048669

Shams, N., Alain, C., and Strother, S. (2015). Comparison of BCG artifact removal methods for evoked responses in simultaneous EEG-fMRI. *J. Neurosci. Methods* 245, 137–146. doi: 10.1016/j.jneumeth.2015.02.018

Shepherd, M., and Müller, H. J. (1989). Movement versus focusing of visual attention. *Percept. Psychophys.* 46, 146–154. doi: 10.3758/BF03204974

Spence, C. (2010). Crossmodal spatial attention. *Ann. N. Y. Acad. Sci.* 1191, 182–200. doi: 10.1111/j.1749-6632.2010.05440.x

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Ann. N. Y. Acad. Sci.* 1296, 31–49. doi: 10.1111/nyas.12121

Spence, C., and Santangelo, V. (2010). Auditory development and learning. *Oxf. Handb. Audit. Sci.* 3:249. doi: 10.1093/oxfordhb/9780199233557.013.0013

Stein, B. E., and Meredith, M. A. (1990). Multisensory integration. Neural and behavioral solutions for dealing with stimuli from different sensory modalities. *Ann. N. Y. Acad. Sci.* 608, 51–70. doi: 10.1111/j.1749-6632.1990.tb48891.x

Stein, B. E., Meredith, M. A., Huneycutt, W. S., and Mcdade, L. (1989). Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli. *J. Cogn. Neurosci.* 1, 12–24. doi: 10.1162/jocn.1989.1.1.12

Stekelenburg, J. J., Vroomen, J., and De, G. B. (2004). Illusory sound shifts induced by the ventriloquist illusion evoke the mismatch negativity. *Neurosci. Lett.* 357, 163–166. doi: 10.1016/j.neulet.2003.12.085

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., and Wallace, M. T. (2012). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. *Exp. Brain Res.* 219, 121–137. doi: 10.1007/s00221-012-3072-1

Stoep, N., Der, V., Stigchel, S., Der, V., and Nijboer, T. C. W. (2015). Exogenous spatial attention decreases audiovisual integration. *Atten. Percept. Psychophys.* 77, 464–482. doi: 10.3758/s13414-014-0785-1

Talsma, D., Doty, T. J., and Woldorff, M. G. (2007). Selective attention and audivisual integration: is attending to both modalities a prerequisite for optimal early integration? *Cereb. Cortex* 17, 691–701. doi: 10.1093/cercor/bhk020

Talsma, D., Senkowski, D. F. S., and Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14, 400–410. doi: 10.1016/j.tics.2010.06.008

Tang, X., Wu, J., and Shen, Y. (2016). The interactions of multisensory integration with endogenous andexogenous attention. *Neurosci. Biobehav. Rev.* 61, 208–224. doi: 10.1016/j.neubiorev.2015.11.002

Tedersälejärvi, W. A., Di, R. F., Mcdonald, J. J., and Hillyard, S. A. (2005). Effects of spatial congruity on audio-visual multimodal integration. *J. Cogn. Neurosci.* 17, 1396–1409. doi: 10.1162/0898929054985383

Tian, Y., Klein, R. M., Satel, J., Xu, P., and Yao, D. (2011). Electrophysiological explorations of the cause and effect of inhibition of return in a Cue-target paradigm. *Brain Topogr.* 24, 164–182. doi: 10.1007/s10548-011-0172-3

Tian, Y., Liang, S., and Yao, D. (2014). Attentional orienting and response inhibition: insights from spatial-temporal neuroimaging. *Neurosci. Bull.* 30, 141–152. doi: 10.1007/s12264-013-1372-5

Tian, Y., and Ma, L. (2020). Auditory attention tracking states in a cocktail party environment can be decoded by deep convolutional neural networks. *J. Neural Eng.* 17:036013. doi: 10.1088/1741-2552/ab92b2

Tian, Y., Xu, W., and Yang, L. (2018a). Cortical classification with rhythm entropy for error processing in cocktail party environment based on scalp EEG recording. *Sci. Rep.* 8:6070. doi: 10.1038/s41598-018-24535-4

Tian, Y., Xu, W., Zhang, H., Tam, K. Y., Zhang, H., Yang, L., et al. (2018b). The scalp time-varying networks of N170: reference, latency, and information flow. *Front. Neurosci.* 12:250. doi: 10.3389/fnins.2018.00250

Tian, Y., and Yao, D. (2008). A study on the neural mechanism of inhibition of return by the event-related potential in the go/Nogo task. *Biol. Psychol.* 79, 171–178. doi: 10.1016/j.biopsycho.2008.04.006

Tian, Y., and Yao, D. (2013). Why do we need to use a zero reference? Reference influences on the ERPs of audiovisual effects. *Psychophysiology* 50, 1282–1290. doi: 10.1111/psyp.12130

Treichler, D. (1967). Are you missing the boat in training aids. *Film AV Commun.* 1, 14–16.

Tucker, D. M. (1993). Spatial sampling of head electrical fields: the geodesic sensor net. *Electroencephalogr. Clin. Neurophysiol.* 87, 154–163. doi: 10.1016/0013-4694(93)90121-B

Van der Stoep, N., Van der Stigchel, S., Nijboer, T. C., and Spence, C. (2017). Visually induced inhibition of return affects the integration of auditory and visual information. *Perception* 46, 6–17. doi: 10.1177/0301006616661934

Vroomen, J., Bertelson, P., and Gelder, B. D. (2001). Directing spatial attention towards the illusory location of a ventriloquized sound. *Acta Psychol.* 108, 21–33. doi: 10.1016/S0001-6918(00)00068-8

Wang, K., Li, W., and Dong, L. (2018). Clustering-constrained ICA for Ballistocardiogram artifacts removal in simultaneous EEG-fMRI. *Front. Neurosci.* 12:59. doi: 10.3389/fnins.2018.00059

Wang, P., Fuentes, L. J., Vivas, A. B., and Chen, Q. (2013). Behavioral and neural interaction between spatial inhibition of return and the Simon effect. *Front. Hum. Neurosci.* 7:572. doi: 10.3389/fnhum.2013.00572

Wang, W., Wang, Y., Sun, J., Liu, Q., Liang, J., and Li, T. (2020). Speech driven talking head generation via attentional landmarks based representation. *Proc. Interspeech* 2020, 1326–1330. doi: 10.21437/Interspeech.2020-2304

Wen, H., You, S., and Fu, Y. (2021). Cross-modal dynamic convolution for multi-modal emotion recognition. *J. Vis. Commun. Image Represent.* 78:103178. doi: 10.1016/j.jvcir.2021.103178

Wilke, C., Ding, L., and He, B. (2008). Estimation of time-varying connectivity patterns through the use of an adaptive directed transfer function. *IEEE Trans. Biomed. Eng.* 55, 2557–2564. doi: 10.1109/TBME.2008.919885

Wu, X., Wang, A., and Zhang, M. (2020). Cross-modal nonspatial repetition inhibition: an ERP study, neuroscience letters. *ISSN* 734:135096. doi: 10.1016/j.neulet.2020.135096

Xu, Z., Yang, W., and Zhou, Z. (2020). Cue–target onset asynchrony modulates interaction between exogenous attention and audiovisual integration. *Cogn. Process.* 21, 261–270. doi: 10.1007/s10339-020-00950-2

Yan, T., Geng, Y., Wu, J., and Li, C. (2015). Interactions between multisensory inputs with voluntary spatial attention: an fMRI study. *Neuroreport* 26, 605–612. doi: 10.1097/WNR.0000000000000368

Yao, D., Li, W., Oostenveld, R., Nielsen, K. D., Arendt-Nielsen, L., and Chen, C. A. N. (2005). A comparative study of different references for EEG spectral mapping: the issue of the neutral reference and the use of the infinity reference. *Physiol. Meas.* 26, 173–184. doi: 10.1088/0967-3334/26/3/003

Zhang, J., Li, Y., Li, T., Xun, L., and Shan, C. (2019). License plate localization in unconstrained scenes using a two-stage CNN-RNN. *IEEE Sensors J.* 19, 5256–5265. doi: 10.1109/JSEN.2019.2900257

Zhang, T., Gao, Y., and Hu, S. (2022). Focused attention: its key role in gaze and arrow cues for determining where attention is directed. *Psychol. Res.* 2022, 1–15. doi: 10.1007/s00426-022-01781-w

Zhang, X., Wang, Z., and Liu, T. (2022). Anti-disturbance integrated position synchronous control of a dual permanent magnet synchronous motor system. *Energies* 15:6697. doi: 10.3390/en15186697

Zhang, X., Zhang, J., and Liu, G. (2021). Comprehensive framework for the integration and analysis of geo-environmental data for urban geohazards. *Earth Sci. Inf.* 14, 2387–2399. doi: 10.1007/s12145-021-00642-1

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 10 August 2023
DOI 10.3389/fnins.2023.1243847



OPEN ACCESS

EDITED BY
Qieshi Zhang,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Ziliang Ren,
Dongguan University of Technology, China
Jing Ji,
Xidian University, China
Ketsuseki Cyou,
Waseda University, Japan

*CORRESPONDENCE
Lilan Liu
lancy@shu.edu.cn

RECEIVED 21 June 2023
ACCEPTED 31 July 2023
PUBLISHED 10 August 2023

CITATION
Sun J, Su J, Yan Z, Gao Z, Sun Y and Liu L (2023) Truck model recognition for an automatic overload detection system based on the improved MMAL-Net.
*Front. Neurosci.* 17:1243847.
doi: 10.3389/fnins.2023.1243847



# Truck model recognition for an automatic overload detection system based on the improved MMAL-Net

Jiachen Sun[1], Jin Su[2], Zhenhao Yan[1], Zenggui Gao[1], Yanning Sun[1] and Lilan Liu[1]*

[1]Shanghai Key Laboratory of Intelligent Manufacturing and Robotics, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China, [2]College of Information Engineering, Lanzhou University of Finance and Economics, Lanzhou, China

Efficient and reliable transportation of goods through trucks is crucial for road logistics. However, the overloading of trucks poses serious challenges to road infrastructure and traffic safety. Detecting and preventing truck overloading is of utmost importance for maintaining road conditions and ensuring the safety of both road users and goods transported. This paper introduces a novel method for detecting truck overloading. The method utilizes the improved MMAL-Net for truck model recognition. Vehicle identification involves using frontal and side truck images, while APPM is applied for local segmentation of the side image to recognize individual parts. The proposed method analyzes the captured images to precisely identify the models of trucks passing through automatic weighing stations on the highway. The improved MMAL-Net achieved an accuracy of 95.03% on the competitive benchmark dataset, Stanford Cars, demonstrating its superiority over other established methods. Furthermore, our method also demonstrated outstanding performance on a small-scale dataset. In our experimental evaluation, our method achieved a recognition accuracy of 85% when the training set consisted of 20 sets of photos, and it reached 100% as the training set gradually increased to 50 sets of samples. Through the integration of this recognition system with weight data obtained from weighing stations and license plates information, the method enables real-time assessment of truck overloading. The implementation of the proposed method is of vital importance for multiple aspects related to road traffic safety.

KEYWORDS

**overload detection, truck model recognition, automatic weighing station, fine-grained visual categorization, MMAL-Net**

## 1. Introduction

With the rapid development of the global economy and the acceleration of urbanization processes, highways play a crucial role in connecting different regions and cities. In the realm of road transportation, trucks serve as vital transportation tools, undertaking the task of transporting a substantial amount of goods. However, the issue of truck overloading has become one of the primary challenges in road traffic safety and road damage. Overloaded trucks exert significant pressure on road infrastructure, increasing the risk of traffic accidents and potentially

leading to severe road collapse incidents. Therefore, the development of an accurate and efficient truckload monitoring method holds significant practical significance.

Traditional methods for truckload monitoring mainly rely on static weight measurement equipment such as weighbridges (Zhou et al., 2005) and fixed scales. However, these devices have several limitations, including the need for trucks to stop for measurement and high time and labor costs. Additionally, static measurement methods cannot provide real-time monitoring and detection capabilities for violations, limiting their effectiveness in practical applications.

To address these challenges, a promising solution has emerged: utilizing camera images from highway weigh stations for truck model recognition and combining them with weighing information obtained from a dynamic weighing system. By leveraging truck photos captured near these stations and employing advanced image processing and pattern recognition techniques, truck models can be identified accurately. Regrettably, the current network architectures utilized for recognition in this context often exhibit a simplistic nature, leading to suboptimal accuracy in the identification process. As a consequence, determining whether a truck is overloaded becomes inaccurate. Moreover, the ability to perform real-time recognition using captured photographs poses an unresolved challenge that demands urgent attention.

This paper aims to propose a truck model recognition method based on highway automatic weighing station camera images, with the objective of accurately identifying truck models. Consequently, the maximum load capacity of the trucks is determined. Through the integration of license plates and weighing information, the system can accurately determine if a truck is carrying excessive load. By doing so, it can prevent the occurrence of misjudgments caused by incidents of license plates damage, which are likely to happen in schemes that rely solely on license plates recognition to obtain vehicle models. Through this method, precise truck information can be provided for freight management, transportation safety, and highway planning, promoting the development of the logistics industry and enhancing traffic safety.

## 2. Literature review

### 2.1. FGVC

Fine-Grained Visual Categorization (FGVC) refers to the task of classifying objects into different subcategories or fine-grained classes within a broader category. In FGVC, the goal is to achieve detailed discrimination and classification among visually similar objects, such as different species of birds, breeds of dogs, or models of cars. This field of research focuses on developing computer vision algorithms and techniques to accurately recognize and classify objects at a fine-grained level, where subtle differences between subclasses need to be distinguished.

Fine-grained image classification, in contrast to conventional image classification tasks, encompasses a low signal-to-noise ratio, restricting the presence of highly discriminating information to minuscule local regions. Thus, the crux of achieving success in fine-grained image classification algorithms lies in the identification and efficient utilization of these valuable local region insights. Presently, most classification algorithms adhere to a common workflow: initial localization of the foreground object and its distinct local regions, followed by individual feature extraction from these regions. The processed features are subsequently utilized for classifier training and prediction purposes. To attain satisfactory classification results, numerous existing algorithms heavily depend on manual annotation information (Wei et al., 2018), such as bounding boxes and part locations. The annotation frame aids in foreground object detection, effectively mitigating background noise interference. Local region positions serve to identify valuable regions or align perspectives, facilitating the extraction of local features. Nevertheless, the costly acquisition of manual annotation information severely limits the practicality of these classification algorithms. In recent years, an increasing number of studies have opted to exclude such labeling information, relying solely on labels to accomplish image classification tasks (Lin et al., 2005; Zhang et al., 2016), resulting in commendable outcomes.

In the research and development of FGVC, traditional classification algorithms based on handcrafted features were initially employed. These algorithms typically begin by extracting local features, such as Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005), from the images. Subsequently, an encoding model like Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2010) is employed for feature encoding, resulting in the desired feature representation. However, the limited descriptive power of handcrafted features often leads to suboptimal classification performance. In the early stages of fine-grained visual categorization research, the representation capacity of features became a primary bottleneck hindering performance improvement.

In recent years, Convolutional Neural Network (CNN)-based methods for fine-grained image recognition (Wang et al., 2017; Xie et al., 2017) have significantly matured. Donahue et al. (2014) conducted an analysis of a CNN model trained on the ImageNet dataset, revealing that the features extracted from the CNN possess more robust semantic characteristics and exhibit superior differentiation compared to artificial features. Building upon these findings, the researchers applied the convolution features to various domain-specific tasks, including fine-grained classification, resulting in improved classification performance. Nevertheless, the crucial components of the task tend to be subtle and were not adequately captured by conventional CNN approaches. Consequently, researchers have directed their attention toward internal enhancements within the framework. Zhang et al. (2014) introduced the Part R-CNN algorithm, which leverages R-CNN (Girshick et al., 2014) for image detection. This methodology aims to achieve precise localization of crucial components and enhance feature representation. Branson et al. (2014) proposed the Pose Normalized Convolutional Neural Network (Pose Normalized CNN) algorithm. Their approach comprises several steps: localization detection is performed on local regions for each input image, followed by cropping the image based on the detected annotation boxes, extracting hierarchical local information, and conducting pose alignment. Subsequently, distinct layers of convolutional features are extracted for different body parts. Finally, these convolutional features are concatenated into a feature vector and utilized for SVM model training. These approaches have demonstrated robust feature representation capabilities and yielded promising results in fine-grained image recognition tasks.

Compared to regular classification tasks, acquiring fine-grained image databases poses greater challenges and requires stronger domain expertise for data collection and annotation. However, in recent years, there has been a significant increase in the availability of fine-grained image databases, which reflects the flourishing development trend and strong real-world demand in this field. Currently, commonly used fine-grained image databases include (1) CUB200-2011: It comprises a total of 11,788 bird images belonging to 200 different categories. This database provides rich manual annotations, including 15 local part locations, 312 binary attributes, 1 bounding box, and semantic segmentation images, (2) Stanford Dogs: This database offers a collection of 20,580 images featuring 120 different breeds of dogs. It provides only bounding box annotations, (3) Oxford Flowers: This database is divided into two scales, containing 17 and 102 categories of flowers, respectively. The 102-category database is more commonly used, with each category containing 40 to 258 images. In total, there are 8,189 images in this database, which provides only semantic segmentation images without any additional annotations, (4) Cars: This database provides a collection of 16,185 vehicle images belonging to 196 different categories, encompassing various brands, years, and models. Only bounding box annotations are provided, and (5) FGVC-Aircraft: This database consists of 10,200 images of 102 different aircraft categories, with each category containing 100 distinct photos. Only bounding box annotations are provided. In recent years, extensive research has been conducted on fine-grained image databases. DCL (Chen et al., 2019) employed a deconstruction and reconstruction approach to learn semantic correlations among local regions in input images. API-Net (Zhuang et al., 2020) progressively recognized pairs of fine-grained images through iterative interaction. GCP (Song et al., 2022) introduced a dedicated network branch to magnify the importance of small eigenvalues. MSHQP (Tan et al., 2022) effectively modeled intra and inter-layer feature interactions, integrating multi-layer features to enhance part responses. These methods primarily focus on locating and utilizing key regions for final recognition, yielding promising performance. However, they tend to overlook the potential contribution of complementary regions that can also play a positive role in the recognition process.

## 2.2. Vehicle recognition and classification

Vehicle recognition and classification are essential components of FGVC field. In the context of vehicles, this entails distinguishing between closely related classes such as different car models, brands, and types, where subtle visual differences in features become crucial for accurate classification. Currently, research on vehicle recognition and classification primarily centers around three main approaches: pattern recognition based on matching method, pattern recognition based on machine learning and pattern recognition based on deep learning.

The first approach involves the identification of vehicles through license plates and vehicle tag detection using a matching method. While the license plate number and label characteristics can directly identify the vehicle's brand and model (Psyllos et al., 2010; Huang et al., 2015), this method has a limitation: it does not encompass all the fine-grained features associated with the vehicle brand and model. Apart from the license plates and labels, vehicle lights and other textural information also bear the characteristics of the vehicle model. Relying solely on license plates and tags is insufficient. Additionally, the license plates of trucks are prone to being contaminated by dirt and dust, which leads to reduced visibility and clarity. In such scenarios, this method becomes ineffective.

The second approach involves using machine learning to classify vehicle brands and models. The traditional machine learning method comprises two steps: feature extraction and classifier classification. Fraz et al. proposed a method for recognizing vehicle brands and models based on a SIFT feature dictionary (Fraz et al., 2014). In this method, SIFT features of pictures from the training set's vehicles were treated as "words" to create a dictionary of vehicle brands and models. However, this method necessitates extensive computation and takes a considerable amount of time to identify each image, making it unsuitable for real-time vehicle brand and model classification in practical scenarios. Abdul et al. proposed a method employing a cascade classifier (Siddiqui et al., 2016). Initially, representative features were extracted from the samples instead of using all features. Subsequently, a cascade-based SVM classifier was employed, resulting in significant improvements in real-time recognition. Biglari et al. (2019) introduced an algorithm based on the histogram of gradient directions feature and cascade classifier. Multiple vehicle brand models were trained first, followed by classification using a cascade SVM classifier, achieving an impressive classification accuracy of up to 96.78%. However, this method still requires hardware acceleration for real-time classification.

The third approach involves vehicle pattern recognition based on deep learning. Yang et al. proposed a method for recognizing vehicle brands and models based on the joint attributes of vehicles (Yang et al., 2015). This method extracts vehicle features from multiple perspectives and angles, fuses the extracted features, and performs recognition. While this method is well-suited for recognizing vehicle brands and models in complex scenes, its real-time performance is compromised due to the abundance of features. Huang et al. suggested randomly discarding certain layers during the training of ResNet to obtain a convolutional neural network with random depth (Huang et al., 2016), thereby addressing the issue of gradient vanishing caused by excessively deep networks. Fang et al. introduced a fine-grained method for recognizing vehicle brands and models (Fang et al., 2016), utilizing a CNN model to extract local and overall features of vehicles, and combining them for classification. Wang et al. (2020) proposed a method based on structural graph to learn discriminative representations for vehicle recognition. This approach first constructs a global structural graph from the features generated by a convolutional network. Then, it utilizes this structural graph as guidance to generate effective vehicle representations. Mo et al. (2020) analyzed the relationship between the number and distribution of vehicle axles and the weight limit of trucks. They proposed a circular detection method based on an improved Hough and clustering algorithm to identify the axles of trucks. Presently, most studies on deep learning for vehicle brand recognition rely on a single convolutional neural network model. However, for the intricate task of truck brand classification, a single model falls short in achieving satisfactory classification accuracy. Consequently, integrating multiple convolutional neural network models to develop a fusion model suitable for truck brand classification becomes a problem that requires resolution in this study.

# 3. Method

In typical scenarios, automatic weighing stations on highways are equipped with multiple cameras to capture frontal and side images of trucks. When utilizing these images for model recognition, the initial step involves utilizing the frontal image (front view) for identification. Analyzing the frontal image allows for the determination of the truck's model. Additionally, the side image is utilized to enhance accuracy in identifying the frontal view. The side image provides supplementary perspectives and details, thereby improving the accuracy of frontal view recognition. Moreover, the side image enables the segmentation of the truck into multiple parts, further refining model recognition precision. Through comprehensive analysis of both frontal and side images, we can achieve more accurate truck identification and conduct additional analysis based on its body features. Knowing the truck's model provides information regarding its rated load capacity. The weight measurement data obtained in the automatic weighing area enables straightforward determination of whether the truck is overloaded. Moreover, the inclusion of license plates information enables efficient monitoring and regulation by traffic authorities. Figure 1 illustrates the process described above.

## 3.1. The improved MMAL-Net

We improved MMAL-Net (Zhang et al., 2021) and employed it for truck recognition and classification. In Figure 2, we illustrate the network architecture that was constructed during the training phase, consisting of three branches: frontal, side, and part branches. The frontal branch is responsible for recognizing and classifying frontal truck images, while the side branch receives side images and segments them into multiple parts using the Attention Part Proposal Module (APPM). The part branch, on the other hand, specializes in recognizing and classifying part images. All three branches utilize a ResNet-50 (He et al., 2016) for feature extraction and employ a Fully Connected (FC) layer for classification, employing cross-entropy loss as the classification loss function.

ResNet-50 is a CNN architecture that belongs to the ResNet family. The ResNet family of architectures was specifically developed to address the problem of vanishing gradients in deep neural networks. In ResNet-50, the numerical suffix "50" indicates that the network consists of a total of 50 layers, including convolutional layers, pooling layers, fully connected layers, and shortcut connections. The key innovation of ResNet lies in the introduction of residual or skip connections, which allow information to bypass certain layers. This enables the network to learn more effectively by facilitating the propagation of gradients during training and enabling the acquisition of deeper and more complex representations. These skip connections also mitigate the problem of degradation, wherein the accuracy of a deep network decreases as its depth increases, by facilitating the training of deeper networks. ResNet-50 has been widely utilized and has achieved significant success in various machine vision tasks, such as image classification, object detection, and image segmentation. It has proven to be a powerful architecture that has advanced the field of computer vision and deep learning.

Formulas 1, 2, and 3 represent the loss function of the three branches, respectively.

$$L_{frontal} = -\log\left(P_f\left(c\right)\right) \tag{1}$$

$$L_{side} = -\log\left(P_s\left(c\right)\right) \tag{2}$$

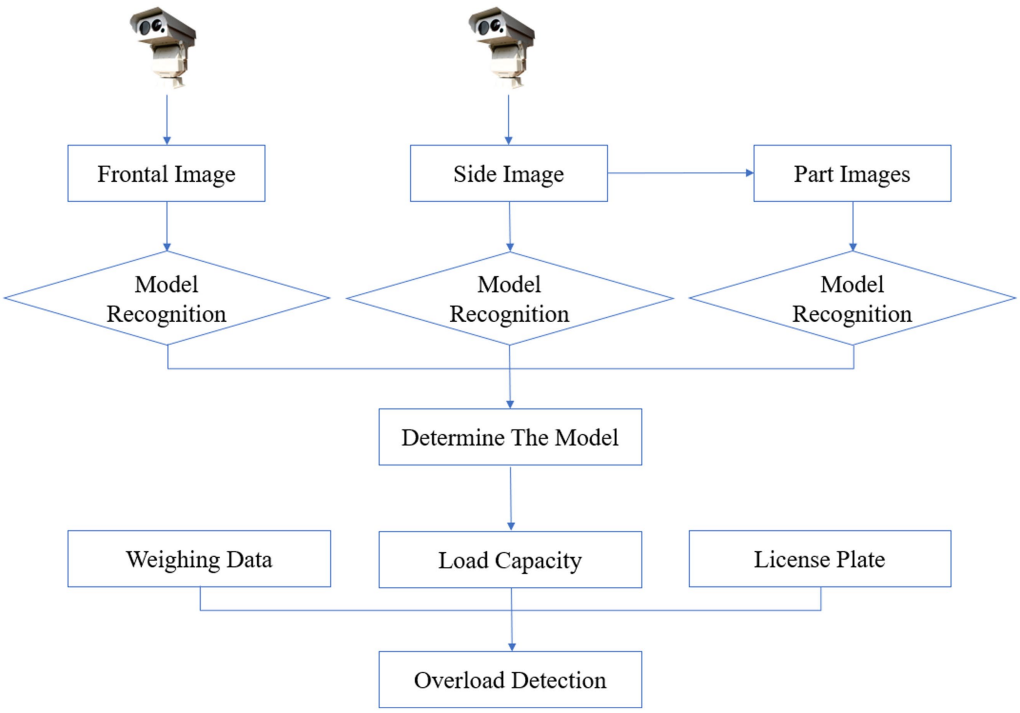



FIGURE 1
The process of the overload detection.

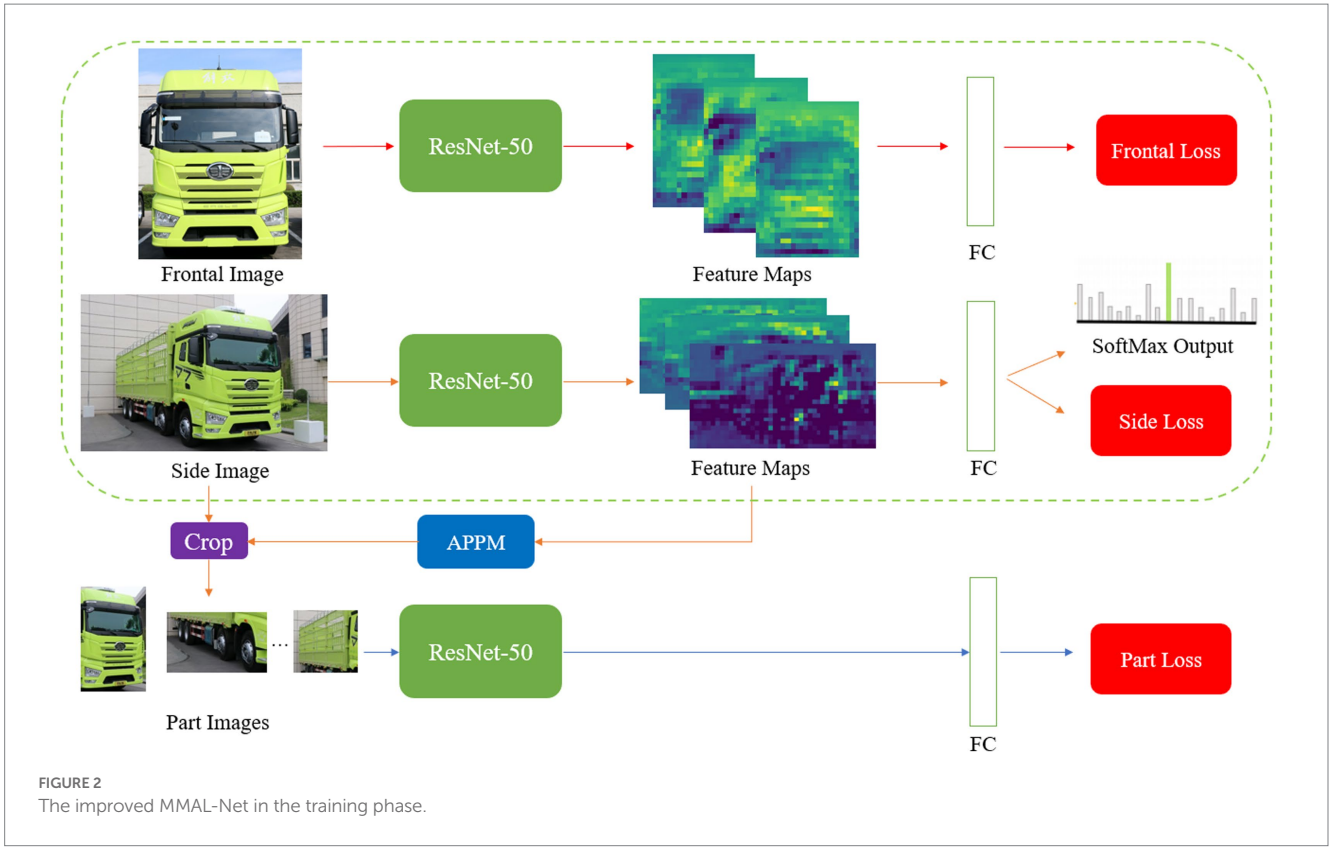



FIGURE 2
The improved MMAL-Net in the training phase.

$$L_{part} = -\sum_{n=0}^{N-1} \log\left(P_{p(n)}(c)\right) \tag{3}$$

Where $c$ represents the ground truth label of the input image, while $P_f$ and $P_s$ denote the category probabilities obtained from the last softmax layer outputs of the frontal and side branches, respectively. $P_{p(n)}$ refers to the output of the softmax layer in the part branch that corresponds to the $n$th part image. $N$ represents the total count of part images.

The total loss is defined as Formula 4:

$$L_{total} = L_{frontal} + L_{side} + L_{part} \tag{4}$$

The total loss is calculated as the cumulative sum of losses from the three branches, collaborating to enhance the model's performance during backpropagation. This enables the final converged model to generate classification predictions by considering both the global structural attributes of the object and its detailed features. During the testing phase, the part branch was excluded to minimize computational complexity, ensuring efficient prediction times for practical applications of our method.

## 3.2. APPM

By analyzing the activation map $A$, we observed that areas with high activation values corresponded to key parts, such as the front area of the truck. To identify these informative regions, we adopted a sliding window approach inspired by object detection techniques. This approach allowed us to extract part images from windows containing relevant information. Additionally, we employed a modified version of the traditional sliding window method using a fully convolutional network, similar to the approach used in Overfeat (Sermanet et al., 2013). This method involved obtaining feature maps for different windows from the output feature map of the previous network branch. Subsequently, we aggregated the activation maps $A_w$ of each window along the channel dimension and computed their mean activation value $\overline{a}_w$, as described in Formula 5. Here, $H_w$ and $W_w$ denote the height and width of a window's feature map, respectively. We then ranked the windows based on their $\overline{a}_w$ values, with higher values indicating more informative regions, as illustrated in Figure 3.

$$\overline{a}_w = \frac{\sum_{x=0}^{W_w-1}\sum_{y=0}^{H_w-1} A_w(x,y)}{H_w \times W_w} \tag{5}$$

However, we cannot directly select the initial windows because they are often adjacent to the windows with the highest average activation values $\overline{a}_w$ and contain nearly identical parts. Nonetheless, our objective is to choose a diverse range of parts. To minimize redundancy in the regions, we employ Non-Maximum Suppression (NMS) to select a fixed number of windows as part images at different scales. The visualization of the module's output in Figure 4 demonstrates that the proposed method effectively identifies distinct part regions with varying levels of importance. We utilize red, orange, yellow, and green rectangles to highlight the regions proposed by APPM that have the highest average activation values at various

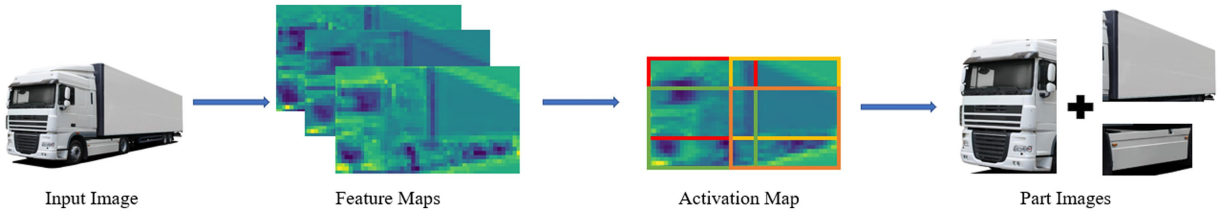



FIGURE 3
The simple pipeline of the APPM. We use red, orange, yellow and green colors to indicate the order of windows' $\bar{a}_w$.

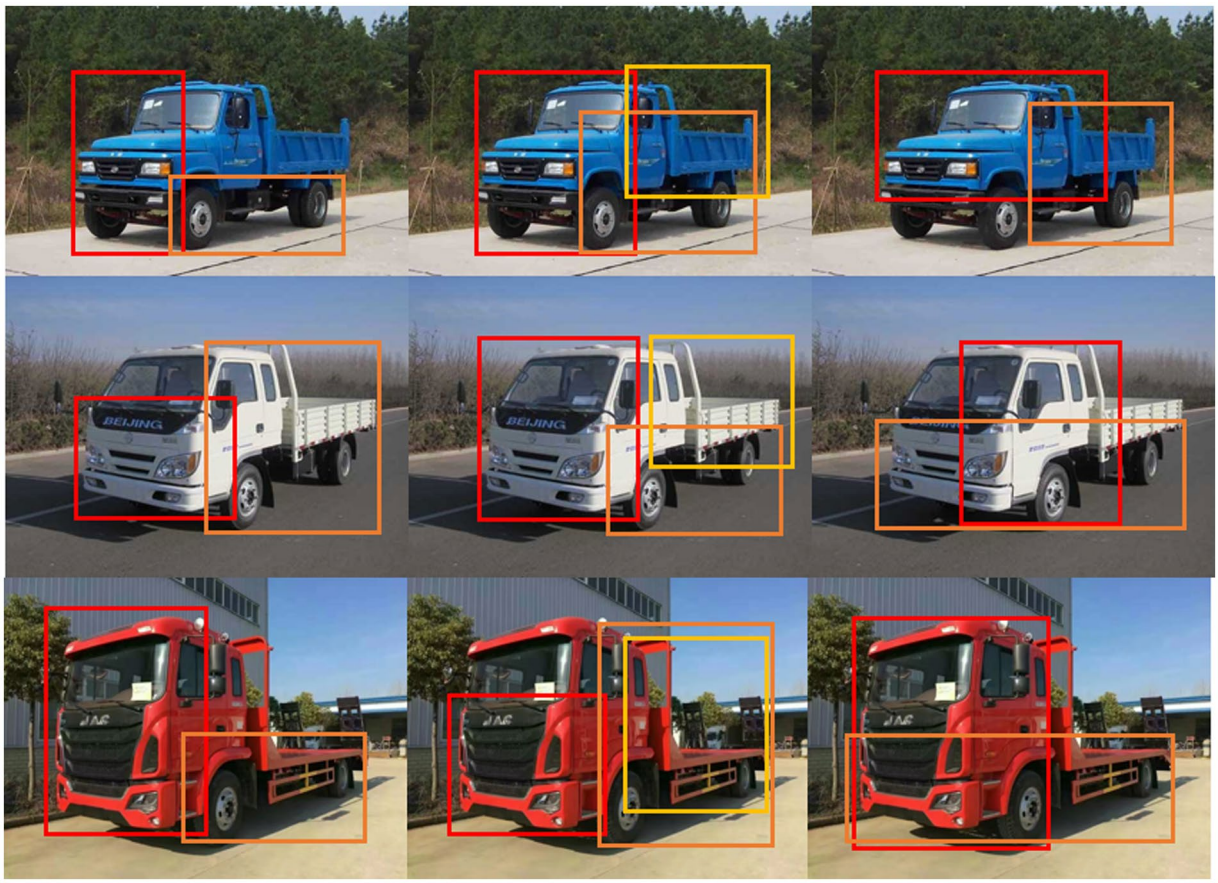


FIGURE 4
Visualization of part regions.

scales, with the red rectangle indicating the highest value. Figure 4 illustrates that the proposed approach captures detailed information and exhibits a more logical ordering at the same scale, thus significantly enhancing the model's robustness to scale variations. Notably, the head region stands out as the most discriminative region for truck recognition.

## 4. Results and discussion

To validate the advantages of the enhanced MMAL-Net, we conducted an evaluation of our method on the well-established and competitive benchmark dataset, Stanford Cars (Krause et al., 2013).

In our experiments, we adopted a consistent preprocessing approach. Initially, we resized the images to dimensions of $512 \times 512$, serving as inputs for both the frontal and side branches. Additionally, all part images were uniformly resized to $256 \times 256$ for the part branch. To ensure efficient initialization, we pre-trained ResNet-50 on the widely used ImageNet dataset, allowing us to effectively obtain the network's initial weights. Throughout both the training and testing phases, we exclusively relied on image-level labels, refraining from employing any additional annotations. Our optimization process involved utilizing SGD with specific hyperparameters: a momentum

value of 0.9 and a weight decay of 0.0001. To enhance training efficiency, we employed a mini-batch size of 6, utilizing a Tesla P100 GPU for computation. For fine-tuning the learning process, we set the initial learning rate to 0.001, which we later scaled down by a factor of 0.1 after 60 epochs. This step was instrumental in facilitating smoother convergence during training. We utilized PyTorch as the foundational framework.

In the experiments, we compared the proposed method to several baseline approaches and achieved competitive results, as shown in Table 1. By comparison, we can observe that our method attains the highest accuracy 95.03%.

In practice, there is a continuous emergence of new truck models. Given their recent introduction, it becomes challenging to obtain an adequate number of instances for constructing a comprehensive dataset. Hence, we utilized a customized dataset on a smaller scale to validate the applicability and effectiveness of our method. The personalized truck dataset includes four truck models: FAW J7, Shaanxi Delong X3000, Dongfeng Dorica D6, and JAC Junling V6. FAW J7 and Shaanxi Delong X3000 are heavy-duty trucks, whereas Dongfeng Dorica D6 and JAC Junling V6 are light-duty trucks. Each truck category is composed of 50 sets of training images and 20 sets of test images and each set comprises one frontal image and one side image. An example is depicted in Figure 5.

Thereafter, the overall network structure with specific features was fine-tuned to achieve fine-grained recognition of multiple target models. Moreover, we evaluated the effect of varied training samples on the recognition performance of the intelligent identification model by testing the same dataset using different

TABLE 1 Comparison of different methods on the Stanford Cars dataset.

| Methods | Backbone | Source | Accuracy (%) |
|---|---|---|---|
| RA-CNN Fu et al. (2017) | VGGNet-19 | CVPR'2017 | 92.5 |
| MA-CNN Zheng et al. (2017) | VGGNet-19 | ICCV'2017 | 92.8 |
| NTS-Net Yang et al. (2018) | ResNet-50 | ECCV'2018 | 93.9 |
| MAMC Sun et al. (2018) | ResNet-101 | ECCV'2018 | 93.0 |
| TASN Zheng et al. (2019) | ResNet-50 | CVPR'2019 | 93.8 |
| DCL Chen et al. (2019) | ResNet-50 | CVPR'2019 | 94.5 |
| API-Net Zhuang et al. (2020) | ResNet-50 | AAAI'2020 | 94.8 |
| DP-Net Wang et al. (2021) | ResNet-50 | AAAI'2021 | 94.8 |
| SAM Shu et al. (2022) | ResNet-50 | ECCV'2022 | 94.18 |
| MSHQP Tan et al. (2022) | ResNet-152 | TOMM'2022 | 94.9 |
| The Improved MMAL-Net | ResNet-50 | This paper | 95.03 |

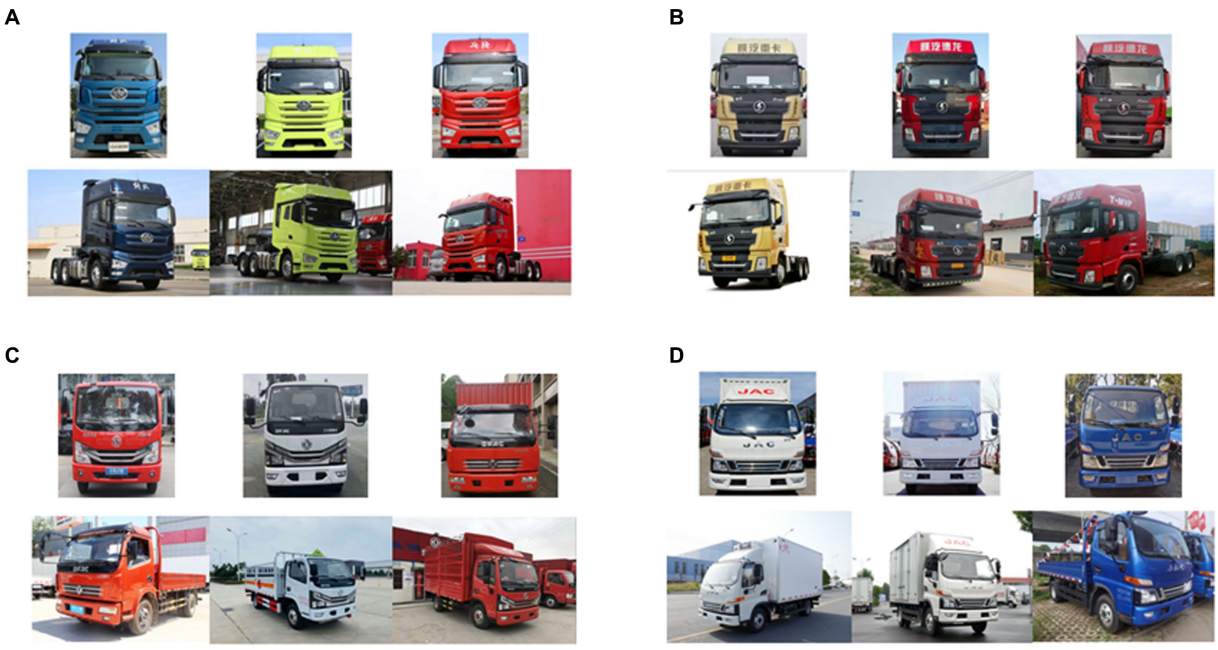



FIGURE 5
The personalized truck dataset. **(A–D)** represent FAW J7, Shaanxi Delong X3000, Dongfeng Dorica D6, and JAC Junling V6.

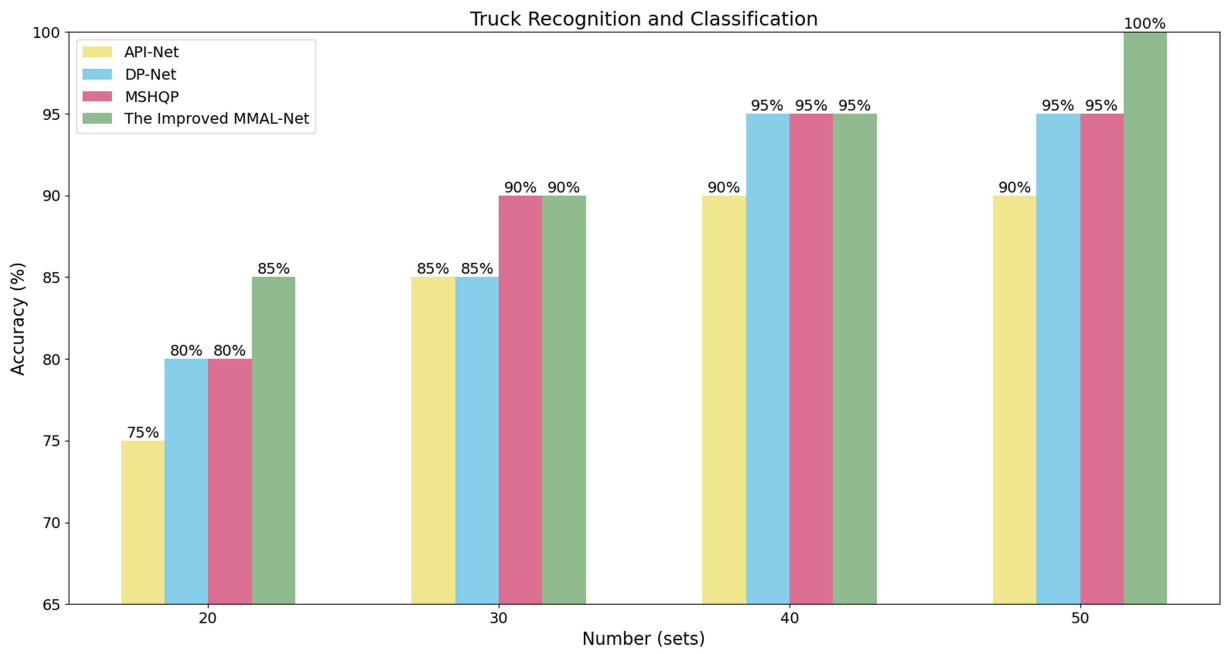



FIGURE 6
The comparison result on different number of training sets.

incremental levels of training data. This simulation emulated the impact of increasing the number of target truck images collected in actual scenarios on the enhancement of the recognition model's performance.

In our experiment, we selected sets of 20, 30, 40, and 50 images for each classifier as training datasets and used the same number of test set to compare the performance of API-Net, DP-Net, MSHQP and the improved MMAL-Net. It is worth mentioning that API-Net, DP-Net, and MSHQP were the top three performing methods in our experiments on the Stanford Cars dataset, excluding our proposed method. The results indicate that as the training data increases, the network's ability to identify and extract features from target trucks gradually improves, suggesting that larger datasets can effectively enhance the model's capability to extract potential features. The improved MMAL-Net exhibits comparable or superior performance to other methods across all numbers of training sets, demonstrating its superior ability to extract fine-grained features of target trucks (see Figure 6).

In our small-scale custom dataset, it is evident that the recognition accuracy reaches 85% when the training set consists of 20 sets of photos. This greatly addresses the practical issue of scarce images of a particular type of truck. The improved MMAL-Net demonstrated remarkable resilience to image quality and scene noise, as evidenced by its recognition accuracy of 100% when trained on a dataset comprising 50 sets of samples. This noteworthy achievement further supports the superior performance of the enhanced network.

Confusion matrices in Figure 7 illustrate the test results of the improved MMAL-Net. At a training set size of 40 sets of images, the improved MMAL-Net had a single misclassification on the test set, misclassifying a Dongfeng Dorica D6 as a JAC Junling V6. However, at a training set size of 50 sets, all classifications were accurate. The improved MMAL-Net accurately classified heavy-duty trucks, avoiding misclassification as light-duty trucks. Similarly, it correctly identified light-duty trucks without misclassification as heavy-duty trucks. This is crucial because misidentifying an overloaded light-duty truck as a heavy-duty truck can result in undetected overweight issues, thus posing safety concerns.

## 5. Conclusion

This paper introduces a method for precise identification of truck models. In our experimental evaluation, this method achieved an accuracy of 95.03% on the competitive benchmark dataset, Stanford Cars. Furthermore, it achieved an accuracy of 100% on our custom truck dataset. When integrated with weighing and license plates systems, it can be applied in highway automatic weighing stations to determine if a truck is overloaded. By providing accurate truck information, this method contributes to freight management, transportation safety, and highway planning, thereby fostering the development of the logistics industry and improving traffic safety. However, the accuracy of truck model recognition may decrease in real-world scenarios due to the reduced data quality. Consequently, future research will focus on addressing this issue, with specific emphasis on long-distance shooting conditions.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

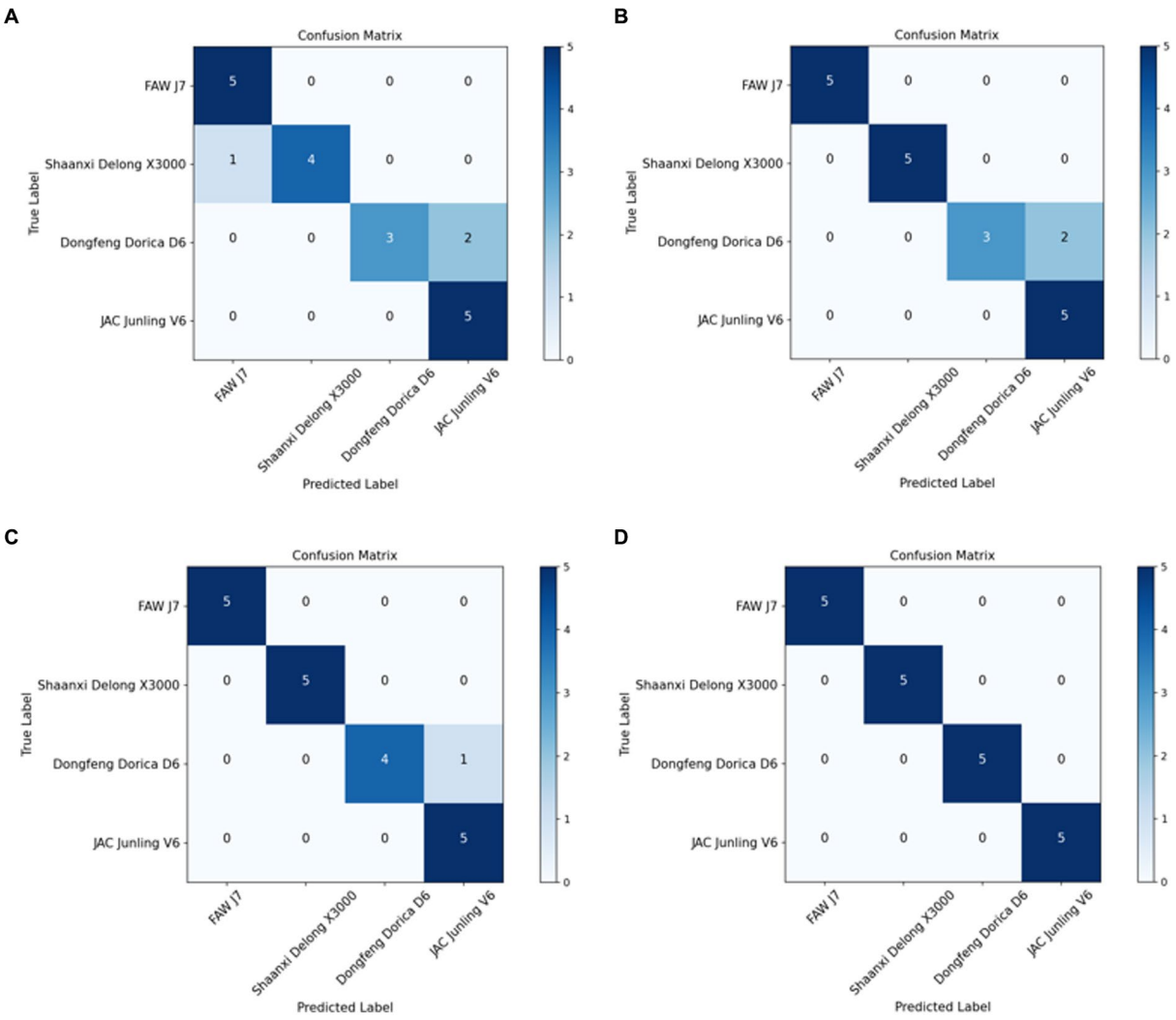



FIGURE 7
Confusion matrices of the improved MMAL-Net. **(A–D)** represent 20, 30, 40 and 50 sets of images, respectively, which are used as the training dataset.

## Author contributions

JiaS contributed to conception and design of the study. JiaS and JinS organized the methodology. JiaS and ZY performed the statistical analysis. JiaS wrote the first draft of the manuscript. JiaS, ZG, and YS contributed to the visualization of the results. JiaS and LL contributed to the supervision of the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This research was supported by National Key R&D Program of China (Grant no. 2021YFB3300503).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Biglari, M., Soleimani, A., and Hassanpour, H. (2019). A cascading scheme for speeding up multiple classifier systems. *Pattern. Anal. Applic.* 22, 375–387. doi: 10.1007/s10044-017-0637-4

Branson, S., Van Horn, G., Belongie, S., and Perona, P. (2014). Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv* 2952:1406. doi: 10.48550/arXiv.1406.2952

Chen, Y., Bai, Y., Zhang, W., and Mei, T. (2019). Destruction and construction learning for fine-grained image recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5157–5166).

Dalal, N., and Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886–893). Ieee.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). "Decaf: a deep convolutional activation feature for generic visual recognition" in *International conference on machine learning* (Amsterdam: PMLR), 647–655.

Fang, J., Zhou, Y., Yu, Y., and Du, S. (2016). Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture. *IEEE Trans. Intell. Transp. Syst.* 18, 1782–1792. doi: 10.1109/TITS.2016.2620495

Fraz, M., Edirisinghe, E. A., and Sarfraz, M. S. (2014). Mid-level-representation based lexicon for vehicle make and model recognition. In 2014 22nd International Conference on Pattern Recognition (pp. 393–398). IEEE.

Fu, J., Zheng, H., and Mei, T. (2017). Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4438–4446).

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580–587).

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. (2016). Deep networks with stochastic depth. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14 (pp. 646–661). Springer International Publishing. United States

Huang, Y., Wu, R., Sun, Y., Wang, W., and Ding, X. (2015). Vehicle logo recognition system based on convolutional neural networks with a pretraining strategy. *IEEE Trans. Intell. Transp. Syst.* 16, 1951–1960. doi: 10.1109/TITS.2014.2387069

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In 2010 IEEE computer society conference on computer vision and pattern recognition (pp. 3304–3311). IEEE.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In Proceedings of the IEEE international conference on computer vision workshops (pp. 554–561).

Lin, T. Y., Roy Chowdhury, A., and Maji, S. (2005) Bilinear CNN models for fine-grained visual recognition. In Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 1449–1457.

Mo, X., Sun, C., Li, D., Huang, S., and Hu, T. (2020). Research on the method of determining highway truck load limit based on image processing. *IEEE Access* 8, 205477–205486. doi: 10.1109/ACCESS.2020.3037195

Psyllos, A. P., Anagnostopoulos, C. N. E., and Kayafas, E. (2010). Vehicle logo recognition using a sift-based enhanced matching scheme. *IEEE Trans. Intell. Transp. Syst.* 11, 322–328. doi: 10.1109/TITS.2010.2042714

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv* 1312:6229. doi: 10.48550/arXiv.1312.6229

Shu, Y., Yu, B., Xu, H., and Liu, L. (2022). Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism. In European Conference on Computer Vision (pp. 449–465). Cham: Springer Nature Switzerland.

Siddiqui, A. J., Mammeri, A., and Boukerche, A. (2016). Real-time vehicle make and model recognition based on a bag of SURF features. *IEEE Trans. Intell. Transp. Syst.* 17, 3205–3219. doi: 10.1109/TITS.2016.2545640

Song, Y., Sebe, N., and Wang, W. (2022). On the eigenvalues of global covariance pooling for fine-grained visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 1–3566. doi: 10.1109/TPAMI.2022.3178802

Sun, M., Yuan, Y., Zhou, F., and Ding, E. (2018). Multi-attention multi-class constraint for fine-grained image recognition. In Proceedings of the European conference on computer vision (ECCV) (pp. 805–821).

Tan, M., Yuan, F., Yu, J., Wang, G., and Gu, X. (2022). Fine-grained image classification via multi-scale selective hierarchical biquadratic pooling. *ACM Trans. Multimedia Comp, Commun, Appl (TOMM)* 18, 1–23. doi: 10.1145/3492221

Wang, C., Fu, H., and Ma, H. (2020). Global structure graph guided fine-grained vehicle recognition. In ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1913–1917). IEEE.

Wang, S., Li, H., Wang, Z., and Ouyang, W. (2021). Dynamic position-aware network for fine-grained image recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35. 2791–2799).

Wang, Q., Li, P., and Zhang, L. (2017). G2DeNet: global Gaussian distribution embedding network and its application to visual recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2730–2739).

Wei, X. S., Xie, C. W., Wu, J., and Shen, C. (2018). Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recogn.* 76, 704–714. doi: 10.1016/j.patcog.2017.10.002

Xie, G. S., Zhang, X. Y., Yang, W., Xu, M., Yan, S., and Liu, C. L. (2017). LG-CNN: from local parts to global discrimination for fine-grained recognition. *Pattern Recogn.* 71, 118–131. doi: 10.1016/j.patcog.2017.06.002

Yang, L., Luo, P., Change Loy, C., and Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3973–3981).

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., and Wang, L. (2018). Learning to navigate for fine-grained classification. In Proceedings of the European conference on computer vision (ECCV) (pp. 420–435).

Zhang, N., Donahue, J., Girshick, R., and Darrell, T. (2014). Part-based R-CNNs for fine-grained category detection. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13 (pp. 834–849). Springer International Publishing.

Zhang, F., Li, M., Zhai, G., and Liu, Y. (2021). Multi-branch and multi-scale attention learning for fine-grained visual categorization. In Multi Media Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part I 27 (pp. 136–147). Springer International Publishing.

Zhang, Y., Wei, X. S., Wu, J. X., Cai, J. F., Lu, J. B., Nguyen, V. A., et al. (2016). Weakly supervised fine-grained categorization with part-based image representation. *IEEE Trans. Image Process.* 25, 1713–1725. doi: 10.1109/TIP.2016.2531289

Zheng, H., Fu, J., Mei, T., and Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE international conference on computer vision (pp. 5209–5217).

Zheng, H., Fu, J., Zha, Z. J., and Luo, J. (2019). Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5012–5021).

Zhou, Z., Liu, J., Li, H., and Ou, J. (2005). A new kind of high durable traffic weighbridge based on fbg sensors. *Proceedings of SPIE - The Int. Soc. Optical Eng.* 5855, 735–738. doi: 10.1117/12.623436

Zhuang, P., Wang, Y., and Qiao, Y. (2020). Learning attentive pairwise interaction for fine-grained classification. In Proceedings of the AAAI conference on artificial intelligence. 34, 13130–13137).

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 21 September 2023
DOI 10.3389/fnins.2023.1273686



OPEN ACCESS

EDITED BY
Teng Li,
Anhui University, China

REVIEWED BY
Yu Liu,
Changchun University of Science and Technology, China
Yaw Missah,
Kwame Nkrumah University of Science and Technology, Ghana

*CORRESPONDENCE
Yuanfa Wang
✉ wangyf@cqupt.edu.cn

RECEIVED 07 August 2023
ACCEPTED 04 September 2023
PUBLISHED 21 September 2023

CITATION
Wang H, Wang K, Yan T, Zhou H, Cao E, Lu Y, Wang Y, Luo J and Pang Y (2023) Endoscopic image classification algorithm based on Poolformer.
*Front. Neurosci.* 17:1273686.
doi: 10.3389/fnins.2023.1273686



# Endoscopic image classification algorithm based on Poolformer

Huiqian Wang[1,2], Kun Wang[1], Tian Yan[1], Hekai Zhou[1], Enling Cao[1], Yi Lu[1], Yuanfa Wang[1,2]*, Jiasai Luo[1] and Yu Pang[1]

[1]Postdoctoral Research Station, Chongqing Key Laboratory of Photoelectronic Information Sensing and Transmitting Technology, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]Chongqing Xishan Science & Technology Co., Ltd., Chongqing, China

Image desmoking is a significant aspect of endoscopic image processing, effectively mitigating visual field obstructions without the need for additional surgical interventions. However, current smoke removal techniques tend to apply comprehensive video enhancement to all frames, encompassing both smoke-free and smoke-affected images, which not only escalates computational costs but also introduces potential noise during the enhancement of smoke-free images. In response to this challenge, this paper introduces an approach for classifying images that contain surgical smoke within endoscopic scenes. This classification method provides crucial target frame information for enhancing surgical smoke removal, improving the scientific robustness, and enhancing the real-time processing capabilities of image-based smoke removal method. The proposed endoscopic smoke image classification algorithm based on the improved Poolformer model, augments the model's capacity for endoscopic image feature extraction. This enhancement is achieved by transforming the Token Mixer within the encoder into a multi-branch structure akin to ConvNeXt, a pure convolutional neural network. Moreover, the conversion to a single-path topology during the prediction phase elevates processing speed. Experiments use the endoscopic dataset sourced from the Hamlyn Centre Laparoscopic/Endoscopic Video Dataset, augmented by Blender software rendering. The dataset comprises 3,800 training images and 1,200 test images, distributed in a 4:1 ratio of smoke-free to smoke-containing images. The outcomes affirm the superior performance of this paper's approach across multiple parameters. Comparative assessments against existing models, such as mobilenet_v3, efficientnet_b7, and ViT-B/16, substantiate that the proposed method excels in accuracy, sensitivity, and inference speed. Notably, when contrasted with the Poolformer_s12 network, the proposed method achieves a 2.3% enhancement in accuracy, an 8.2% boost in sensitivity, while incurring a mere 6.4 frames per second reduction in processing speed, maintaining 87 frames per second. The results authenticate the improved performance of the refined Poolformer model in endoscopic smoke image classification tasks. This advancement presents a lightweight yet effective solution for the automatic detection of smoke-containing images in endoscopy. This approach strikes a balance between the accuracy and real-time processing requirements of endoscopic image analysis, offering valuable insights for targeted desmoking process.

KEYWORDS

**endoscopic image, image classification, Poolformer, token mixer, ConvNeXt, single-path topology only during inference**

# 1. Introduction

Endoscopes are essential tools that utilize the body's natural cavities or tiny incisions to provide real-time visualization of internal organs and tissues (Fu et al., 2021; Boese et al., 2022; Chadebecq et al., 2023). This minimizes the need for larger incisions during surgery, leading to shorter patient recovery periods. Consequently, endoscopy is now extensively employed in examining and treating various diseases involving the gastrointestinal tract (Aceves et al., 2022; Niknam et al., 2022), ear, nose, throat (Bastier et al., 2022; Poutoglidis et al., 2022), spine (Ahn, 2020; Simpson et al., 2022) and urinary system(Zou et al., 2020; Yamashita et al., 2022). Despite the benefits of endoscopy, challenges arise during procedures: the generation of smoke due to the destruction and vaporization of tissue proteins and fat by the instruments (Yi et al., 2023). This smoke hinders the visibility of tissue structures in endoscopic images, obstructing the physician's vision and impeding accurate judgment and treatment. To address this challenge, image-based surgical smoke analysis and processing have emerged as a promising solution. Not constrained by hardware limitations, this approach reduces the reliance on surgical aids and assists physicians in obtaining clearer views for more precise diagnoses and treatments. Consequently, it holds immense potential and value for clinical applications.

However, the existing methods for intelligent analysis and processing of surgical smoke primarily focus on desmoking endoscopic images. For instance, Wang et al. (2019) proposed an improved convolutional neural network (CNN) with an encoder-decoder architecture for real-time surgical smoke removal. Their network takes an image with smoke along with its laplacian image pyramid decomposition as input and produces an image with smoke removed. To create the synthetic dataset, they utilized Blender and Adobe Photoshop to add rendered smoke to clean images. Similarly, Lin et al. (2021) introduced a supervised UNet-based network where the Laplace pyramid is fused at the encoder, and the CBAM module is integrated at the decoder. They employed Blender to generate datasets of laparoscopic images with varying levels of light and dense smoke. Their method achieved a high structural similarity of 0.945 and a peak signal-to-noise ratio of 29.27 for the test data. Furthermore, Bolkar et al. (2018) constructed a synthetic surgical desmoking dataset. They adapted the integrated desmoking network, AOD-Net, initially designed for outdoor desmoking, and their proposed supervisory model comprises five convolutional layers with ReLU activation units and three cascade layers. Azam et al. (2022) removed smoke from laparoscopic images by manual multiple exposure image fusion method. Venkatesh et al. (2020), Pan et al. (2022), Zhou et al. (2022), and Su and Wu (2023) respectively used CycleGAN-based network structure to realize laparoscopic image de-smoking and affirmed the important role of smoke detection in laparoscopic image desmoking, but their main design focus was on the structure of smoke purification network. Additionally, Wang et al. (2023) proposed a desmoking method based on Swin transformer, employing Swin transformer blocks to extract deep features. Most of the aforementioned desmoking techniques process all endoscopic images within the video stream for smoke removal, which is inefficient because smoke is not consistently present throughout the entire surgical procedure, and a substantial portion of the video stream consists of smoke-free images. Processing all video stream images for de-smoking not only increases computational demands but may also introduce new noise into the original smoke-free images. Hence, it becomes imperative to differentiate between smoked and smoke-free images, enabling the smoke cleaning algorithms to selectively focus on desmoking only the images containing smoke, while leaving the smoke-free images unaltered. This targeted approach ensures more efficient and precise desmoking, preserving the clarity and integrity of the original smoke-free images. This approach would significantly reduce equipment resource requirements, improve processing speed, and enhance the real-time, accuracy, and scientificity of desmoking in endoscopic scenarios.

To date, few studies specifically focus on the classification of endoscopic images containing smoke. Nevertheless, endoscopic image classification aligns with the fundamental principles of other image binary/multi-classification problems, wherein the objective is to predict input images into multiple categories based on their distinctive features. In the early stages, researchers employed algorithms like k-nearest neighbors, Support Vector Machine, and Random Forest for such tasks. These methods typically involved feature extraction prior to classification, necessitating human selection of one or more features that influenced the classification quality. In recent years, CNNs have gained prevalence for image classification due to their ability to automatically extract relevant image features and demonstrate exceptional performance on large-scale datasets. Lecun et al. (1998) proposed an early CNN architecture, comprising two convolutional layers, two pooling layers, and three fully connected layers, which facilitated the classification and recognition of handwritten digits and laid the groundwork for subsequent image classification models. Notably, Krizhevsky et al. (2012) introduced AlexNet, which achieved groundbreaking results in the ImageNet image classification competition. Their work significantly improved performance on large-scale image datasets. Additionally, Tan and Le (2019) introduced EfficientNet, a CNN structure optimized through neural network search technology. Furthermore, ResNet was proposed as an innovative deep residual learning framework to address the issue of gradient explosion in deep network training (He et al., 2016). Howard et al. (2017) proposed MobileNet, a lightweight deep neural network designed for embedded devices. MobileNet utilizes depth-wise separable convolution to efficiently reduce the number of model operations and parameters, making it well-suited for resource-constrained environments. Dosovitskiy et al. (2021) made a significant breakthrough in image classification by directly applying the transformer architecture to this domain, introducing the vision transformer (ViT) model. The ViT model utilizes the transformer's encoder to extract essential features from images, resulting in remarkable advancements in image classification. In a related development, Yu et al. (2022) proposed the Poolformer model. Instead of employing the attention module, the Poolformer model utilizes a straightforward spatial pooling operation. Even with fewer parameters, the Poolformer model achieves competitive performance in image classification tasks. Furthermore, Almeida et al. (2022), Dewangan et al. (2022), and Zhao et al. (2022) individually explored lightweight CNNs for smoke detection in images of natural scenes.

Among existing image classification methods, network models like Poolformer have demonstrated the capability to achieve highly accurate real-time recognition in natural images. They hold significant potential for extending their effectiveness to the detection of endoscopic smoke-containing images. However, compared to natural images, endoscopic images face distinctive challenges in feature

extraction and recognition. This is primarily due to the non-Lambertian reflective properties of human tissues, resulting in weak texture features and a lack of salient features. Furthermore, the classification of endoscopic smoke-containing images necessitates real-time performance during surgical procedures, where achieving a high level of real-time efficiency is critical for successful implementation. The characteristics inherent in endoscopic scenes introduce complexity to the task of automatic feature extraction and recognition.

To enhance real-time performance while maintaining accuracy in smoke detection on endoscopic images with weak textures, this paper proposes a method for endoscopic smoke image classification using Poolformer. The primary enhancement of the algorithm lies in the model's encoder, where the Token Mixer is upgraded from a basic pooling layer to a multiplexed branching structure akin to the purely convolutional neural network ConvNeXt (Liu et al., 2022). During prediction, it is further transformed into a single-path topology to bolster the model's inference speed.

# 2. The proposed method

## 2.1. Overview

The Poolformer-based network for endoscopic image classification proposed in this paper is depicted in Figure 1. In terms of the network structure, the original Poolformer replaces the Multi-head Attention module in the encoder block of the conventional vision transformer with a simple pooling layer. To further enhance the feature extraction capabilities for weakly textured images, this paper proposes the design of a multi-branch pure convolutional neural network structure similar to ConvNeXt, aiming to optimize the pooling layer in the original Poolformer model. This enhancement improves the model's feature extraction ability. Furthermore, to ensure real-time processing in endoscopic video streaming, the model's structure is transformed into a one-way model to obtain classification results through predictive reasoning during the testing process.

## 2.2. Convolution module

In the Vision Transformer (ViT) module (Dosovitskiy et al., 2021), input tokens (vectors) are essential for processing images of various sizes. As an example, in the ViT-B/16 model, the input image, $x \in \mathbb{R}^{h \times w \times c}$, where $h$ denotes the height, $w$ signifies the width, and $c$ represents the number of channels, undergoes convolution with a kernel size of $16 \times 16$, a stride of 16, and employs 768 convolution kernels to accomplish this operation. This process involves partitioning the input image $x$ into patches of size $16 \times 16$. While increasing the convolutional kernel and step size in large datasets can expand the receptive field, allowing for feature maps over a wider area and obtaining superior global features, in smaller datasets, such as medical datasets like endoscopes, this advantage may lead to the loss of detailed information between patches.

To tackle this issue, this paper adopts the convolution-based patching method, which effectively mitigates the loss of detailed information. This approach removes the constraint that each patch size must be a multiple of the image's dimensions, enabling adaptation to datasets with varying size dimensions. As illustrated in Figure 2, the preprocessed input vector $x$ undergoes feature extraction through convolution, activation function, and maximum pooling operations. A downsampling operation is applied to meet the input specifications of the subsequent Positional Embedding layer. The GELU activation function is integrated in order to introduce randomness by combining it with the concept of dropout, thereby enhancing the robustness of the model. Additionally, to address the degradation problem, a residual module based on ResNet (He et al., 2016) is employed. Finally, a positional embedding layer vector of the same size as ViT is obtained through a convolution and flattening operation.

## 2.3. Improved Poolformer encoder

The encoder of the fundamental ViT model primarily comprises two components: an attention module, also known as the token mixer, which facilitates information exchange between tokens, and subsequent elements such as channel MLP and residual connections. Abstracting the architecture while disregarding the specifics of how the token mixer is implemented with an attention module, the aforementioned design can be represented as the MetaFormer architecture (Yu et al., 2022), depicted in the first panel of Figure 3A. Contrasting with the conventional ViT model, the Poolformer model transforms the multi-head attention mechanism into a simple pool pooling layer, as illustrated in Figure 3B. Leveraging the overall superiority of the entire MetaFormer framework and the inclusion of the pooling layer, it significantly reduces the computation burden, machine load, and required video memory.

The pooling layer, in the process of dimensionality reduction, may lead to the loss of local information, which is particularly critical in weak texture endoscopic images where local information plays a crucial role. It is essential to minimize information loss as much as possible. Convolutional neural networks excel at retaining local information compared to pooling layers. Leveraging this advantage, the token mixer part is optimized to adopt a ConvNeXt-like multiplexed branching structure, as depicted in Figure 3C. ConvNeXt is a pure convolutional neural network architecture that competes with transformer networks. In comparison to the transformer model, ConvNeXt significantly reduces the number of parameters, introduces spatial inductive bias, and eliminates positional bias. Consequently, this acceleration of network convergence leads to a more stable network training process. Through modifications involving stage proportions, grouping convolutions, an anti-bottleneck design, utilization of larger convolutional kernels in finer details, and replacing the activation function, ConvNeXt achieves faster inference speed and higher accuracy than the Swin Transformer.

For the improved Poolformer encoder, the 2D matrix $x_1$ obtained from the input image through the convolution operation and flattening operation in Figure 2 serves as the input sequence. The specific structure and steps, for example, using ViT-B/16 (where the 2D matrix $x_1$ is in the format of [197,768]), are illustrated in Figure 4. In step (1), $x_1$ undergoes mapping to interchange the $H$ (height) dimension and $C$ (channel) dimension, resulting in the matrix $x_2$. A similar operation is performed in step (2), where the height dimension containing class categorization information is considered as the channel dimension.

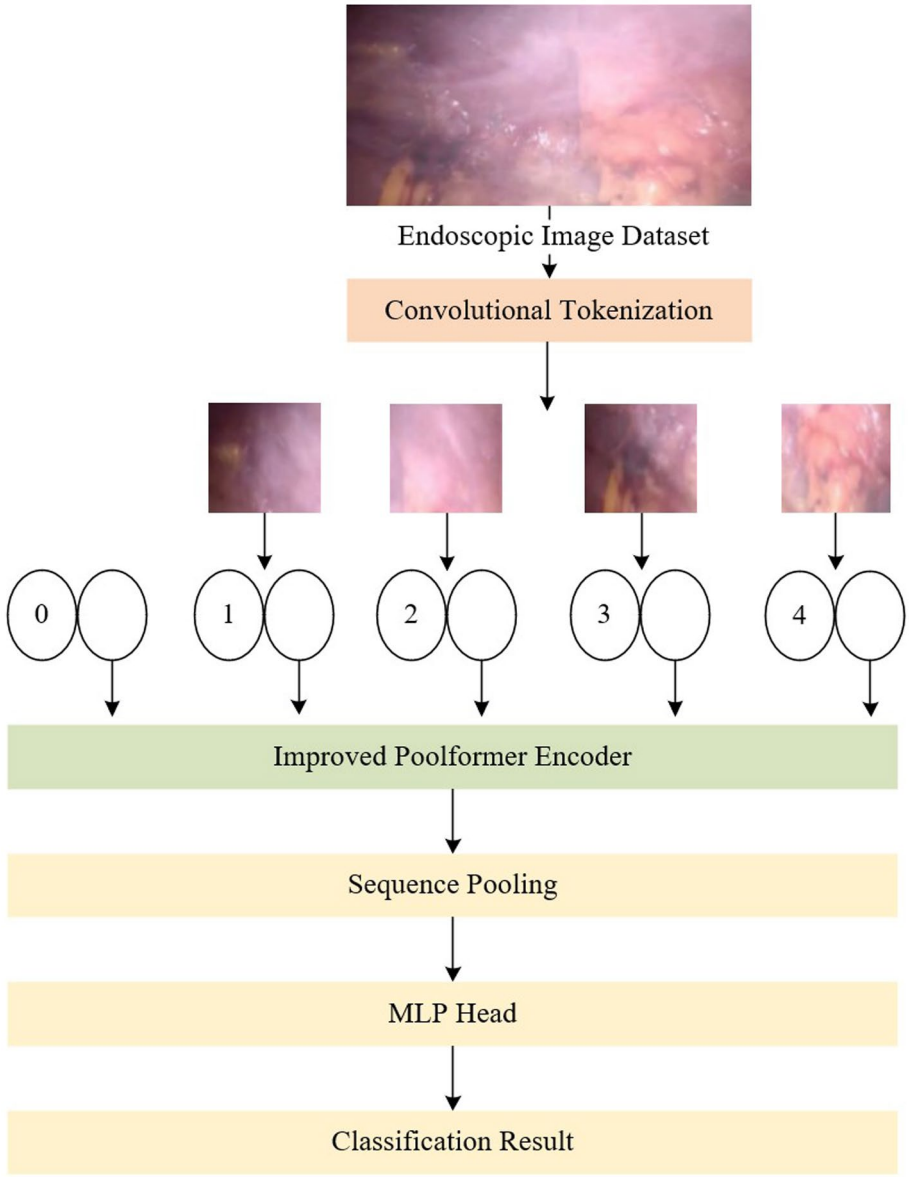



FIGURE 1
An overview of the proposed network, which consists of convolution module, improved Poolformer encoder, sequence pooling and MLP head.

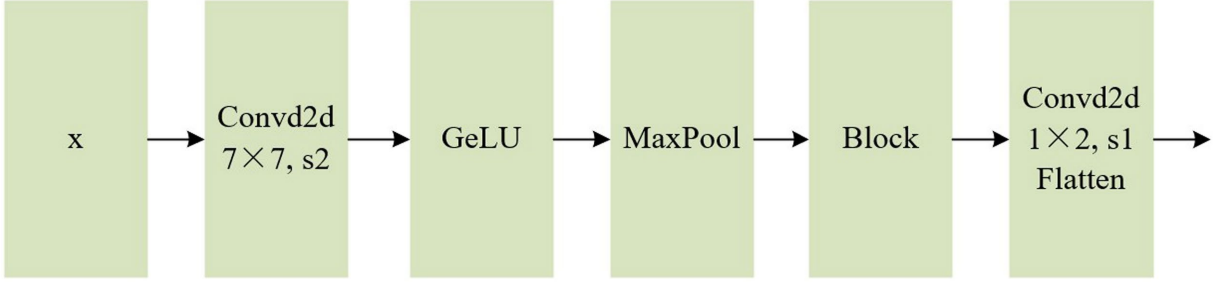



FIGURE 2
Flow chart of convolution module.

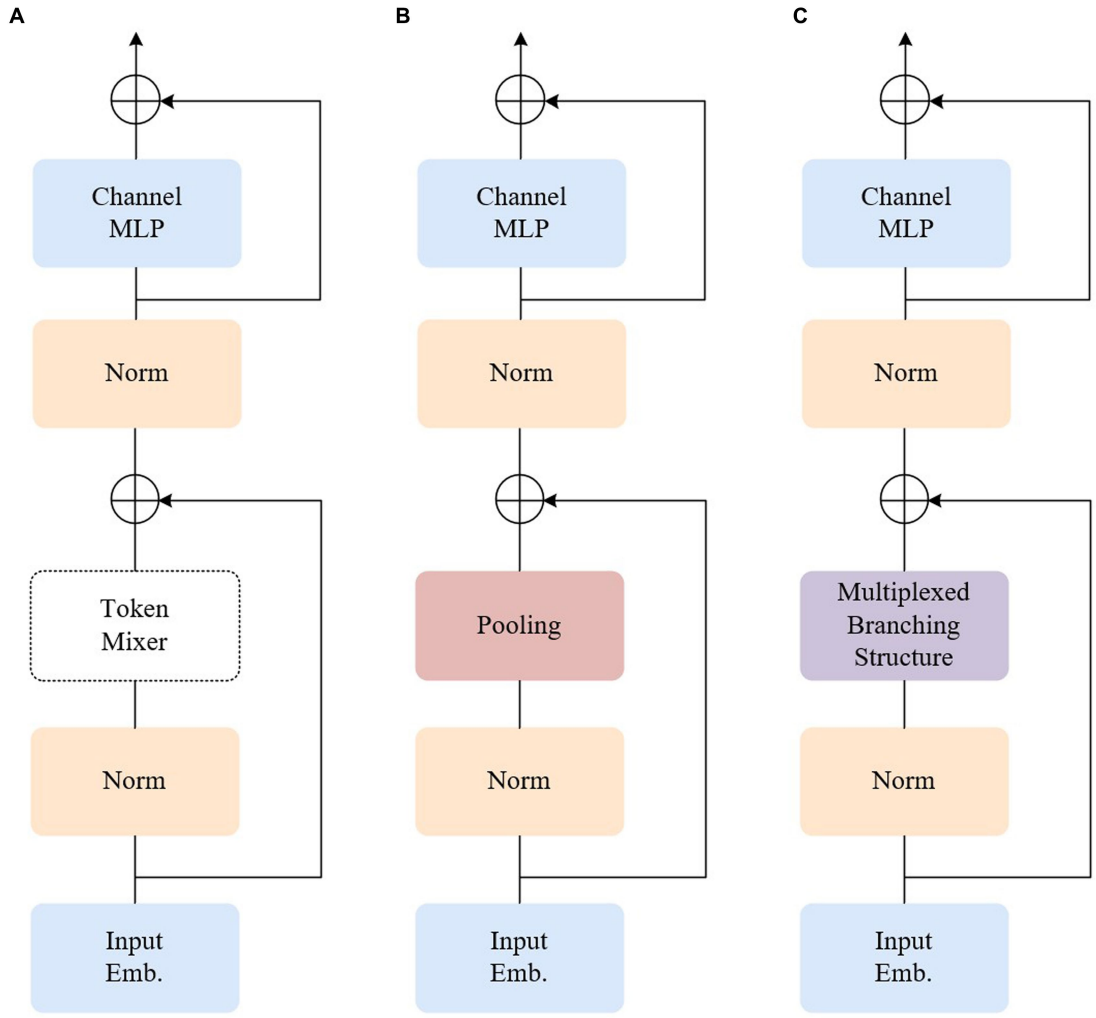



FIGURE 3
Illustrations of the architecture of different encoders. **(A)** MetaFormer. **(B)** Poolformer. **(C)** Our model.

## 2.4. RepConvNeXt block

The proposed module, transforms the ConvNeXt Block into a RepConvNeXt Block—a one-way structure resembling RepVgg (Ding et al., 2021)—during prediction process to further enhance real-time performance, as depicted in Figure 5. During training, using multi-branch structures such as ResNet or models like DenseNet (Huang et al., 2017) generally increases the model's representational capacity by parallelizing multiple branches.

Converting the multi-branch into a single-path topology during inference offers several advantages: Firstly, it enhances speed. Considering the degree of parallelism in hardware computation and MAC (memory access cost) during model reasoning, multi-branch models require separate computation of results for each branch. Some branches may compute faster while others compute more slowly, leading to potential underutilization of hardware arithmetic and insufficient parallelism. Additionally, each branch necessitates memory access and storage, resulting in substantial time wasted on IO operations. Secondly, it improves memory efficiency. The residual module depicted in Figure 6A, assuming the convolutional layer does not alter the number of channels, requires storing the respective feature maps on both the main branch and the shortcut branch, leading to roughly twice the memory consumption of the input activation before the add operation. Conversely, the structure shown in Figure 6B maintains the same memory usage throughout.

## 2.5. Classification

Through enhancements made to the Poolformer encoder, the output of the Transformer encoder after sequence pooling to the L-layer differs from the traditional ViT model. Instead of generating classification results by slicing the class token separately, the improved model utilizes data sequences containing both input image and class information. As a result, the model becomes more compact, and the sequence pooling output of the Transformer encoder produces sequential embedding in the latent

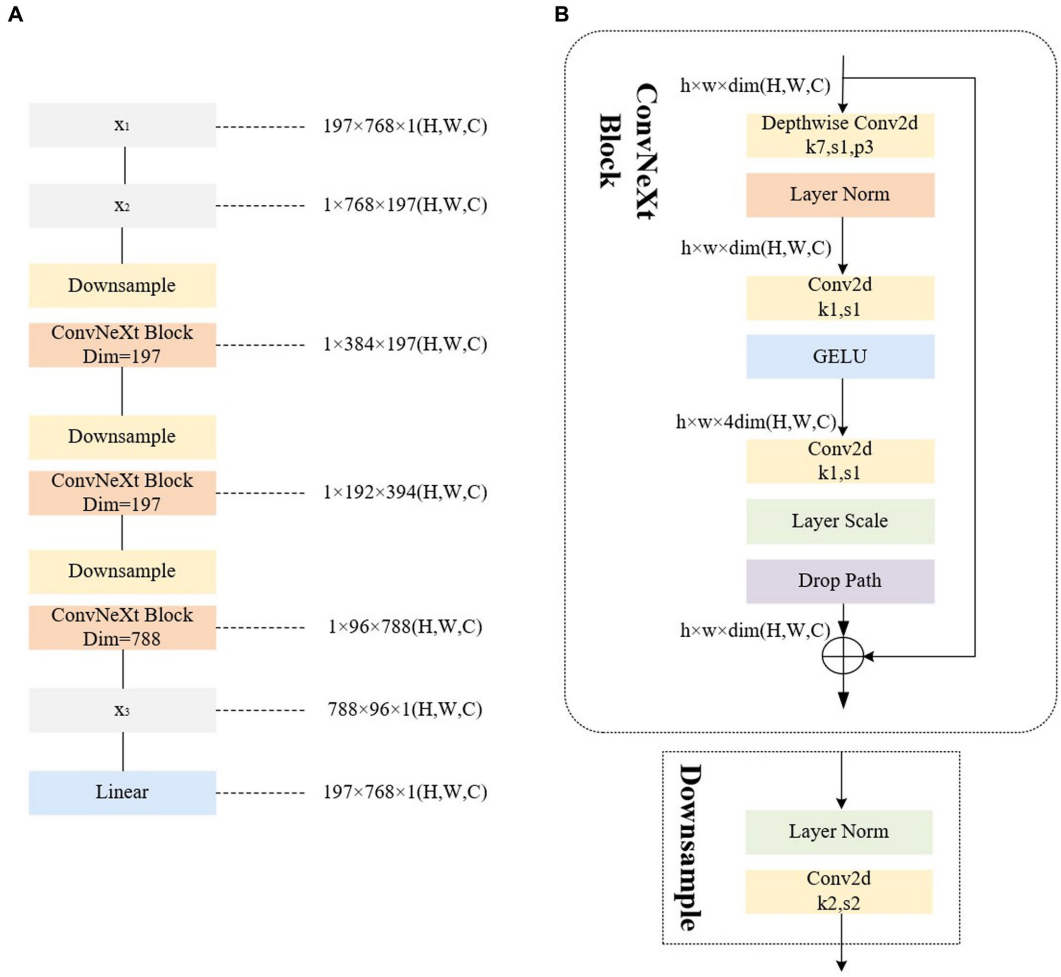



FIGURE 4
Illustrations of the architecture of our improved encoder. **(A)** The improved encoder. **(B)** The ConvNeXt block.

space, enhancing the association with the input data. The final output obtained after sequence pooling can be utilized to derive results through a linear classifier.

## 3. Experiments and results

### 3.1. Dataset

For the experiment, real laparoscopic images from the Hamlyn Centre Laparoscopic/Endoscopic Video Dataset[1] are employed, comprising 5,000 endoscopic images with dimensions of 384 × 192 pixels. As the images constitute a continuous video sequence with minimal differences between adjacent frames, to ensure the robustness of model training and the accuracy of model testing, we adopted a sampling approach. Specifically, we selected 5,000 images from the video dataset at irregular intervals and rendered 1,000 of them to generate a dataset comprising smoke-containing images, as illustrated in Figure 7. The remaining 4,000 images constitute the smoke-free dataset. The selected images is further partitioned into a training set (3,800 images) and a test set (1,200 images), maintaining a 4:1 ratio between smoke-free and smoke-containing images in each set. This balanced distribution ensures effective model training and evaluation.

This paper introduces Blender,[2] a 3D graphic image engine, for software rendering to generate smoke-containing images, which enhances the neural network training dataset. The integration of software rendering addresses the limitation of smoke images in the real endoscopy image dataset. The Blender physical rendering engine is utilized to create realistic and accurate smoke textures, enabling the generation of simulated smoke with random shapes and densities. The rendered smoke possesses local color and transparency, with its position controlled by input parameters: random intensity ($T_{rand}$), density ($D_{rand}$), and position of smoke generation ($P_{rand}$). The smoke image is defined as follow:

1 http://hamlyn.doc.ic.ac.uk/vision/

2 https://www.blender.org/

$$I_{smoke'}(x,y) = \text{Blender}(T_{\text{rand}}, D_{\text{rand}}, P_{\text{rand}}) \quad (1)$$

The smoke image, denoted as $I_{smoke}(x, y)$, is synthesized by utilizing the luminance values of RGB channels. By fusing this rendered smoke with the laparoscopic image, the smoke-containing image is defined as follow:

$$I_{simage}(x,y) = I_{original}(x,y) + I_{smoke}(x,y) \quad (2)$$

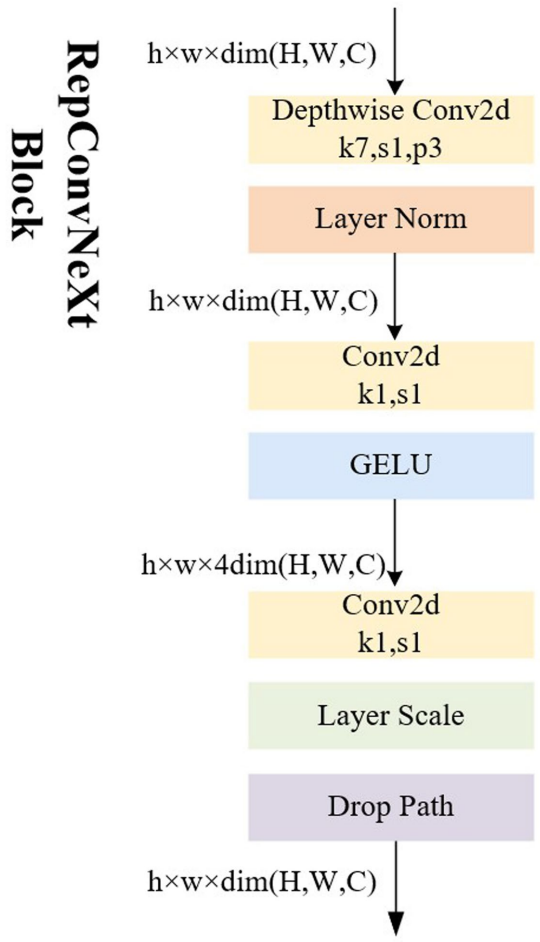



FIGURE 5
Illustrations of the architecture of RepConvNeXt Block.

## 3.2. Experimental platforms

The experimental platform used in this study consists of a Windows 10 operating system, 8 GB RAM, a single NVIDIA 2080Ti 11 GB GPU, and a sixth-generation Intel® Core™ i5 (4C4T) processor. CUDA 10.2, the computing platform provided by NVIDIA, is installed on this platform. The PyTorch 1.8.1 framework is employed to implement the endoscopic smoke image classification algorithm presented in this paper.

## 3.3. Experimental setup

In the training process of endoscopic smoke image classification, the hyperparameters for image training were set as follows: The dataset images were resized to a size of 224 × 224 using the transforms. Resize function as input to the Convolutional Tokenization layer. An exponential decay method was applied to adjust the learning rate, starting with an initial learning rate of 0.001. To enhance the number of Poolformer encoders and prevent overfitting, L = 10 was employed, and data augmentation was implemented through random level inversion. The training was conducted using a 10-fold cross-validation method with 50 epochs.

The experiments were conducted by the controlled variable method on endoscopic images for multiple separate groups, including the following network architectures: mobilenet_v3 (Howard et al., 2019), efficientnet_b7 (Tan and Le, 2019), the ViT network (ViT-B/16) (Dosovitskiy et al., 2021), Poolformer network with Token Mixer changed from attention to pooling layer (Poolformer_s12) (Yu et al., 2022), improved Poolformer network with the utilization of multiplexed branching structure akin to ConvNeXt during training, and improved Poolformer network with the utilization of multi-branch structure during training and single-path structure during prediction.

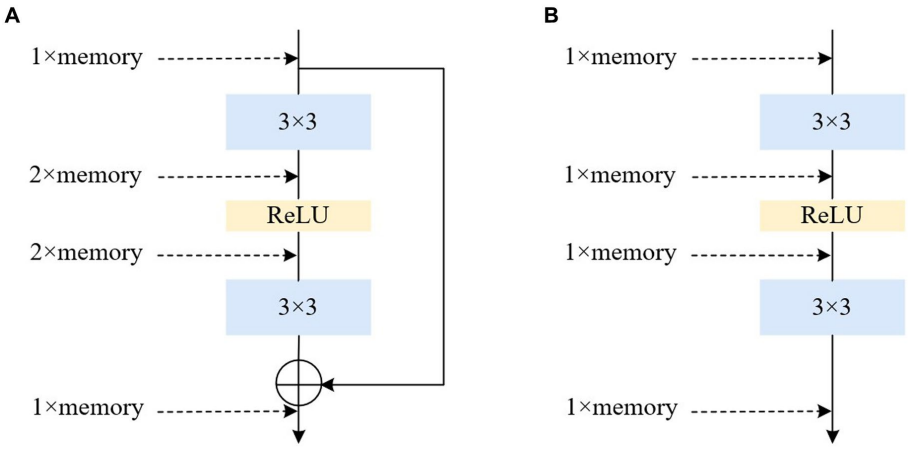



FIGURE 6
Illustrations of the architecture of multiple and single path. **(A)** Residual. **(B)** Plain.

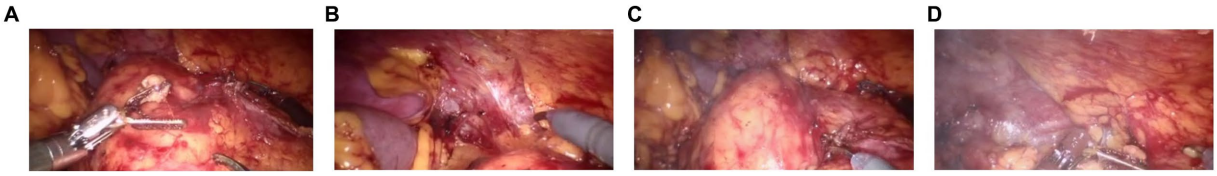



FIGURE 7
Experimental data set. **(A,B)** Original image. **(C,D)** Synthesized image with smoke.

TABLE 1 The results of comparable experiments on different classification model.

| Model | Acc/% | Sens/% | Inference Speed /fps |
|---|---|---|---|
| mobilenet_v3 | 93.9 | 78.6 | 56.2 |
| efficientnet_b7 | 94.3 | 80.3 | 47.8 |
| VIT-B/16 | 94.5 | 80.8 | 42.6 |
| Our method | 95.9 | 83.5 | 87.1 |

# 4. Results

## 4.1. Evaluation metrics

For the classification algorithm of endoscopic images based on Poolformer, which is adopted in this paper, the metrics used for evaluation include Accuracy (Acc), Sensitivity (Sens), and inference speed/ frames per second (fps).

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{3}$$

$$Sens = \frac{TP}{TP + FN} \tag{4}$$

where TP represents the number of true positive samples (images with smoke correctly predicted as images with smoke), FP represents the number of false positive samples (smoke-free images incorrectly predicted as images with smoke), FN represents the number of false negative samples (images with smoke incorrectly predicted as smoke-free images), and TN represents the number of true negative samples (smoke-free images correctly predicted as smoke-free images).

## 4.2. Method comparison

To verify the effectiveness of the model, multiple sets of comparison experiments were conducted using the same endoscopic image dataset and smoke rendering scenarios, along with consistent settings for the remaining experimental parameters. The results were averaged over five runs, and the performance of different detection models on the dataset is presented in Table 1. Among the networks for comparison, all are lightweight neural networks designed for low-power devices, except for the classic ViT-B/16 network. The results reveal that in comparison to the mobilenet_v3, efficientnet_b7, and ViT-B/16 models, the proposed model demonstrates improvements in accuracy by 2, 1.6, and 1.4%, along with enhancements in sensitivity by 4.9, 3.2, and 2.7%, respectively. Furthermore, the proposed model achieves superior processing speed performance, with a frame rate increase of 30.9, 39.3, and 44.5 fps when compared to the mentioned models. These comparative experiments highlight the efficacy of the paper's approach in conducting more accurate, comprehensive, and expeditious screening of smoke-containing images within endoscopic scenes, surpassing these existing modeling methodologies.

## 4.3. Ablation experiment

To evaluate the effectiveness of the improved multi-branch structure and the single-path inference process, we compare the performance of the original Poolformer model with versions that incorporate the multi-path structure alone and in combination with the single-path structure for real endoscopic image classification. The comparative experiments are presented in Table 2. The results demonstrate that the enhanced model, which incorporates a multiplexed branching structure, surpasses the original Poolformer model in terms of classification performance on the dataset. Specifically, the enhanced model exhibited a 2.8% enhancement in accuracy and a notable 9.6% increment in sensitivity. This outcome substantiates the efficacy of replacing the conventional pooling layer with a multiplexed branching structure within the Poolformer architecture, effectively bolstering detail retention within the endoscopic environment. However, the incorporation of this structure introduced a minor drawback, resulting in a reduction of processing speed by 26.3 fps. Further refinement of the model, encompassing a training process enriched with the multiplexed branching structure and a prediction network strengthened by a single-path topology, yielded commendable results. This adaptation yielded a 2.3% enhancement in accuracy and an 8.2% augmentation in sensitivity. Remarkably, this performance boost incurred only a marginal 6.4 fps decline in processing speed compared to the original

TABLE 2 Ablations study for each component of our method.

| Seq | Poolformer_s12 | Multi-branch Structure | Single-path Structure | Acc/% | Sens/% | Inference Speed /fps |
|---|---|---|---|---|---|---|
| 1 | √ | | | 93.6 | 75.3 | 93.5 |
| 2 | √ | √ | | 96.4 | 84.9 | 67.2 |
| 3 | √ | √ | √ | 95.9 | 83.5 | 87.1 |

Poolformer model. Thus, the strategic integration of the multiplexed branching structure into the training network emerged as a viable approach to amplify detail retention in the endoscopic environment. The incorporation of RepConvNeXt structure concurrently elevated processing speed, thereby enhancing endoscopic smoke classification performance and reducing processing time. Conclusively, the experimental results demonstrate the significant capability of the approach proposed in this study. This approach effectively enhances the detection prowess of the Poolformer model in the endoscopic image while concurrently sustaining its efficient real-time operational cadence.

## 5. Conclusion

This paper introduces an improved Poolformer model for the automatic classification and recognition of endoscopic images containing smoke. The proposed model enhances the Token Mixer in the encoder by replacing the simple pooling layer with a multiplexed branching structure, similar to the pure convolutional neural network ConvNeXt. During the prediction process, the structure transforms into single-way, further improving the inference speed.

The experimental findings establish the superiority of our proposed method in the field of endoscopic image classification. In comparison to the traditional ViT-B16 network and the newer, lightweight networks including mobilenet_v3 and efficientnet_b7, our model exhibits substantial improvements. Specifically, it achieves an enhanced accuracy of 1.4, 2, and 1.6%, alongside sensitivity improvements of 2.7, 4.9, and 3.2%, respectively. Notably, these enhancements are accompanied by a significant boost in inference speed, with improvements of 44.5, 30.9, and 39.3 fps, respectively. These performance gains are attained without any appreciable degradation in image processing speed, underscoring the model's efficiency. Furthermore, in contrast to the Poolformer framework, our model achieves these performance enhancements while maintaining image processing speeds, thus ensuring real-time processing remains unaffected. Comparatively, when compared to Poolformer_s12, our proposed method excels further, achieving an accuracy increase of 2.3% and a sensitivity boost of 8.2%. Although there is a marginal reduction in processing speed by 6.4 fps, these trade-offs emphasize the method's prowess in smoke feature recognition and real-time processing efficiency within endoscopic environments. This method serves as an effective means for real-time screening of smoke-containing images in endoscopes, paving the way for potential integration with smoke removal techniques. Such integration can lead to more targeted and precise desmoking, avoiding the issues arising from the enhancing of smoke-free images, notably mitigating computational overhead. By introducing real-time smoke detection into endoscopic procedures, we aspire to reduce equipment resource requirements, augment processing speed, and enhance the real-time, precision, and scientific validity of smoke removal in endoscopic settings.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

HW: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Writing – original draft, Writing – review & editing. KW: Conceptualization, Data curation, Methodology, Writing – original draft, Software, Validation, Writing – review & editing. TY: Conceptualization, Data curation, Methodology, Validation, Writing – review & editing. HZ: Conceptualization, Data curation, Methodology, Writing – review & editing. EC: Conceptualization, Data curation, Methodology, Writing – review & editing. YL: Conceptualization, Funding acquisition, Investigation, Methodology, Writing – review & editing. YW: Conceptualization, Funding acquisition, Methodology, Writing – review & editing. JL: Conceptualization, Methodology, Writing – review & editing, Funding acquisition. YP: Conceptualization, Funding acquisition, Methodology, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The project was funded by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJQN202100602, KJQN202300637, KJQN202300613 and KJQN202000604; Chongqing Technical Innovation and Application Development Special Project under Grant No. CSTB2022TIAD-KPX0062 and cstc2021jscx-gksbx0051; Project funded by China Postdoctoral Science Foundation under Grant No. 2022MD713702, Special Postdoctoral Support from Chongqing Municipal People's Social Security Bureau under Grant No. 2021XM3010; Nature Science Foundation of Chongqing under Grant No.CSTC2021JCYJ-BSH0221,2022NSCQ-LZX0254andCSTB2022NSCQ-MSX1523, National Natural Science Foundation of China under Grant No. U21A20447 and 61971079, Chongqing Innovation Group Project under Grant No. cstc2020jcyj- cxttX0002.

## Conflict of interest

HW and YW were employed by Chongqing Xishan Science & Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Aceves, S. S., Alexander, J. A., Baron, T. H., Bredenoord, A. J., Day, L., Dellon, E. S., et al. (2022). Endoscopic approach to eosinophilic esophagitis: American Society for Gastrointestinal Endoscopy consensus conference. *Gastrointest. Endosc.* 96, 576–592.e1. doi: 10.1016/j.gie.2022.05.013

Ahn, Y. (2020). The current state of cervical endoscopic spine surgery: an updated literature review and technical considerations. *Expert Rev. Med. Devices* 17, 1285–1292. doi: 10.1080/17434440.2020.1853523

Almeida, J. S., Huang, C., Nogueira, F. G., Bhatia, S., and de Albuquerque, V. H. C. (2022). EdgeFireSmoke: a novel lightweight CNN model for real-time video fire–smoke detection. *IEEE Trans. Industr. Inform.* 18, 7889–7898. doi: 10.1109/TII.2021.3138752

Azam, M. A., Khan, K. B., Rehman, E., and Khan, S. U. (2022). Smoke removal and image enhancement of laparoscopic images by an artificial multi-exposure image fusion method. *Soft. Comput.* 26, 8003–8015. doi: 10.1007/s00500-022-06990-4

Bastier, P. L., Gallet de Santerre, O., Bartier, S., De Jong, A., Trzepizur, W., Nouette-Gaulain, K., et al. (2022). Guidelines of the French society of ENT (SFORL): drug-induced sleep endoscopy in adult obstructive sleep apnea syndrome. *Eur. Ann. Otorhinolaryngol. Head Neck Dis.* 139, 216–225. doi: 10.1016/j.anorl.2022.05.003

Boese, A., Wex, C., Croner, R., Liehr, U. B., Wendler, J. J., Weigt, J., et al. (2022). Endoscopic imaging technology today. *Diagnostics* 12:1262. doi: 10.3390/diagnostics12051262

Bolkar, S., Wang, C., Cheikh, F. A., and Yildirim, S. (2018). "Deep smoke removal from minimally invasive surgery videos", In 2018 25th IEEE International Conference on Image Processing (ICIP), 3403–3407.

Chadebecq, F., Lovat, L. B., and Stoyanov, D. (2023). Artificial intelligence and automation in endoscopy and surgery. *Nat. Rev. Gastroenterol. Hepatol.* 20, 171–182. doi: 10.1038/s41575-022-00701-y

Dewangan, A., Pande, Y., Braun, H.-W., Vernon, F., Perez, I., Altintas, I., et al. (2022). FIgLib & SmokeyNet: dataset and deep learning model for real-time wildland fire smoke detection. *Remote Sens.* 14:1007. doi: 10.3390/rs14041007

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J. (2021). "Repvgg: making vgg-style convnets great again", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 13733–13742.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations, 2021*.

Fu, Z., Jin, Z., Zhang, C., He, Z., Zha, Z., Hu, C., et al. (2021). The future of endoscopic navigation: a review of advanced endoscopic vision technology. *IEEE Access* 9, 41144–41167. doi: 10.1109/ACCESS.2021.3065104

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L. -C., Chen, B., Tan, M., et al. (2019). "Searching for mobilenetv3", In Proceedings of the IEEE/CVF international conference on computer vision, 1314–1324.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint* [Epub ahead of preprint].

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks", In Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Proces. Syst.* 25, 1097–1105.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lin, J., Jiang, M., Pang, Y., Wang, H., Chen, Z., Yan, C., et al. (2021). A desmoking algorithm for endoscopic images based on improved U-net model. *Concurr. Comput.* 33:e6320. doi: 10.1002/cpe.6320

Liu, Z., Mao, H., Wu, C. -Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). "A convnet for the 2020s", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 11976–11986.

Niknam, N., Obanor, S., and Lee, L. A. (2022). Endoscopic methods for the detection and treatment of gastric cancer. *Curr. Opin. Gastroenterol.* 38, 436–442. doi: 10.1097/MOG.0000000000000867

Pan, Y., Bano, S., Vasconcelos, F., Park, H., Jeong, T. T., and Stoyanov, D. (2022). DeSmoke-LAP: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *Int. J. Comput. Assist. Radiol. Surg.* 17, 885–893. doi: 10.1007/s11548-022-02595-2

Poutoglidis, A., Fyrmpas, G., Vlachtsis, K., Paraskevas, G. K., Lazaridis, N., Keramari, S., et al. (2022). Role of the endoscope in cochlear implantation: a systematic review. *Clin. Otolaryngol.* 47, 708–716. doi: 10.1111/coa.13909

Simpson, A. K., Lightsey, H. M., Xiong, G. X., Crawford, A. M., Minamide, A., and Schoenfeld, A. J. (2022). Spinal endoscopy: evidence, techniques, global trends, and future projections. *Spine J.* 22, 64–74. doi: 10.1016/j.spinee.2021.07.004

Su, X., and Wu, Q. (2023). Multi-stages de-smoking model based on CycleGAN for surgical de-smoking. *Int. J. Mach. Learn. Cybern.* doi: 10.1007/s13042-023-01875-w

Tan, M., and Le, Q. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks", In International conference on machine learning: PMLR, 6105–6114.

Venkatesh, V., Sharma, N., Srivastava, V., and Singh, M. (2020). Unsupervised smoke to desmoked laparoscopic surgery images using contrast driven cyclic-DesmokeGAN. *Comput. Biol. Med.* 123:103873. doi: 10.1016/j.compbiomed.2020.103873

Wang, C., Mohammed, A. K., Cheikh, F. A., Beghdadi, A., and Elle, O. J. (2019). "Multiscale deep desmoking for laparoscopic surgery," in *SPIE medical Imaging: SPIE*. (San Diego, California, US: SPIE Medical Imaging), 505–513.

Wang, F., Sun, X., and Li, J. (2023). Surgical smoke removal via residual Swin transformer network. *Int. J. Comput. Assist. Radiol. Surg.* 18, 1417–1427. doi: 10.1007/s11548-023-02835-z

Yamashita, S., Inoue, T., Kohjimoto, Y., and Hara, I. (2022). Comprehensive endoscopic management of impacted ureteral stones: literature review and expert opinions. *Int. J. Urol.* 29, 799–806. doi: 10.1111/iju.14908

Yi, Y., Li, L., Li, J., Shu, X., Kang, H., Wang, C., et al. (2023). Use of lasers in gastrointestinal endoscopy: a review of the literature. *Lasers Med. Sci.* 38:97. doi: 10.1007/s10103-023-03755-9

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., et al. (2022). "Metaformer is actually what you need for vision", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 10819–10829.

Zhao, L., Liu, J., Peters, S., Li, J., Oliver, S., and Mueller, N. (2022). Investigating the impact of using IR bands on early fire smoke detection from Landsat imagery with a lightweight CNN model. *Remote Sens.* 14:3047. doi: 10.3390/rs14133047

Zhou, Y., Hu, Z., Xuan, Z., Wang, Y., and Hu, X. (2022). Synchronizing detection and removal of smoke in endoscopic images with cyclic consistency adversarial nets. *IEEE/ACM Trans. Comput. Biol. Bioinform.* PP, 1–12. doi: 10.1109/TCBB.2022.3204673

Zou, X., Zhang, G., Xie, T., Yuan, Y., Xiao, R., Wu, G., et al. (2020). Natural orifice transluminal endoscopic surgery in urology: the Chinese experience. *Asian J. Urol.* 7, 1–9. doi: 10.1016/j.ajur.2019.07.001

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 05 October 2023
DOI 10.3389/fnins.2023.1270850



OPEN ACCESS

EDITED BY
Teng Li,
Anhui University, China

REVIEWED BY
Li Cao,
Wuhan Polytechnic University, China
Xiaomin Yang,
Sichuan University, China

*CORRESPONDENCE
Sen Zhou
✉ cquzhousen@163.com

RECEIVED 01 August 2023
ACCEPTED 04 September 2023
PUBLISHED 05 October 2023

CITATION
Bai T, Zhou S, Pang Y, Luo J, Wang H and Du Y (2023) An image caption model based on attention mechanism and deep reinforcement learning.
*Front. Neurosci.* 17:1270850.
doi: 10.3389/fnins.2023.1270850



# An image caption model based on attention mechanism and deep reinforcement learning

Tong Bai[1], Sen Zhou[2*], Yu Pang[1], Jiasai Luo[1], Huiqian Wang[1] and Ya Du[3]

[1]School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China, [2]Chongqing Academy of Metrology and Quality Inspection, Chongqing, China, [3]Department of Peripheral Vascular (Wound Repair), Chongqing Hospital of Traditional Chinese Medicine, Chongqing, China

Image caption technology aims to convert visual features of images, extracted by computers, into meaningful semantic information. Therefore, the computers can generate text descriptions that resemble human perception, enabling tasks such as image classification, retrieval, and analysis. In recent years, the performance of image caption has been significantly enhanced with the introduction of encoder-decoder architecture in machine translation and the utilization of deep neural networks. However, several challenges still persist in this domain. Therefore, this paper proposes a novel method to address the issue of visual information loss and non-dynamic adjustment of input images during decoding. We introduce a guided decoding network that establishes a connection between the encoding and decoding parts. Through this connection, encoding information can provide guidance to the decoding process, facilitating automatic adjustment of the decoding information. In addition, Dense Convolutional Network (DenseNet) and Multiple Instance Learning (MIL) are adopted in the image encoder, and Nested Long Short-Term Memory (NLSTM) is utilized as the decoder to enhance the extraction and parsing capability of image information during the encoding and decoding process. In order to further improve the performance of our image caption model, this study incorporates an attention mechanism to focus details and constructs a double-layer decoding structure, which facilitates the enhancement of the model in terms of providing more detailed descriptions and enriched semantic information. Furthermore, the Deep Reinforcement Learning (DRL) method is employed to train the model by directly optimizing the identical set of evaluation indexes, which solves the problem of inconsistent training and evaluation standards. Finally, the model is trained and tested on MS COCO and Flickr 30 k datasets, and the results show that the model has improved compared with commonly used models in the evaluation indicators such as BLEU, METEOR and CIDEr.

KEYWORDS

**image caption, encoder-decoder architecture, deep neural networks, attention mechanism, deep reinforcement learning**

## 1. Introduction

In recent years, profound advances have been made in deep learning technology due to the breakthrough in computing power of computers and the surge in data (LeCun et al., 2015). Meanwhile, image caption based on deep learning has also seen significant improvements (Bai and An, 2018; Srivastava and Srivastava, 2018; Liu et al., 2019). Image caption is

the intersection of the fields of computer vision and natural language processing, along with its potential value in terms of contributing to visually impaired individuals' daily life assistance, graphic conversion, automatic title generation and machine intelligence (Hossain et al., 2019; Kang and Hu, 2022). Fundamentally, it involves utilizing techniques grounded in deep learning to interpret a given image and automatically generate descriptive text as if the machine is looking at an image and speaking. Despite its intuitive nature for humans, this process is highly challenging for machines, requiring the accurate interpretation of image content, object relationships and the synthesis of appropriate language. As such, significant research efforts are still required to achieve reliable and effective image caption models that match human-level performance (Anderson et al., 2016; Bernardi et al., 2016).

The advancement of image caption technology is of profound importance in terms of both research and practical application. Its significance is particularly evident in the following areas: firstly, in the field of visual assistance systems, image caption can play a vital role in helping the visually impaired access crucial visual information (Jing et al., 2020; Bhalekar and Bedekar, 2022). By expressing image content comprehensively and concretely, this technology can reduce the obstacles that the visually impaired face in their learning and daily life. Secondly, due to the widespread deployment of cameras and the increasing amount of monitoring data being acquired, the workloads of surveillance personnel have become overwhelming. A system based on image caption can provide summarized information of the monitoring data leading to more efficient work processes (Nivedita et al., 2021). Overall, with the continuous development and maturity of deep learning theory, image caption technology will undoubtedly have an increasingly significant impact on people's lifestyles, advancing progress across society and industry (Amritkar and Jabade, 2018; Kinghorn et al., 2018).

Image caption has broad application prospects, and more and more researchers begin to study this challenging task. Before the introduction of encoder-decoder architecture, two primary approaches had emerged in the early stages, template-based method and search-based method. The template-based approach generates the final caption from a pre-set sentence template. Farhadi et al. (2010) use detectors to detect objects to form descriptions of images based on language templates. Other researchers use independent corpus construction and more effective semantic analysis models to describe the images. Elliott and de Vries (2015) express target objects in images by means of visual dependency representation, selects the target objects corresponding to the most appropriate features, and fills them in the template. After continuous improvement of the template-based method, although the main object of the image can be recognized accurately, the generated sentences are monotonous and lack some semantic information. The search-based method involves using similarity algorithms to compute the similarity between extracted features and the images stored in a constructed image library, to find out the images in line with the algorithm, and these images have been matched with the corresponding sentence descriptions in advance, which can be fine-tuned for appropriate output. Verma et al. (2013) adopt traditional image feature extraction methods to compare the extracted image features with those in the database, so as to determine the maximum joint probability output in the description tuple. Li and Jin (2016) introduce the reordering mechanism which greatly improves the model performance. The search-based method relies heavily on the constructed search image library, and the results have great uncertainty and poor robustness.

The image caption model based on encoder-decoder architecture is derived from the machine translation model (Cho et al., 2014). The encoder-decoder architecture can directly realize the mapping between the images and the descriptions by learning. And the deep neural network model can learn these mappings from a large amount of data to generate a more accurate descriptions, which makes this method have greater improvement in performance compared with the previous methods. The Multimodal Recurrent Neural Network (M-RNN) model is proposed in Mao et al. (2014), stands out as a pioneering approach utilizing an encoder-decoder architecture, effectively bridging the gap between image and text features through modal fusion. The Neural Image Caption (NIC) model proposed in Vinyals et al. (2015) adopt Long Short-Term Memory (LSTM) to replace RNN, which effectively improves performance and is also the baseline model for many subsequent methods. Deng et al. (2020) introduce an adaptive attention model with a visual sentinel, and introduces the Dense Convolutional Network (DenseNet) to extract the global features of the image in the encoding phase, which significantly improves the quality of image caption generation. Fei (2021) propose a memory-augmented method, which extends an existing image caption model by incorporating extra explicit knowledge from a memory bank, and the experiments demonstrate that this method holds the capability for efficiently adapting to larger training datasets. In Shakarami and Tarrah (2020), an efficient image caption method using machine learning and deep learning is proposed. The experimental results demonstrate the superiority of the offered method compared to existing methods by improving the accuracy. Huang et al. (2019) propose an Attention on Attention (AoA) network for both the encoder and the decoder of the image caption model, which extends the conventional attention mechanisms to determine the relevance between attention results and queries. Krause et al. (2017) use faster-RCNN to acquire regional features and combine them, and then uses multi-layer recurrent neural networks to get the image caption. There are several other improvements (Yang et al., 2019; Liu et al., 2020; Parikh et al., 2020; Singh et al., 2021) that are based on this encoder-decoder architecture. This kind of method is characterized by its flexibility and strong generalization ability. At present, most improvements are based on encoder-decoder architecture.

With the development of technology, the performance of image caption has been made substantial advancements compared with traditional methods (Liu et al., 2020). However, there are several challenges persist, including shortcomings in the encoding and decoding processes, loss of visual information during decoding, insufficient attention to detail information, and discrepancies between training objectives and evaluation indicators. To address these issues, this paper studies and optimizes the image caption model with encoder-decoder architecture. The structure of the paper is arranged as follows: section 2 puts forward the image caption model based on guided decoding and feature fusion. Section 3 further improves the performance of the image caption

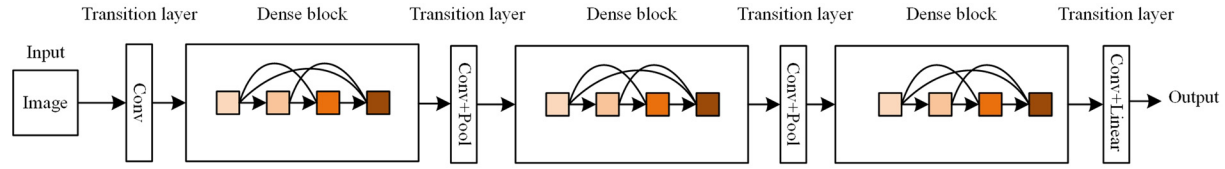



FIGURE 1
DenseNet model.

model. Section 4 provides the experimental process and result analysis. Finally, the conclusion of our image caption model is in section 5.

## 2. Image caption model based on guided decoding and feature fusion

In order to solve the problems in image caption technology, this paper proposes an image caption model based on guided decoding and feature fusion. Based on the encoder-decoder architecture, DenseNet model is used to encode image features, and the Multiple Instance Learning (MIL) method is used to extract the image visual information. The two parts together constitute the encoding process of image visual information, and the guided decoding module is adopted to dynamically adjust the input image visual information during the decoding process. The decoder uses a Nested Long Short-Term Memory (NLSTM) network, which can learn more hidden information by increasing the depth of the network model.

### 2.1. Encoder design based on feature fusion

Convolutional Neural Network (CNN) is a crucial model for processing visual image problems and have significantly improved with each architecture iteration. Typically, lower-level features are utilized to distinguish between various classes of basic contour information, while higher-level features are more abstract and effectively differentiate between different varieties of semantic information for the same target. From this perspective, the deeper the layers of the network model, the richer the information extracted. However, the consequent problem is that the increase in model depth causes the gradient to diminish until it disappears during the transfer process. The problem of gradient disappearance can be solved to some extent by using the Batch Normalization (BN) method (Bjorck et al., 2018). Residual Network (ResNet) and highway network also address the problem of gradient disappearance and model degradation by using bypass settings and gating units (Shaked and Wolf, 2017). Nevertheless, these models are prone to excessive parameters and depth redundancy. In image caption tasks, where image scenes are rich, it is necessary not only to identify targets but also to be able to abstractly describe the interconnections between targets, so fusing the base feature map with higher-level feature maps is a good way to handle this problem. In this paper, we employ the DenseNet model for image feature extraction, which is based on the architecture as illustrated in Figure 1. The fundamental concept of DenseNet resides in establishing connections between varied depth feature maps, enabling the utilization of both high-level and low-level features to their fullest potential.

DenseNet has been identified to improve feature multiplexing by means of bypass and this not only deepens the network's layer depth, but also amplifies image information availability. Furthermore, it mitigates problems related to gradient disappearance and model degradation while also keeping the number of parameters less than those of deep neural networks such as ResNet. Meanwhile, with the increase in layer depth, optimization of the network does not become more convoluted. The model's accuracy increases proportionally with an increase in parameters, devoid of overfitting occurrences.

For the image caption tasks, the object, attribute and relation detector are trained separately by independent hand-labeled training data. We train our image caption models on datasets that contain multiple images and descriptive sentences corresponding to each image. Different from the tasks of image classification and object detection, in the task of image caption, there are not only nouns, but also verbs, adjectives, adverbs and other parts of speech in the description generated by an image. Therefore, in order to describe the needs of the tasks, it is necessary to construct a word set $D$ composed of 1,200 common words, which basically contains more than 95% of the words that need to be used in the training set, and the remaining words are treated as non-essential words.

Then, we need to extract the corresponding word from the image through the constructed word set. Because the datasets used in this paper did not define and label corresponding words with corresponding bounding boxes, at the same time, the parts of speech are not even marked, typical supervised learning methods are not suitable for this task. Certainly, while image classification can provide corresponding words for a whole image, many words or semantics are only applicable to the subregions of the image. Such generic classifications often fail to enhance model performance. Therefore, this study applies the MIL method to tackle tasks with one-to-many relationships (Dietterich et al., 1997).

In the image caption tasks, each image corresponds to a packet. For each word $w$ in the word set $D$, the packets are divided into positive packets and negative packets according to different image areas, thus forming the input set of the whole MIL model. The classification method is as follows: if the word $w$ in the word set $D$ appears in the corresponding description sentence of an image $I$, then the packet is marked as a positive packet; if the word in the word set has no corresponding word in the description sentence, the

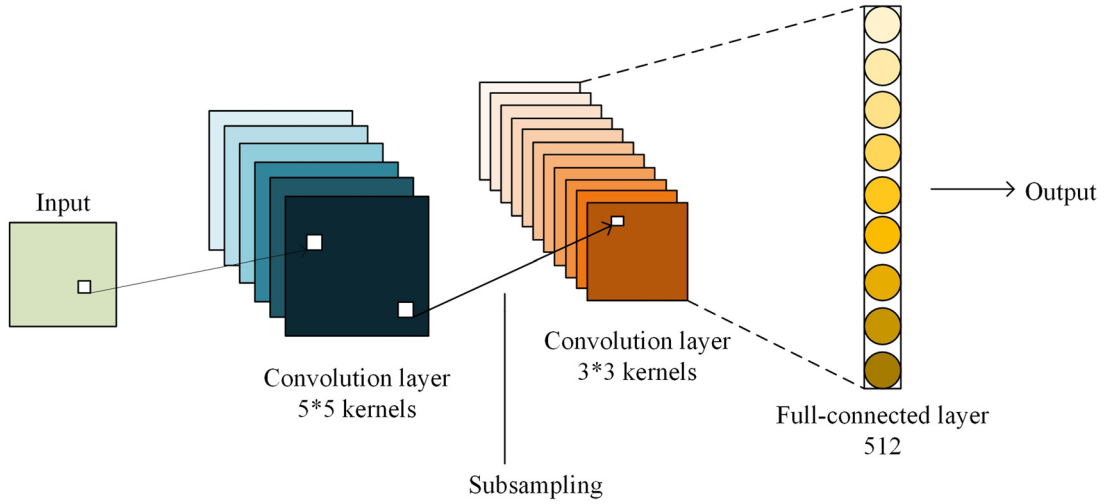



FIGURE 2
The model structure of guided decoding network.

packet is marked as a negative packet. The training set is represented in formula (1).

$$\{(x_1 y_1)\ (x_2 y_2)\ \ldots\ (x_l y_l)\} \tag{1}$$

For the input packet in the training set $x_i$, when $y_i = 1$, it is the positive packet, and when $y_i = -1$, it is the negative packet. Using the MIL model, the probability $P_w$ that each packet $b_i$ contains the word $w$ in the word set $D$ is calculated by the following formula:

$$P_w = 1 - \prod_{j \in b_i} \left(1 - x_{ij}^w\right) \tag{2}$$

Where $x_{ij}^w$ represents the probability that a particular region $j$ in an image $i$ corresponds to the word $w$ in the word set. Since it is image information, the Visual Geometry Group Network (VGGNet) model is used here for calculation. VGG16 model has a total of 16 layers, including 5 convolutional layers, each convolutional layer is followed by a pooling layer, generally using the maximum pooling method. After the convolutional layers, there are 3 fully connected layers, and finally the SoftMax layer is used for classification. The input of the network model is a 224*224 RGB image. The specific calculation process of $x_{ij}^w$ is to adopt a fully connected layer with a sigmoid nonlinear activation function, and the formula is as follows.

$$x_{ij}^w = \frac{1}{1 + \exp\left(-W_w^t \theta\left(b_{ij}\right) + b_w\right)} \tag{3}$$

Where $\theta\left(b_{ij}\right)$ represents the features of region $j$ in the image $i$ extracted by the seventh fully connected layer in the model, $W_w$ and $b_w$, respectively, represent the weight and bias of the word $w$, which can be obtained by learning in model training.

After the operation of the model, a spatial feature map of the image will be obtained in the last fully connected layer, which is corresponding to the position of the input image, that is, the features of different regions in the image. The visual text information of the images in datasets is generated by the MIL model. Generally, the top 10 words with the highest probability after being processed by the MIL model are selected.

In this paper, the image feature extraction module and visual information extraction module will be fused by guiding the decoding module to provide a basis for the subsequent decoding process. In the NIC model of image caption, visual information is only input to the decoder at the beginning of decoding, and the strength of its information features will gradually diminish during the decoding process. The ideal decoder should be able to balance the two-input information of image vision and description, so as to avoid the reduction of decoding accuracy because one information dominates the decoding. Therefore, a CNN model for guided decoding is constructed in this paper. By inputting the learned features into the network for modeling, the modeled guidance vector is sent into each time sequence of the decoder, and at the same time, it can accept the error signal feedback from each time sequence of the decoder and make corresponding adjustments. The introduction of the model structure can realize the complete end-to-end training process. The guided decoding network is a deep neural network composed of two convolutional layers and one fully connected layer, represented by CNN-g. Its model structure is shown in Figure 2.

## 2.2. Decoder design based on NLSTM model

Text information is a critical component of training datasets and plays a vital role in the effectiveness of decoding. To ensure optimal feature extraction and expression, it is necessary to structure raw unstructured text data using a text representation model. This allows for efficient participation in the decoder's training process.

Word to Vector (Word2Vec), a highly effective word embedding model built using shallow neural networks, consists of two main structures: skip-gram and CBOW (Continuous Bag of Words). While skip-gram predicts the probability of generating surrounding words based on the current word, CBOW predicts the generation probability of the current word based on surrounding words. The complexity and variation of the semantic environment in image

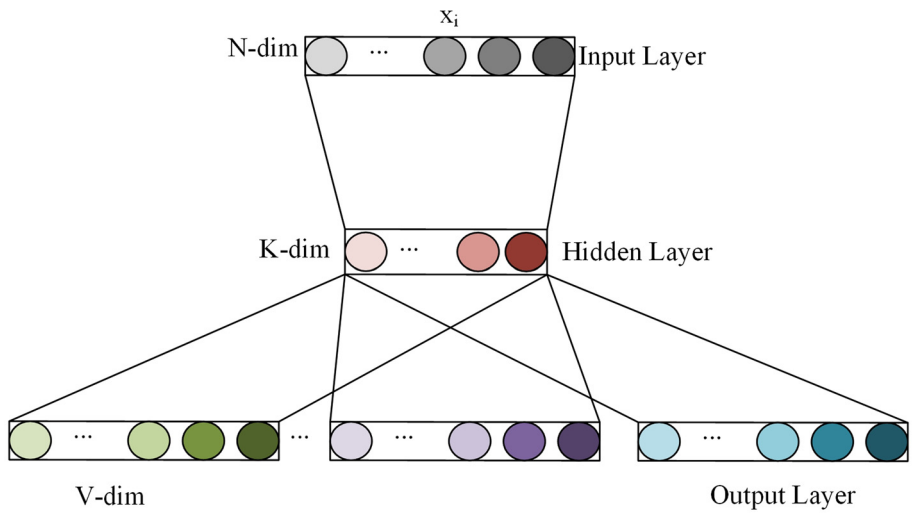



FIGURE 3
The simplified structure of skip-gram.

caption require more precise word embedding inputs. To address this need, this paper adopts the skip-gram model. Skip-gram is a shallow neural network model composed of the input layer, hidden layer and output layer, and its simplified structure is shown in Figure 3. Wherein, each word in the input layer uses one-hot encoding, the size of the training set thesaurus is $N$, and the hidden layer has $K$ hidden units. After the training is completed, any word $x_i$ in the thesaurus can be calculated to get the feature vector with this word as the central word.

In the actual model training process, managing the number of output feature vectors can pose a challenge due to the large volume of training data involved. To address this issue, the hierarchical SoftMax method is leveraged in this paper. This method entails constructing a Huffman coded binary tree based on word frequencies, where high-frequency words are placed at the root node to minimize computations. The tree is organized hierarchically from top to bottom, with each node classified by a sigmoid activation function. The sigmoid activation function determines the probability of the left and right branches of the tree, and the goal of model training is to multiply the probability on the passed branches to reach the maximum value.

In the context of processing and predicting sequence data, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) networks are commonly employed. When it comes to image caption tasks, RNN and LSTM serve as decoders. Among them, LSTM has proven effective in addressing the long-term dependence issue. In this paper, an enhanced NLSTM model is utilized as a decoder to decode input image features. Different from the general LSTM model, in NLSTM, the memory function $c_t$can be obtained through model training as shown in formula (4).

$$c_t = m_t\left(f_t \odot c_{t-1} i_t \odot \mathrm{Tanh}\left(w_c x_t + u_c h_{t-1}\right)\right) \tag{4}$$

Where $m_t$ is a state function learned from NLSTM, and represents the state $m$ at time $t$. $h_t$ and $x_t$ are the input and hidden states of the memory function, respectively. $i_t$ and $f_t$ respectively represent the input gate and forgetting gate. $w_c$ and $u_c$ are learned during training.

In the NLSTM model, the specific calculation method of internal LSTM is obtained by the following formulas:

$$\tilde{i}_t = \tilde{\sigma}_i\left(\tilde{w}_i \tilde{x}_t + \tilde{u}_i \tilde{h}_{t-1} + \tilde{b}_i\right) \tag{5}$$

$$\tilde{f}_t = \tilde{\sigma}_f\left(\tilde{w}_f \tilde{x}_t + \tilde{u}_f \tilde{h}_{t-1} + \tilde{b}_f\right) \tag{6}$$

$$\tilde{o}_t = \tilde{\sigma}_o\left(\tilde{w}_o \tilde{x}_t + \tilde{u}_o \tilde{h}_{t-1} + \tilde{b}_o\right) \tag{7}$$

$$\tilde{c}_t = \tilde{f}_t \odot \tilde{c}_{t-1} + \tilde{i}_t \odot \mathrm{Tanh}\left(\tilde{w}_c \tilde{x}_t + \tilde{u}_c \tilde{h}_{t-1} + \tilde{b}_c\right) \tag{8}$$

$$\tilde{h}_t = \tilde{o}_t \odot \tilde{\sigma}_h\left(\tilde{c}_t\right) \tag{9}$$

Where $\tilde{c}_t$ is the internal memory function, $\tilde{x}_t$ and $\tilde{h}_t$ are the input layer and hidden layer states of the memory function, respectively. $\tilde{i}_t$, $\tilde{f}_t$, and $\tilde{o}_t$ respectively represent the input gate, forgetting gate and output gate of the internal LSTM. To achieve the gating effect in the neural network, the sigmoid function $\tilde{\sigma}$ is commonly used as the activation function, and the Tanh function is utilized as the candidate memory function. The parameters $\tilde{w}$, $\tilde{u}$, and $\tilde{b}$ are learned during training.

The memory unit of the external LSTM is updated according to formula (10).

$$c_t = \tilde{h}_t \tag{10}$$

The value of $h_t$ is then updated through the memory unit $c_t$ of the external LSTM as shown in formula (11).

$$h_t = o_t \odot \mathrm{Tanh}\left(c_t\right) \tag{11}$$

NLSTM uses the standard LSTM network as a gating unit to input relevant information into its memory unit, reducing internal memory burden. This enables a more deterministic time hierarchy and better handling of time series problems compared to stacked models. Finally, a SoftMax layer is used in the model to predict the output words obtained by the final model through the probability distribution of words at time $t$. The structure of the image caption model is shown in Figure 4.

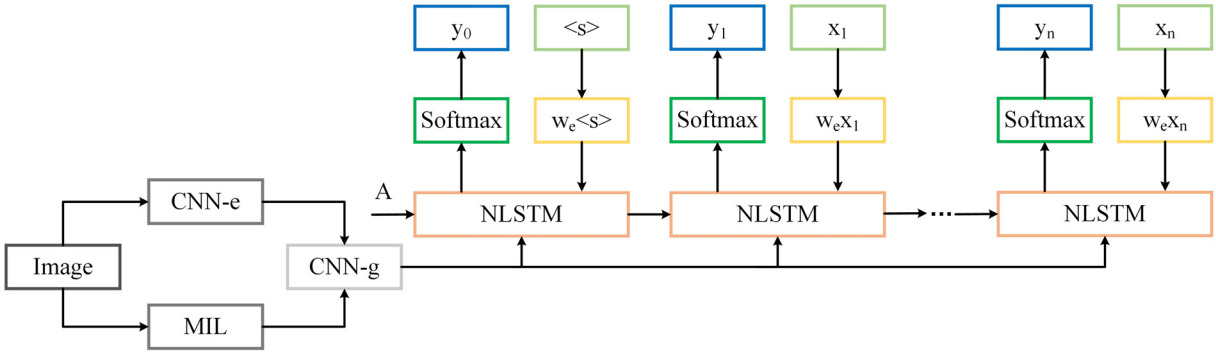



FIGURE 4
The structure of the image caption model.

In Figure 4, CNN-e represents the DenseNet model used in the coding process, and CNN-g is the guided decoding network. The extracted image fusion features are represented by the formula (12).

$$v = f_g(A + M) \tag{12}$$

Where $A$ represents the global image feature, $M$ stands for the visual text information learned from multiple instances, and $f_g$ represents the model function learned by guiding the decoding model.

The decoded output $y_t$ at time $t$ is calculated by formula (13).

$$y_t = w_v v + w_e x_t \tag{13}$$

## 3. Image caption combining attention mechanism and deep reinforcement learning

In order to further improve the performance of the image caption model, we build a double-layer decoding network by introducing the attention mechanism on the basis of the model proposed above. The output of the first layer and the image features are sent to an attention module to extract important detail features. The output of the module is fused with the output of the first layer as the input of the second layer for the second decoding. Meanwhile, considering the powerful perception and decision abilities of Deep Reinforcement Learning (DRL), this paper constructs a training optimization method based on DRL to improve the overall performance of the model.

### 3.1. Attention mechanism

Although the traditional encoder-decoder based image caption model can describe the content of the image in a short text description, it often ignores some local and detailed information in the image during the description process. However, this information is very important to the richness and accuracy of the description. When the attention mechanism was introduced into the image caption task for the first time, which effectively improved the performance of the NIC model. The attention mechanism is inspired by the human process of observing things, people immediately focus on the important information in an image while paying less attention or ignoring irrelevant information or background information. In deep learning, the formation of attention is basically through the way of masks, that is, important information in the image is distinguished by giving different weights. After continuous training of the model, it can learn which regions are important in the image and form more attention to these regions. There are two main types of attention mechanisms: hard attention and soft attention. Here, we represent the feature vector $v$ extracted by the encoder as shown in formula (14).

$$v = \{v_1 v_2 \ \dots \ v_k\}, v_i \in \mathbb{R}^g \tag{14}$$

The output of the last convolutional layer of the DenseNet is used to represent the features of different positions in the image. At different moments of decoding, the attention weights for different regions of the image can be calculated by formula (15).

$$\hat{v}_{it} = f_{att}(v_i h_{t-1}) \tag{15}$$

Where $h_{t-1}$ represents the state of the hidden layer on the decoder LSTM at time $t-1$, $f_{att}$ represents a function that assigns different weights to each region of the image.

The SoftMax function is used to normalize formula (15) so that the weight range is [0,1] and the weighted sum is 1, as shown in formula (16).

$$a_{it} = f_{softmax}(\hat{v}_{it}) \tag{16}$$

Finally, the visual context vectors of different regions of the image are calculated by weight. Its visual context features $\hat{v}_t$ are expressed as shown in the formula (17).

$$\hat{v}_t = \sum_{i=1}^{k} h_{it} v_i \tag{17}$$

Where $h_{it}$ is the multivariate two-point distribution of the input vector $v$, $a_{it}$ is the weight of the different regions of the image in the input decoder at time $t$, as shown in formula (18).

$$(h_{it} v) = a_{it} \tag{18}$$

To obtain local image details during the decoding phase, we propose a double-layer stacked decoding structure, based on the

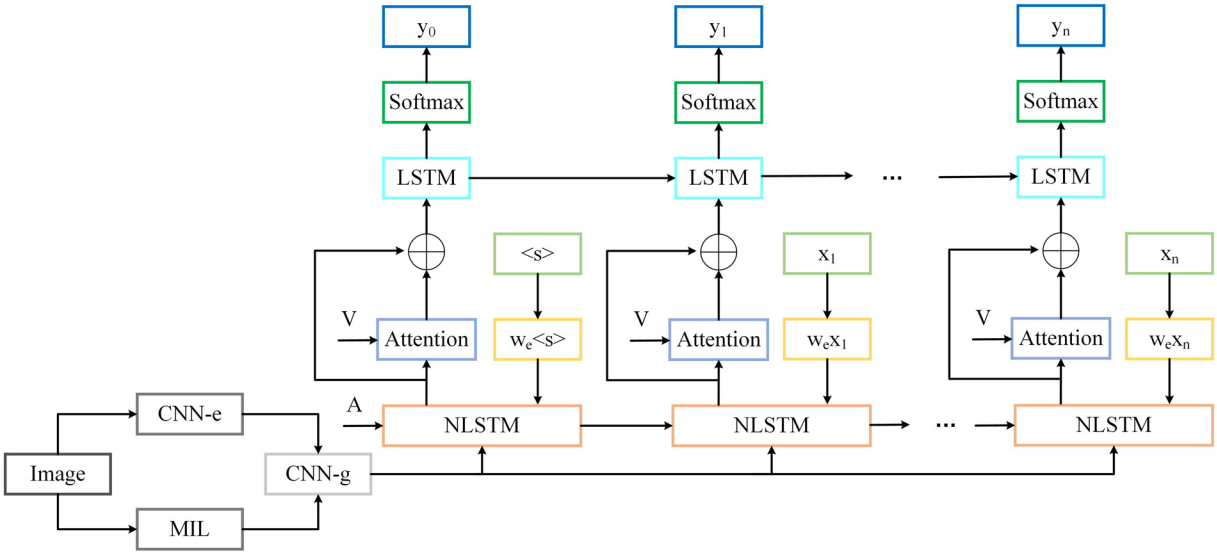



FIGURE 5
The improved structure of the image caption model.

previous model in Figure 4 as the first layer decoding. The new model is depicted in Figure 5. After the output of the first layer decoder and the visual features of the image are calculated by the attention module, they are used as the input of the second layer LSTM decoder by means of residual connection. The introduction of the attention mechanism can effectively improve the performance of the model. The feature vector of the image is represented and calculated by formula (19).

$$v = w^{vi} f_{cnn}(I) \tag{19}$$

Where $I$ represents the input original image after preprocessing, $f_{cnn}$ represents the computational model of DenseNet.

In this model, the last fully connected layer in Figure 4 is removed, and the output of the convolution model is reduced dimensionality by the matrix. The state of the hidden layer of the first layer decoder at time $t$ is calculated by formula (20).

$$h_t^1 = f_{nlstm}\left(x_t h_{t-1} v_g\right) \tag{20}$$

Where $x_t$ represents the input feature vector of word embedding, $h_{t-1}$ represents the hidden layer state at the moment $t-1$, $v_g$ represents the input vector to guide the decoding, and $f_{nlstm}$ stands for the NLSTM network used by the decoder of the first layer.

In the attention module, the image features and the hidden layer state of the first layer decoder are used as inputs, and unlike the hidden layer state of the $t-1$ moment used by the soft attention mechanism, the hidden layer state of the $t$ moment used here is shown in formula (21).

$$\hat{v}_{it} = \mathrm{Tanh}\left(w^v v \oplus w^h h_t^1\right) \tag{21}$$

Where $w^v$ and $w^h$ represent the parameter matrix to be learned by the model, $\oplus$ represents the summation operation of the matrix.

The weight of the attention module is calculated as shown in formula (22).

$$a_t = f_{softmax}\left(w^a \hat{v}_{it}\right) \tag{22}$$

Where $w^a$ represents the parameter matrix to be learned by the model, $f_{softmax}$ represents the SoftMax operation.

Based on the weight of the attention module, we can get the visual attention features of the image $\hat{v}_t$, as shown in formula (23).

$$\hat{v}_t = a_t v \tag{23}$$

Then, by means of residual connection, the visual attention feature is added and fused with the corresponding subscript element of the hidden layer state $h_t$ at $t$ moment of the first layer decoder, as shown in formula (24), and it is used as the input of the second layer decoder.

$$x_t^2 = \hat{v}_t \oplus h_t^1 \tag{24}$$

LSTM is used as the second layer decoder for the final processing of sequence information. The hidden layer state of the second layer decoder is obtained by formula (25).

$$h_t^2 = f_{lstm}\left(x_t^2 h_{t-1}^2\right) \tag{25}$$

Where $h_{t-1}^2$ represents the hidden layer state of the second layer decoder at time $t-1$, $f_{lstm}$ represents the model calculation function of the second layer LSTM.

After the second hidden layer state is obtained, an evaluation module is used to predict the possibility of output words, which is mainly composed of linear layer, fully connected layer and SoftMax layer. The linear layer is used for dimensionality reduction of words output by LSTM, and the fully connected layer is used for the upsampling of vectors after dimensionality reduction. Finally, the probability distribution $y_t$ of word output is calculated through the

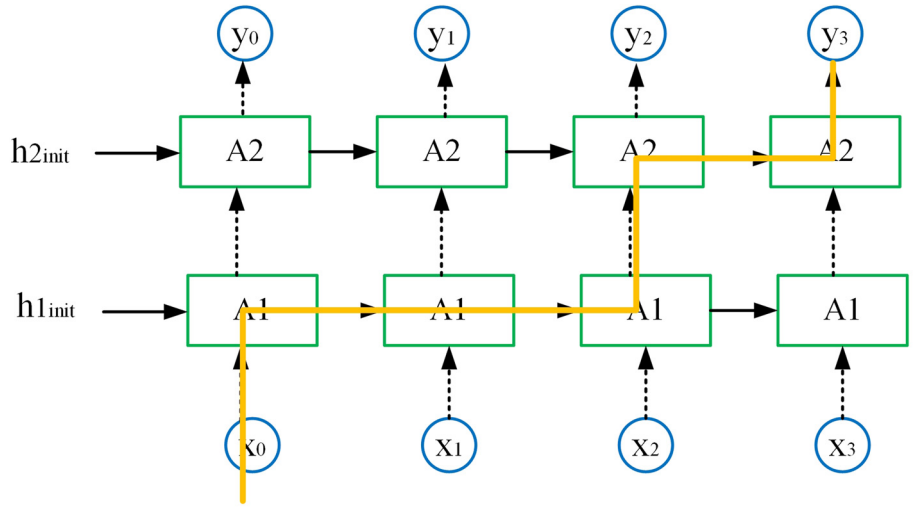



FIGURE 6
Dropout operation in the double-layer decoding structure.

SoftMax layer, as shown in formula (26).

$$y_t = f_{softmax}\left(w^N h_t^2 + b^N\right) \tag{26}$$

With the increase of the number of model layers, the expressiveness of the model is also enhanced. However, this also leads to overfitting problems. To address this issue, this paper adopts the dropout method in the double-layer decoding structure that reduces overfitting. The main idea of this method is to deactivate part of the computing units and keep the other part of the computing units working on the data that flows into each unit. Figure 6 illustrates the implementation of dropout operation in the double-layer decoding structure, at time $t = 0$, input $x_0$ is passed into the first layer of RNN, and then transmission continues in the first layer until time $t = 2$, during which there is no dropout operation. At time $t = 2$, the dropout operation is performed when the first layer passes to the second layer, which is always coherent in timing. The dropout operation helps greatly in improving the robustness of the model.

## 3.2. Deep reinforcement learning

Reinforcement learning is an artificial intelligence learning method. Different from supervised learning and unsupervised learning, reinforcement learning will only make different rewards or punishments according to the quality of actions. DRL not only has the understanding ability of deep learning, but also makes use of reinforcement learning to make decisions and judgments on the environment, and realizes the response and treatment of complex problems through the end-to-end learning process. The framework of DRL is mainly derived from Markov Decision Process (MDP).

The policy gradient algorithm is a frequently adopted technique for DRL. It offers a direct approach to optimize the expected reward of the policy, without relying on intermediate stages, and enables the determination of an optimal policy within the given policy space. The method utilizes an approximation function to directly optimize the policy and achieve the highest expected total reward. The actor-critic architecture diagram for this algorithm is illustrated in Figure 7, with its policy gradient being expressed through the formula (27).

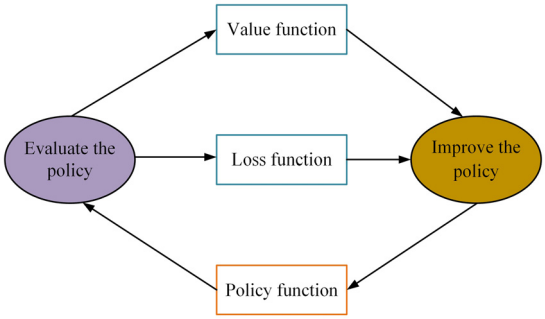



FIGURE 7
Policy gradient learning method based on actor-critic architecture.

$$g_p = \mathrm{E}\left(\sum_{t=0}^{\infty} \psi_t \nabla_\theta log\pi_\theta\left(a_t s_t\right)\right) \tag{27}$$

Where $\pi_\theta\left(a_t s_t\right)$ represents the policy function, which is learned by the neural network in DRL, $\psi_t$ represents the evaluation function, which is approximated by a neural network.

The policy function can guide the agent's actions. The guidance process is calculated according to the probability of taking an action in a certain state, and it is a mapping function from state to action. At the same time, the optimal policy is selected to guide the value function through policy evaluation. The value function is the state value function under the guidance of the policy. The policy function $\theta_t$ is updated by formula (28) during the learning process. The value function $w_t$ is updated by the formula (29).

$$\theta_{t+1} = \theta_t + a\delta\nabla_\theta log\pi\left(a_t s_t \theta_t\right) \tag{28}$$

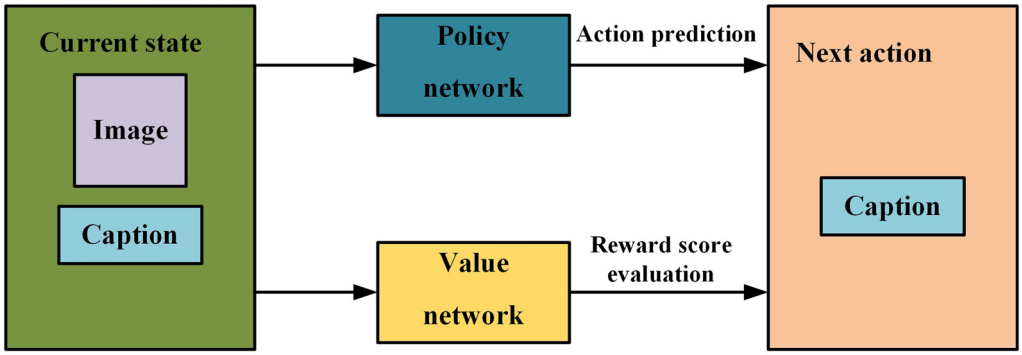



FIGURE 8
The model structure and prediction process based on DRL.

$$w_{t+1} = w_t + \beta\delta\nabla_w \hat{v}\left(s_t w_t\right) \tag{29}$$

Where $a_t$ and $s_t$, respectively, represent the action and state at time $t$.

Considering the powerful perception and decision abilities of DRL, we use it to further optimize our image caption model. And on the basis of the actor-critic structure, two kinds of deep neural networks, policy network and value network, are used to construct models for predicting words that best describe the image in each state. Specifically, the policy network evaluates the confidence of the next predicted word based on the current state, and thus suggests the next possible action to be taken. The value network evaluates the reward scores of the actions predicted by the policy network in the current state, and decides whether to choose the actions given by the policy network according to these reward scores. In other words, the model's predictions are constantly adjusted according to the actual situation for producing the better image caption. The model structure and prediction process are shown in Figure 8.

The whole process consists of four main elements, including agent, environment, action and goal. In the image caption tasks, the policy network and the value network are the agents and also the main parts of the model. The input image $I$ and its description sentence $s_t = \{x_1 x_2 \ldots x_t\}$ represent the actual environment of the agents. The next predicted word $x_{t+1}$ is the next action, and the thesaurus of all the words in the caption is the space for the actions. Generating the image caption is the goal of this process.

The policy network adopts the encoder-decoder architecture mentioned above in this paper. We use $s_t$ to represent the current state, $e = \{Ix_1 x_2 \ldots x_t\}$ to represent the environment, and $a_t = x_{t+1}$ to represent the next action based on the environment. The visual feature $v_g$ of image $I$ is extracted by CNN, as shown in formula (30).

$$v_g = f_{cnn}(I) \tag{30}$$

Using $v_g$ as the input of the decoder NLSTM, the action $a_t$ at time $t$ is predicted according to the hidden layer state $h_t$ at time $t$ and the input word $x_{t-1}$ at time $t-1$. Because the decoder adopts a sequential processing mode, the prediction word $x_t$ will also be used as the input for time $t+1$, and the hidden layer state at the next time will also be updated as the input is updated. The formulas are shown as follows.

$$h_t = NLSTM\left(\psi\left(x_{t-1} v_g\right) h_{t-1}\right), \quad t \in N^* \tag{31}$$

$$p_\varepsilon\left(a_t s_t\right) = \phi\left(h_t\right) \tag{32}$$

Where $\psi$ and $\phi$ represent the input and output of the decoder, respectively. $p_\varepsilon\left(a_t s_t\right)$ represents the possibility of taking action $a_t$ in the case of determining state $s_t$.

In the value network, the value function $v_p$ under the policy $p$ is first defined, which represents the prediction of the total reward $r$ in the state $s_t$, expressed by formula (33).

$$v_p(s) = \mathrm{E}\left(rs_t a_{t\ldots T} \sim p\right) \tag{33}$$

In this paper, the output $v_\epsilon(s)$ of the value network is constructed to fit the value function. The value network is based on the deep neural network, and its structure is shown in Figure 9. It mainly consists of three parts: CNN module, RNN module and fully connected network module. The CNN module is used to extract the visual features of the image, and the Inception-v3 model is selected in this paper. RNN module adopts LSTM structure to extract semantic features of descriptions. The fully connected network module uses the linear regression method to obtain the reward score of the generated semantic descriptions.

In the value network, when the agents complete a goal, the total reward is used to motivate the actions taken. Here, the linear mapping method implemented by the fully connected module maps the image and the corresponding description into a semantic embedding space, to calculate the vector distance between them. The loss function $m_{loss}$ of this mapping can be expressed by the formula (34).

$$m_{loss} = \sum_{f_{cnn}} \sum_{s} \alpha \left[\max\left(0 h_{T-1}(s) \cdot f_m\left(f_{cnn}\right)\right) - h_T(s) \cdot f_m\left(f_{cnn}\right)\right] \tag{34}$$

Where $\alpha$ is the penalty coefficient with the range of (0,1), $f_{cnn}$ is the image feature extracted by the DenseNet, and $f_m$ is the mapping function.

For a given description sentence $s$, whose embedded characteristics depend on the final state $h_T$ of the hidden layer, and

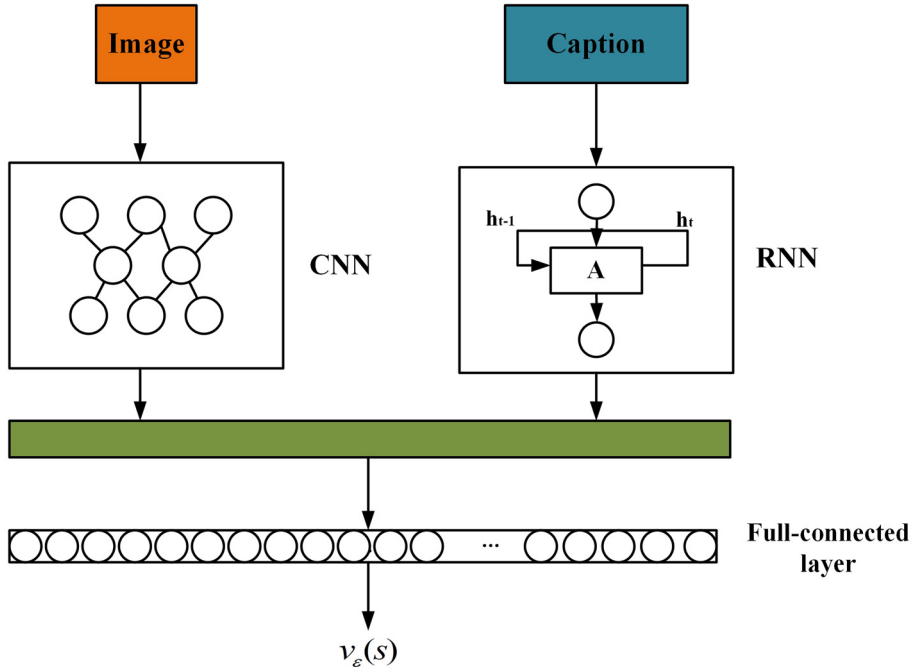



FIGURE 9
The structure of value network.

the total reward is defined as shown in formula (35).

$$r_T = \frac{h_{T-1}(s) \cdot f_m(f_{cnn})}{\| h_{T-1}(s) \cdot f_m(f_{cnn}) \|} \tag{35}$$

According to formula (35), the total loss $r_{loss}$ is calculated in formula (36).

$$r_{loss} = \beta (m_{loss} + r_T) \tag{36}$$

Where $\beta$ is the hyperparameter with the range of (0,1).

## 4. Experimental process and result analysis

We assess the effectiveness of the image caption model presented in this paper by means of a deliberate experimental process, including thorough comparative analysis of the experimental results. The experimental environment and datasets deployed in the experiment are introduced in detail. Additionally, the data preprocessing method, specific model training methodology, and optimization of model parameters are also comprehensively discussed. Finally, through comparative analysis, the performance and advantages of the proposed model are evaluated in depth for maximum objectivity and credibility.

In the tasks of image caption, the most popular datasets adopted by most researchers include MS COCO (Lin et al., 2014) and Flickr 30 k (Young et al., 2014). The Flickr dataset is primarily a description of human activity scenarios. We use 29,000 of the Flickr data as a training set, 1,000 as a validation set, and the remaining 1,000 as a test set. In addition, 40,775 images and 30,775 data of the corresponding image descriptions from the MS COCO dataset are added to the training set to increase the number of training samples. The deep learning framework used is TensorFlow.

First of all, it is necessary to preprocess the data in the datasets, including the images and the descriptions. The image size is uniformly adjusted to 256*256, then trimmed to 224*224 to fit the model input. And the image is normalized to scale each pixel with the range of (0,1). Firstly, the description sentences need to segment, convert all letters to lower case, and remove spaces and punctuation. Then, the number of occurrences of all words in the datasets is counted, and words that appear less than 5 times are tagged *UNK* which have little effect on predicting outcomes. Finally, it is stipulated that the length of the sentences is not more than 15 words, each sentence only intercepts the characteristic values corresponding to the first 15 words. For sentences with less than 15 words, we supplement the number of characteristic values to 15, and the supplementary characteristic values are 0. At the same time, the tag *start* and *end, respectively,* placed at the beginning and end of the description sentences, to mark the beginning and end of the sentences.

In this paper, we adopt BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015), which are commonly used evaluation indicators. In the model testing phase, this paper uses the method of beam search to choose a better generated sentence. The five sentences with the highest probability value are output at each decoding moment, that is, the value of beam size is set to 5.

Given that dropout operation is used during model training, the impact of different dropout ratios on model performance can vary. To determine the optimal dropout ratio for the model, this

TABLE 1 Comparison of model performance on MS COCO dataset.

| Models | BLEU-1 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| Google NIC | 0.666 | 0.277 | 0.237 | - | 0.855 |
| Soft attention | 0.707 | 0.243 | 0.239 | - | - |
| g-LSTM | 0.670 | 0.264 | 0.227 | - | 0.813 |
| RIC | 0.734 | 0.299 | 0.254 | - | - |
| RHN | 0.723 | 0.306 | 0.252 | - | 0.989 |
| LSTM-A5 | 0.730 | 0.325 | 0.251 | 0.538 | 0.986 |
| This paper (basic model with no DRL and attention mechanism) | 0.716 | 0.289 | 0.244 | 0.456 | 0.893 |
| This paper (final model with DRL but no attention mechanism) | 0.746 | 0.339 | 0.284 | 0.583 | 0.991 |
| This paper (final model with DRL and attention mechanism) | 0.752 | 0.344 | 0.289 | 0.588 | 1.066 |

TABLE 2 Comparison of model performance on the Flickr 30 k dataset.

| Models | BLEU-1 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| Google NIC | 0.663 | 0.183 | - | - | - |
| Soft attention | 0.669 | 0.199 | 0.185 | - | - |
| g-LSTM | 0.646 | 0.206 | 0.180 | - | - |
| RIC | 0.745 | 0.244 | 0.202 | - | - |
| RHN | 0.738 | 0.307 | 0.216 | - | - |
| This paper (basic model with no DRL and attention mechanism) | 0.718 | 0.242 | 0.191 | 0.352 | 0.886 |
| This paper (final model with DRL but no attention mechanism) | 0.734 | 0.320 | 0.215 | 0.492 | 0.885 |
| This paper (final model with DRL and attention mechanism) | 0.738 | 0.335 | 0.222 | 0.504 | 0.921 |

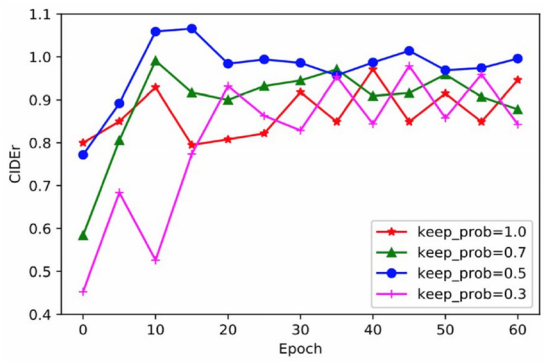



FIGURE 10
A comparison of the scores of different dropout ratio on CIDEr.

paper compares model scores across different dropout ratios using the CIDEr evaluation indicator and presents a comparative graph in Figure 10. Analysis of the results indicates that when dropout operation is not performed, the score of the model fluctuates greatly, which indicates that the model is too complex and overfitting has occurred. Similarly, when the dropout ratio is 0.3, the fluctuation remains high and the model convergence score is low suggestive of underfitting arising from insufficient involvement of neurons in training. In contrast, when the dropout ratio is set at either 0.5 or 0.7, the curve remains relatively stable with a better CIDEr score when the dropout ratio is 0.5. Thus, the appropriate dropout ratio for the model is determined to be 0.5.

In this study, we conducted a comparative analysis of our model's performance against other mainstream models, namely Google NIC, Soft attention, g-LSTM, RIC, RHN, and LSTM-A5. We evaluated the models using different metrics on MS COCO and Flickr 30 k. The comparison results are presented in Tables 1, 2.

As shown in Table 1, on the MS COCO dataset, the basic model proposed in this paper has improved the scores of BLEU-1 and BLEU-4, which measure sentence coherence and accuracy, by nearly 0.05 and 0.03, respectively, compared with the g-LSTM model, due to the use of the guided decoding network. At the same time, using DenseNet and MIL to process image information also improved the score of CIDEr evaluation index reflecting semantic richness by nearly 0.04 compared with Google NIC which only used the Inception-v3 structure as the image information extraction model. However, compared with more advanced models such as RIC and LSTM-A5, the proposed basic model still has a certain gap in the scores of various evaluation indexes. The reason is that the attention mechanism is not introduced, so the details are not enough. And the decoder only uses a single layer structure, so the decoding process is not sufficient.

As can be seen from the results in Table 1, on the MS COCO dataset, the performance of the final model in this paper is superior to the comparison models on various evaluation indicators even when without attention mechanism. Therefore, the use of DRL can significantly improve the performance of the image caption model, and when the attention mechanism is added, the model certainly performs better. Specifically, the BLEU scores of the proposed model are improved by 0.018 and 0.019, respectively, compared with the


FIGURE 11
The effect diagram of the attention mechanism.

best results in the comparison models, which indicates that the output sentences of the proposed model have better coherence and accuracy. In terms of the METEOR scores, the proposed model also has an improvement of more than 0.03 compared with other models. In addition, without the attention mechanism, the model in this paper is also improved by more than 0.05 compared with the g-LSTM model, so the end-to-end model structure in this paper has greater advantages than the static adjustment of g-LSTM. Compared with the Soft attention model, which also uses the attention mechanism, the performance is improved by 0.05 due to the double-layer mechanism guiding the decoding and the optimization of DRL. In terms of CIDEr scores, which measures semantic richness and description consistency, there is also an improvement of 0.077 compared with the best results in the comparison models, which shows the excellent performance of the model designed in this paper.

As shown in Table 2, because the Flickr 30 k dataset contains much less data than the MS COCO dataset, the evaluation index scores of the proposed basic model and final model are basically decreased compared with those in Table 1. However, the basic model presented in this paper has higher evaluation index scores than the Google NIC, Soft attention, and g-LSTM models. And the scores of the final model are better than the comparison models in most evaluation indicators, however, the scores of some indicators are slightly lower than those of some models, which may be caused by the poor generalization ability of the model due to too small amount of data.

After the attention mechanism is used to improve the proposed model, in order to verify the actual effect, the extracted image features and the hidden layer state of the first layer decoder are processed by the attention module, then the words corresponding to different regions in the image are determined according to the corresponding weights, and the effect diagram is shown in Figure 11. Figure 11 shows the corresponding focus of each word in the sentence in the image. The white highlights in each image from left to right correspond to each word from left to right in the sentence below, and the whiter part of the highlights indicates the greater attention weight assigned. As can be seen from the images, the attribute word "green" about color focuses on the position of the bird's body, and the target subject "bird" focuses on the head of the bird, because the head is the area that can best reflect the characteristics of the bird. The phrase "standing on" focuses on the bird's feet, which is characteristic of the action. The word "grass" focuses on the green area where the bird is standing. Through the above analysis, it can be seen that the double-layer decoding structure model with the introduction of the attention mechanism is very accurate in extracting and matching key information and local information in the image, and it is also helpful in improving the performance of the image caption model.

## 5. Conclusion

Aiming at the problems of existing image caption models, this paper proposes an image caption model based on deep learning. Firstly, based on the NIC model, the encoder and decoder are optimized through DenseNet and NLSTM networks. Meanwhile, this paper also introduces a guided decoding network to realize the dynamic adjustment of encoded information in the decoding process and avoid the loss of image information. The experimental results show that compared with several common models, the performance of the basic model designed in this paper is improved. Then, on the basis of the proposed image caption model, we introduce the attention mechanism to construct a double-layer decoding structure and improve the decoding depth to obtain the details of the image. The powerful perception and decision abilities of DRL are adopted to optimize the model, which solve the problem of discrepancies between training objectives and evaluation indicators, and improve the expressive ability of the image caption model. Through the comparison and analysis of the experimental results with several common models, our image caption model further improves the scores of each evaluation index, and the output description of the image is more accurate and semantic rich. In future work, we will design the image caption model based on expression ways in different scenes and language habits of different people, so that the sentences output by the model will be closer to the expression ways of humans in real scenes. Meanwhile, we will continue to expand the datasets to include richer content, and further design a better model to enable zero-sample learning through textual inference.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

TB: Methodology, Project administration, Writing – original draft, Writing – review & editing. SZ: Software, Supervision, Validation, Writing – review & editing. YP: Data curation, Supervision, Writing – review & editing. JL: Validation, Visualization, Writing – original draft. HW: Data curation, Writing – original draft, Writing – review & editing. YD: Investigation, Writing – original draft.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The project was funded by the National Natural Science Foundation of China (62171073, 61971079, and U21A20447), Department of Science and Technology of Sichuan Province (2020YFQ0025 and 2020YJ0151), Project of Central Nervous System Drug Key Laboratory of Sichuan Province (210022-01SZ, 200020-01SZ, 200028-01SZ, and 200027-01SZ), Natural Science Foundation of Chongqing (CSTB2022NSCQ-MSX1523, cstc2019jcyj-msxmX0275, cstc2019jcyj-msxmX0737, cstc2020jcyj-cxttX0002, cstc2019jcyjmsxmX0666, cstc2021jscx-gksbx0051, and cstc2021jcyj-bsh0221), China Postdoctoral Science Foundation (2022MD713702), Chongqing Technical Innovation and Application Development Special Project (CSTB2022TIAD-KPX0062), Chongqing Scientific Institution Incentive Performance Guiding Special Projects (cstc2022jxj120036), Science and Technology Research Project of Chongqing Education Commission (KJZD-k202000604, KJQN202100602, KJQN202100602, and KJQN202000604), SAMR Science and Technology Program (2022MK105), Key Research Project of Southwest Medical University (2021ZKZD019), Special support for Chongqing Postdoctoral Research Project (2021XM3010 and 2021XM2051), Project funded by China Postdoctoral Science Foundation (2022MD713702, 2021MD703941, and 2021M693931).

## Acknowledgments

The authors thank the School of Optoelectronic Engineering of Chongqing University of Posts and Telecommunications for their assistance in the research.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Amritkar, C. and Jabade, V. (2018). Image caption generation using deep learning technique. In: *2018 fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1–4). IEEE.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: semantic propositional image caption evaluation. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14* (pp. 382–398). Springer International Publishing.

Bai, S. and An, S. (2018). A survey on automatic image caption generation. *Neurocomputing* 311, 291–304. doi: 10.1016/j.neucom.2018.05.080

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., et al. (2016). Automatic description generation from images: a survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.* 55, 409–442. doi: 10.1613/jair.4900

Bhalekar, M. and Bedekar, M. (2022). D-CNN: a new model for generating image captions with text extraction using deep learning for visually challenged individuals. *Engineer Technol Appl Sci Res* 12, 8366–8373. doi: 10.48084/etasr.4772

Bjorck, N., Gomes, C. P., Selman, B., and Weinberger, K. Q. (2018). Understanding batch normalization. *Adv. Neural Inf. Proces. Syst.* 31:2375. doi: 10.48550/arXiv.1806.02375

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Computer Science.

Deng, Z., Jiang, Z., Lan, R., Huang, W., and Luo, X. (2020). Image captioning using DenseNet network and adaptive attention. *Signal Process. Image Commun.* 85:115836. doi: 10.1016/j.image.2020.115836

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* 89, 31–71. doi: 10.1016/S0004-3702(96)00034-3

Elliott, D. and de Vries, A. (2015). Describing images using inferred visual dependency representations. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 42–52).

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., et al. (2010). Every picture tells a story: generating sentences from images. In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11* (pp. 15–29). Springer Berlin Heidelberg.

Fei, Z. (2021). Memory-augmented image captioning. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1317–1324).

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Comput Surveys* 51, 1–36. doi: 10.1145/3295748

Huang, L., Wang, W., Chen, J., and Wei, X. Y. (2019). Attention on attention for image captioning. In: *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4634–4643).

Jing, Y., Zhiwei, X., and Guanglai, G. (2020). Context-driven image caption with global semantic relations of the named entities. *IEEE Access* 8, 143584–143594. doi: 10.1109/ACCESS.2020.3013321

Kang, W. and Hu, W. (2022). A survey of image caption tasks. In: *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)* (pp. 71–74). IEEE.

Kinghorn, P., Zhang, L., and Shao, L. (2018). A region-based image caption generator with refined descriptions. *Neurocomputing* 272, 416–424. doi: 10.1016/j.neucom.2017.07.014

Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 317–325). IEEE, 214.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, X. and Jin, Q. (2016). Improving image captioning by concept-based sentence reranking. In: *Advances in Multimedia Information Processing-PCM 2016: 17th Pacific-Rim Conference on Multimedia, Xi´ an, China, September 15–16, 2016, Proceedings, Part II* (pp. 231–240). Springer International Publishing.

Lin, C. Y. (2004). "Rouge: a package for automatic evaluation of summaries" in *Text summarization branches out* (Barcelona, Spain: Association for Computational Linguistics), 74–81.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13* (pp. 740–755). Springer International Publishing.

Liu, M., Hu, H., Li, L., Yu, Y., and Guan, W. (2020). Chinese image caption generation via visual attention and topic modeling. *IEEE Trans Cybernet* 52, 1247–1257. doi: 10.1109/TCYB.2020.2997034

Liu, M., Li, L., Hu, H., Guan, W., and Tian, J. (2020). Image caption generation with dual attention mechanism. *Inf. Process. Manag.* 57:102178. doi: 10.1016/j.ipm.2019.102178

Liu, X., Xu, Q., and Wang, N. (2019). A survey on deep neural network-based image captioning. *Vis. Comput.* 35, 445–470. doi: 10.1007/s00371-018-1566-y

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv [Preprint].

Nivedita, M., Chandrashekar, P., Mahapatra, S., Phamila, Y. A. V., and Selvaperumal, S. K. (2021). Image captioning for video surveillance system using neural networks. *Int J Image Graph* 21:2150044. doi: 10.1142/S0219467821500443

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Parikh, H., Sawant, H., Parmar, B., Shah, R., Chapaneri, S., and Jayaswal, D. (2020). Encoder-decoder architecture for image caption generation. In: *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)* (pp. 174–179). IEEE.

Shakarami, A. and Tarrah, H. (2020). An efficient image descriptor for image classification and CBIR. *Optik* 214:164833. doi: 10.1016/j.ijleo.2020.164833

Shaked, A. and Wolf, L. (2017). Improved stereo matching with constant highway networks and reflective confidence learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4641–4650). IEEE.

Singh, A., Singh, T. D., and Bandyopadhyay, S. (2021). An encoder-decoder based framework for hindi image caption generation. *Multimed. Tools Appl.* 80, 35721–35740. doi: 10.1007/s11042-021-11106-5

Srivastava, G. and Srivastava, R. (2018). A survey on automatic image captioning. In: *Mathematics and Computing: 4th International Conference, ICMC 2018, Varanasi, India, January 9–11, 2018* (pp. 74–83). Springer Singapore.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: consensus-based image description evaluation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566–4575).

Verma, Y., Gupta, A., Mannem, P., and Jawahar, C. V. (2013). Generating image descriptions using semantic similarities in the output space. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 288–293).

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: a neural image caption generator. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164). IEEE.

Yang, X., Tang, K., Zhang, H., and Cai, J. (2019). Auto-encoding scene graphs for image captioning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10685–10694).

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist* 2, 67–78. doi: 10.1162/tacla00166

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 19 October 2023
DOI 10.3389/fnins.2023.1291674



OPEN ACCESS

EDITED BY
Teng Li,
Anhui University, China

REVIEWED BY
Hai Wang,
Murdoch University, Australia
Ming Yu,
Hefei University of Technology, China

*CORRESPONDENCE
Dong Zhang
✉ Zhangdong17@mails.jlu.edu.cn

RECEIVED 09 September 2023
ACCEPTED 06 October 2023
PUBLISHED 19 October 2023

CITATION
Liu D, Zhang D, Wang L and Wang J (2023) Semantic segmentation of autonomous driving scenes based on multi-scale adaptive attention mechanism.
*Front. Neurosci.* 17:1291674.
doi: 10.3389/fnins.2023.1291674



# Semantic segmentation of autonomous driving scenes based on multi-scale adaptive attention mechanism

Danping Liu[1], Dong Zhang[2]*, Lei Wang[1] and Jun Wang[1]

[1]School of Advanced Manufacturing Engineering, Hefei University, Hefei, China, [2]State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun, China

**Introduction:** Semantic segmentation is a crucial visual representation learning task for autonomous driving systems, as it enables the perception of surrounding objects and road conditions to ensure safe and efficient navigation.

**Methods:** In this paper, we present a novel semantic segmentation approach for autonomous driving scenes using a Multi-Scale Adaptive Mechanism (MSAAM). The proposed method addresses the challenges associated with complex driving environments, including large-scale variations, occlusions, and diverse object appearances. Our MSAAM integrates multiple scale features and adaptively selects the most relevant features for precise segmentation. We introduce a novel attention module that incorporates spatial, channel-wise and scale-wise attention mechanisms to effectively enhance the discriminative power of features.

**Results:** The experimental results of the model on key objectives in the Cityscapes dataset are: ClassAvg:81.13, mIoU:71.46. The experimental results on comprehensive evaluation metrics are: AUROC:98.79, AP:68.46, FPR95:5.72. The experimental results in terms of computational cost are: GFLOPs:2117.01, Infer. Time (ms):61.06. All experimental results data are superior to the comparative method model.

**Discussion:** The proposed method achieves superior performance compared to state-of-the-art techniques on several benchmark datasets demonstrating its efficacy in addressing the challenges of autonomous driving scene understanding.

KEYWORDS

**semantic segmentation, attention mechanism, autonomous driving, convolutional neural networks, deep learning**

## 1. Introduction

Over the past several decades, autonomous driving technology has made remarkable strides. The current bottleneck impeding its mass adoption is safety, as it directly pertains to human life and well-being. Autonomous vehicles are increasingly becoming integral across a multitude of scenarios—from daily living and work commutes to travel and leisure—where safety emerges as a critical factor governing their application. These self-driving platforms are fundamentally built upon sophisticated visual perception systems (Hubmann et al., 2018; Jin et al., 2021; Hu et al., 2023), in which semantic segmentation plays an essential role for pixel-level classification of camera images. While recent research has primarily focused on enhancing the accuracy of semantic segmentation, high-precision pixel-level classification of objects often relies on strong supervised learning methods trained on large, fully-annotated datasets. These models are consequently limited to classifying conventional objects—that is, categories predefined in the

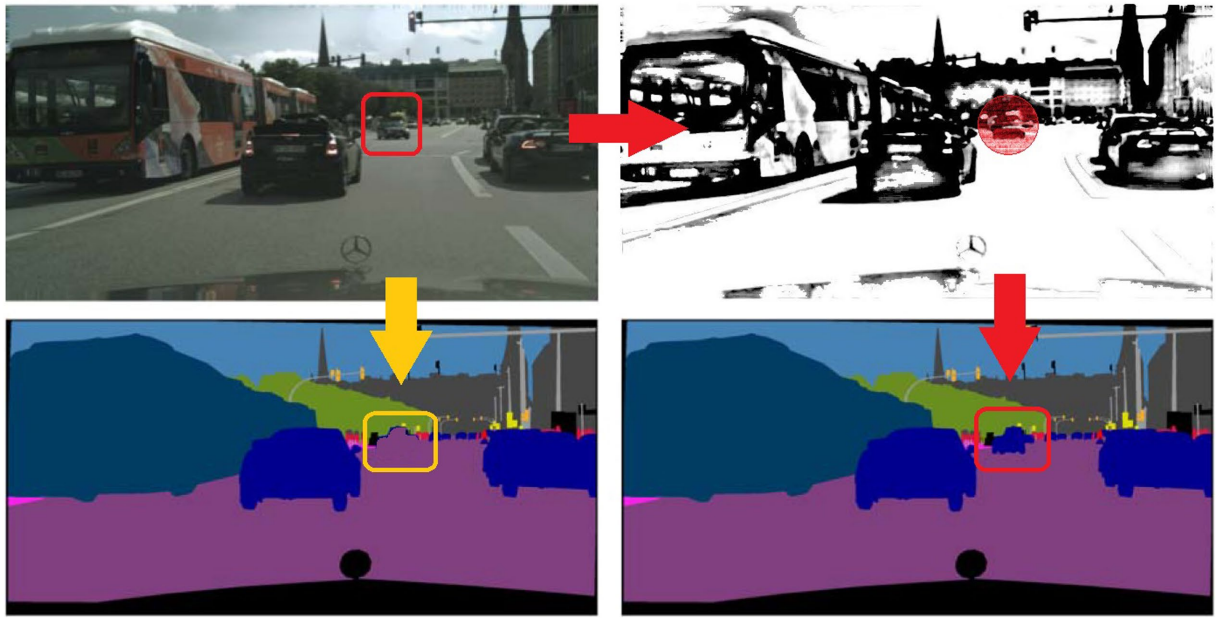


FIGURE 1
Examples of hazardous scenarios.

dataset—operating under the overly idealistic assumption that all objects in real-world driving environments remain constant. Unfortunately, the real world is ever-changing, and unpredictable situations can arise at any moment. For instance, an object with altered characteristics, such as an small obstacle in a driving scene in Figure 1, may not be properly identified by the model, which may overconfidently misclassify it into another category. Such scenarios pose serious safety risks and significantly hamper the practical deployment of deep learning algorithms in autonomous driving. Moreover, collecting a dataset that encompasses every conceivable variation is impractical. Driving environments that present significant challenges due to their dynamic nature fall under the category of hazardous scenarios, where all dynamic elements could be termed 'anomalous obstacles.' Therefore, it is crucial for a perception network to be trained to adapt to variations and anomalies in these risky settings.

Several studies have addressed the challenge of detecting variations and anomalous targets in hazardous driving scenarios (Lis et al., 2019; Doshi and Yilmaz, 2020; Xia et al., 2020; Blum et al., 2021; Vojir et al., 2021). One line of approaches employs uncertainty estimation techniques, intuitively based on the low prediction probabilities associated with anomalous targets. These methods design specific functions to compute uncertainty probabilities and subsequently generate anomaly scores. However, these techniques often yield noisy and imprecise detection results due to the model's overconfidence in identifying anomalous targets. Another primary approach involves augmenting the training pipeline with additional tasks specifically for detecting anomalous obstacles. Some methods employ external out-of-distribution (OoD) datasets as training samples for this category, while others utilize feature reconstruction techniques to either manually design or learn the features of unknown classes to distinguish anomalies. Generative models are then used to resynthesize the input images. Although these methods have proven effective, they are either computationally expensive in terms of inference time or labor-intensive in their implementation. Moreover, the retraining process may compromise the original network's performance in semantic segmentation. Therefore, there is a pressing need for more balanced solutions for perceiving and segmenting variations and anomalous objects in hazardous scenarios. The ideal approach should enhance the performance of uncertainty methods without significantly increasing computational overhead or training complexity, all while preserving the accuracy of semantic segmentation.

Human attention mechanisms serve as the foundation for various cognitive processes, allowing us to selectively focus on specific stimuli from an array of available inputs for deeper processing. While psychology offers critical methodologies for studying these attention mechanisms, neuroscience also stands as a primary field in which they are explored (Desimone and Duncan, 1995). Human attention can be conceptualized as a filtering process, determining which pieces of information merit further consideration and which should be disregarded (Treisman and Gelade, 1980). Psychological research delves into the behavioral aspects of attention, such as its selectivity, concentration, and shifting focus. Extensive inquiries into the operational aspects of attention have been made through experiments, observations, and surveys, covering theories of selective attention, filtering models, theories of attention allocation, and the attentional blink, among others (Broadbent, 1958; Kahneman, 1973; Raymond et al., 1992). Neuroscience examines the neural underpinnings of attention, identifying specific brain regions involved in the attention process. Utilizing functional Magnetic Resonance Imaging (fMRI) and electrophysiological techniques, scientists have identified the prefrontal and parietal cortices as key areas for regulating attention (Corbetta and Shulman, 2002), with additional research focusing on neurotransmitter systems and neural oscillations (Arnsten and Li, 2005; Jensen and Mazaheri, 2010). Given that attention mechanisms

are integral to human cognition and crucial for learning, memory, decision-making, and other cognitive functions, they have inspired research and applications in computer science and artificial intelligence. In fields ranging from resource allocation to state-of-the-art deep learning models—particularly in scenarios dealing with big data and large volumes of information—attention mechanisms have found robust applications (Mnih et al., 2014; Bahdanau et al., 2015; Ma et al., 2019). Drawing inspiration from psychological and neuroscience research into attention mechanisms, significant progress has also been made in developing attention algorithms within the domain of artificial intelligence (Vaswani et al., 2017; Nobre and van Ede, 2018; Cichy and Kaiser, 2019).

Inspired by human attention mechanisms, humans demonstrate remarkable environmental perception skills, effortlessly identifying invariant and ordinary elements amidst variations and anomalies such as large-scale changes in object dimensions, occlusions, and diverse object appearances. This keen attention to the constant and ordinary amidst flux and irregularities equips humans with robust capabilities for environmental perception. How might this attention paradigm be mapped onto the domain of semantic segmentation in autonomous driving scenes? First, by analyzing and constructing the feature attributes associated with variations and anomalies in hazardous scenarios; and second, by aligning these identified feature attributes with the most fitting attention mechanisms.

One of the most pervasive attributes of variation and anomaly in autonomous driving scenarios is the substantial and high-frequency scale change of environmental objects. Objects may vary considerably in size and shape, and can be particularly challenging to recognize at differing image resolutions. For instance, a distant car may appear small in the image, whereas a nearby car would be considerably larger, leading to anomalies such as two objects at different distances with similar scales and contours being misperceived as the same category. To address this issue, we employ a scale attention mechanism that operates over multiple image scales within the network architecture. These results are then integrated to enhance the accuracy and robustness of semantic segmentation, thereby providing more reliable and granular information for autonomous driving scenarios.

Due to the spatially diverse distribution of objects at different scales—for instance, distant vehicles may occupy a diminutive spatial footprint, while nearby pedestrians may occupy a more substantial one—a scale attention mechanism necessitates integration with spatial attention. Without such a fusion, the model may struggle to ascertain the relative spatial positions and importance of differently sized structures or objects. For example, a distant small vehicle might be semantically more critical than a proximal large tree, but in the absence of spatial context, the model might disproportionately focus on the tree. Additionally, spatial attention allows the model to home in on partially obscured yet crucial areas, such as the legs or head of an obstructed pedestrian. Given that different features or attributes may reside in different channels—for instance, some channels may prioritize edge information, while others may focus on texture or color information—structures or objects of different scales may exhibit diverse feature expressions across these channels. For a scale attention mechanism to properly weight these features, channel attention integration becomes necessary, failing which could lead to information loss or confusion at certain scales. Moreover, objects in driving environments display various characteristics owing to changes in lighting, weather, and object types, among other factors. For instance, the same object category—such as a car—can display significant variations in color, model, and design. Since different appearance features may be distributed across different channels, channel attention allows the model to focus on key channels instrumental in identifying specific appearances.

This paper introduces a Multi-Scale Adaptive Attention Mechanism (MSAAM) for Semantic Segmentation in Autonomous Driving Scenes. Initially, a scale attention module is incorporated at the end of the Convolutional Neural Network (CNN) encoder. Subsequently, spatial and channel attention models are synergistically integrated to enhance the performance of the multi-scale attention mechanism. Building on this, a composite weighting model encompassing scale, spatial, and channel attention is established. This model is trained through a compact neural network to meet the requirements for adaptive weighting and employs the Softmax function to ensure the sum of the weights equals one, thereby preventing disproportionately large weights. Finally, an attention-specific loss function is proposed to further amplify the distance between the attention values focused on specific pixels and those on the remaining pixels. These methodologies allow us to train a semantic segmentation network based on MSAAM, effectively addressing the perceptual challenges posed by hazardous scenarios in autonomous driving, such as large-scale variations, occlusions, and diverse object appearances, among others.

The main contributions of our work are as follows:

This paper introduces the Multi-Scale Adaptive Attention Mechanism (MSAAM) specifically designed for semantic segmentation in driving scenarios. It is an attention mechanism that seamlessly integrates three channels—scale, spatial, and channel—and adaptively allocates their weights.

The multi-scale adaptive attention model that fuses multiple channels is adept at handling various attributes encountered in scenes, such as large-scale variations, occlusions, and diverse object appearances. Moreover, this attention model is highly modular and can be flexibly adapted to integrate with various Convolutional Neural Network (CNN) architectures, essentially offering a plug-and-play solution.

Our approach improves the performance of pixel-level semantic segmentation without substantially increasing the number of parameters or complicating the training process.

## 2. Related work

In the realm of hazardous scenario analysis, research work predominantly focuses on two main approaches for detecting variations and abnormal feature attributes: one that leverages uncertainty estimation and another that incorporates additional training tasks. This article also explores studies relevant to multi-scale attention mechanisms, which is the focus of our work. In this section, we provide an overview of research conducted in these three key areas.

### 2.1. Anomaly segmentation via uncertainty estimation

Methods based on uncertainty estimation serve as the most straightforward approach in abnormality detection, where

uncertainty scores are utilized to identify obstacles on the road. Early studies employed Bayesian neural networks and Monte Carlo dropout to assess uncertainty. However, these techniques are often slow in inference and prone to boundary misclassifications (Kendall et al., 2015; Kendall and Gal, 2017; He et al., 2020). Alternative approaches focus on utilizing maximum softmax probabilities or maximum logits to improve uncertainty assessment, but these too suffer from the issue of boundary misclassification (Hendrycks and Gimpel, 2016; Jung et al., 2021; Hendrycks et al., 2022). Generally speaking, without additional fine-tuning using outlier data, methods based on uncertainty tend to perform poorly in terms of overconfidence and false positives at boundaries.

## 2.2. Anomaly segmentation via introducing additional training tasks

Another approach to abnormal segmentation involves incorporating extra training tasks. These tasks primarily fall under three categories: feature reconstruction, leveraging auxiliary datasets, and image re-synthesis. Feature reconstruction methods operate by analyzing the normality and deviations in the input features but are dependent on precise pixel-level segmentation (Creusot and Munawar, 2015; Di Biase et al., 2021). Methods based on auxiliary datasets employ external data to enhance detection accuracy but struggle to capture all potential anomalies, compromising the model's generalizability (Bevandic´ et al., 2019; Chan et al., 2021). Image re-synthesis techniques, such as those employing autoencoders and Generative Adversarial Networks (GANs), create more diverse abnormal samples but at the cost of computational complexity and extended inference time (Ohgushi et al., 2020; Tian et al., 2021). While these additional training tasks contribute to improving abnormality detection, they may also adversely impact the primary task, i.e., semantic segmentation performance.

## 2.3. Multi-scale attention mechanisms for image segmentation or fine-grained image classification

Effective learning of multi-scale attention regions is pivotal in the domains of image segmentation and fine-grained image classification (Ge et al., 2019; Zheng et al., 2019). Earlier research largely relied on manually annotated object bounding boxes, a process that is both time-consuming and impractical. Xiao et al. were the first to introduce a multi-scale attention model that does not depend on manual annotation, incorporating both object-level and part-level attention (Xiao et al., 2015). More recent studies have evolved to be more intricate, involving adaptive region localization, weakly-supervised learning, and Feature Pyramid Networks (Fu et al., 2017; Rao et al., 2019; Ding et al., 2021). These advancements contribute to more precise localization and classification of target areas, thereby enhancing the performance of pixel-level segmentation or fine-grained classification (Li et al., 2016a,b; Nian et al., 2016; Zhang et al., 2019, 2020; Jiang et al., 2020; Liu et al., 2021).

# 3. Methodology

This section elucidates the Multi-Scale Adaptive Attention Mechanism (MSAAM) approach that we employ for semantic segmentation in autonomous driving scenes. Initially, in Subsection 3.1, we articulate the motivations underlying our methodology. Following this, Subsection 3.2 presents an overview of the comprehensive architecture of MSAAM. Subsection 3.3 details the multi-scale attention module, while Subsection 3.4 describes a weight-adaptive fusion attention system.

## 3.1. Motivation

Human attention mechanisms assist us in selecting and focusing on a particular stimulus among various inputs for in-depth processing. This mechanism is not only a focal point in psychological research but also a principal area of study in neuroscience. Psychology investigates the behavioral characteristics of attention, utilizing a range of experiments and questionnaires to understand how attention is selected and allocated. Neuroscience, on the other hand, delves into the brain regions responsible for attention, employing technologies such as fMRI and electrophysiology. Attention plays a crucial role in cognitive functions like learning, memory, and decision-making. Inspired by these insights, the fields of computer science and artificial intelligence have also begun to explore and implement attention mechanisms, especially in contexts that involve large-scale data and high information volume. Advances in attention mechanisms within artificial intelligence have been made by drawing upon foundational research in psychology and neuroscience.

Inspired by human attention mechanisms, we can identify stability and regularity amidst environmental variations and anomalies, thereby perceiving the environment more effectively. How can such an attention paradigm be applied to semantic segmentation in autonomous driving scenarios? First, it involves analyzing and identifying the characteristics of variations and anomalies in hazardous scenes; second, it calls for choosing suitable attention mechanisms tailored for these specific traits.

In autonomous driving scenes, rapid and substantial changes in object scale pose a significant challenge. For instance, cars at varying distances appear drastically different in size within the same image, potentially leading to erroneous identification. To tackle this issue, we employ scale attention mechanisms to process multiple image scales and integrate the results. This enhances the accuracy and robustness of semantic segmentation, making autonomous driving more reliable.

In autonomous driving contexts, both the scale and spatial positioning of objects are of paramount importance. For example, a distant car may hold more significance than a nearby tree, yet the model may overemphasize the tree due to a lack of spatial context. Therefore, scale attention must be combined with spatial attention to comprehend the relative positioning and importance of objects in space. Spatial attention also helps the model focus on partially occluded yet crucial areas. Additionally, object features of different scales and appearances might reside in different channels, such as edge or color information. To avoid losing or confusing these details, the

scale attention model also incorporates channel attention. In this way, the model can more accurately identify a variety of appearances under different lighting conditions, weather, and object types.

## 3.2. Overall architecture

Semantic segmentation models are generally formulated as encoder-decoder architectures. An input image is initially transformed into high-dimensional features via the encoder. Subsequently, with these intermediate features as input, the MSAAM first infers a two-dimensional attention map. Importantly, attention should not be unbounded. A constant-sum constraint on attention values forces pixels within the attention map to compete against each other for maximal gain, thereby circumventing the pitfall of the model setting all attention values unfavorably high. We then select multi-layer, multi-scale features generated by the encoder and fuse them with the attention map. These fused features are fed into the decoder network to produce the predictive output. To widen the gap in attention values between focus pixels and other pixels, we introduce a penalty term in the loss function, termed as MSAAM Loss. Finally, the network's predictive output is combined with MSAAM's attention map to generate the ultimate integrated prediction.

Within the architecture, the MSAAM module situated between the encoder and the decoder serves as the linchpin for the attention mechanism. Initially, a Pyramid Attention Module is integrated at the terminal phase of the encoder. This module employs Pyramid Pooling to capture information across different scales, thereby establishing a multi-scale attention mechanism. Subsequently, we utilize the Convolutional Block Attention Module (CBAM) to concurrently address both spatial and channel attention. CBAM enriches contextual information by employing Global Average Pooling and Global Max Pooling techniques. To precisely calculate the weights across the three dimensions—scale, space, and channel—we have engineered a miniature neural network. This network comprises several fully connected layers and a Softmax layer, designed to learn the aggregate attention weights across different dimensions. As a specific implementation detail, Gated Recurrent Units (GRU) are employed to update the weights for each dimension, thus constructing a weight-adaptive model. The basic architecture of attention is shown in Figure 2.

## 3.3. Multi-scale attention module

Addressing the large-scale variations of objects poses a significant challenge for semantic segmentation in autonomous driving scenarios. Integrating a multi-scale attention mechanism into the segmentation process ameliorates these challenges by enabling the model to focus on regions of varying sizes.

The Pyramid Pooling Attention module (PSA) is specifically designed to capture contextual information across different dimensions and spatial resolutions. Traditional attention mechanisms often operate at a single scale, which could limit their ability to understand either broader or more nuanced details. In contrast, pyramid models, by creating representations at various granularities, can effectively tackle the multi-scale challenges inherent in computer vision. These representations offer a more comprehensive understanding of the scene, which is crucial for enhancing segmentation performance in diverse and dynamically changing environments, such as those encountered in autonomous driving.

The scale-wise attention module $\mathrm{f}^{\mathrm{sc}}$ in our framework is a sophisticated operation that effectively combines the input feature map $F_{in}$ with an attention map produced by the PSA module. Mathematically, it is represented as:

$$\mathrm{f}^{\mathrm{sc}}\left(F_{in}\right) = F_{in} + F_{in} \odot PSA\left(F_{in}\right) \tag{1}$$

in this context, symbolizes the scale-wise attention module, $F_{in}$ is the input feature map, $\odot$ stands for element-wise multiplication, and PSA denotes the Pyramid Pooling Attention module. The essence of this formula is that given an intermediate feature map, our module produces an attention map through the Pyramid Pooling Attention module and then multiplies this attention map with the input feature map, achieving adaptive feature refinement.

The definition of the Pyramid Pooling Attention module PSA is as follows:

$$\mathrm{PSA}\left(F_{in}\right) = softmax\left(\sum_{i=1}^{N} \omega_i P_i * F_{in}\right) \tag{2}$$

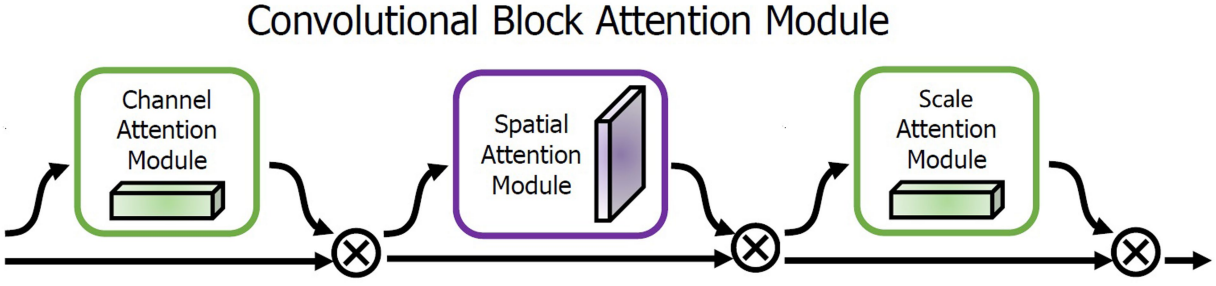



FIGURE 2
Architecture of the MSAAM attention mechanism.

in this equation, N represents the number of layers in the pyramid, $P_i$ refers to the pooling operation at the $i$-th layer, $w_i$ is the weight for that layer, and $*$ denotes the convolution operation. The resulting attention map amalgamates information from different scales by using a weighted combination of pyramid layers.

### 3.4. Weight-adaptive fusion attention system

Following the scale attention layer, we integrate both spatial and channel attention layers, formalized as follows:

$$\mathrm{f}^{\mathrm{sp}}(F)=\sigma\left(Conv_{7\times7}\left(\frac{1}{C(F)}\sum_{\forall j}f\left(F_i,F_j\right)g\left(F_j\right)\right)\right) \quad (3)$$

where, $\mathrm{f}^{\mathrm{sp}}$ represents the spatial attention module, A denotes the sigmoid function, $\mathrm{Conv}_{7\times7}$ stands for a convolutional layer with a kernel size of $7\times7$, $F_i$ and $F_j$ represent the input features from any two positions, $f$ is a function for calculating the relationship between two positions, $g$ is a function to compute the embedding of input features, and $C$ signifies a normalization factor.

$$\mathrm{f}^{\mathrm{c}}\left(F_{in}\right)=F_{in}\odot\sigma\left(W_3\left(\delta\left(W_2\delta\left(W_1F_{in}\right)\right)+b_3\right)\right) \quad (4)$$

here, $\mathrm{f_c}$ indicates the channel-wise attention module, $F_{in}$ is the input feature map, $\odot$ refers to element-wise multiplication, $\sigma$ represents the Sigmoid function, $\delta$ is the ReLU function, $W_1$, $W_2$, and $W_3$ are convolution kernel parameters, and $b_3$ is the bias parameter.

To accurately calculate the weights across three dimensions—scale, space, and channel—a compact neural network is designed. It consists of several fully connected layers and a Softmax layer, employed for learning the composite attention weights across different dimensions. Specifically, Gated Recurrent Units (GRU) are utilized to update the weights for each dimension. The formal definition is:

$$h_t=GRU\begin{pmatrix}W_{sc}\cdot f^{sc}\left(F_{in}\right)+W_{sp}\cdot f^{sp}\left(F\right)\\+W_c\cdot f^{c}\left(F_{in}\right),h_{t-1}\end{pmatrix} \quad (5)$$

here, $h_t$ represents the hidden state at time $t$, employed for weight calculation. $\mathrm{W_{sc}}$, $W_{sp}$, and $W_c$ are weight matrices corresponding to scale, space, and channel, respectively.

The computation of the weights can be realized through a straightforward fully connected layer:

$$\alpha_{\mathrm{sc}},\alpha_{sp},\alpha_c=Softmax\left(W_h\cdot h_t\right) \quad (6)$$

here, $\alpha_{\mathrm{sc}},\alpha_{\mathrm{sp}},and\ \alpha_{\mathrm{c}}$ denote the weights across the three dimensions.

To enlarge the attention-value gap between the focus pixels and the remaining pixels, a penalty term is introduced in the loss function, known as MSAAM Loss, defined as:

$$\mathrm{MSAAMLoss}=\mathrm{CrossEntropy}\left(\mathrm{Y},\hat{Y}\right)+\lambda\left(Var\left(\alpha_{sc}\right)+Var\left(\alpha_{sp}\right)+Var\left(\alpha_c\right)\right) \quad (7)$$

here, $Y$ is the ground truth, $\hat{Y}$ is the model prediction, and $\lambda$ is a hyperparameter that balances the importance of the two terms. $Var(\alpha)$ indicates the variance of the weights; a higher variance implies that the model has allocated significantly different weights across different scales, spaces, or channels—something we wish to encourage.

In summary, the GRU model maintains a hidden state that captures the significance of the scale, space, and channel information observed thus far. These weights are normalized through a Softmax layer for subsequent use in the attention mechanism. The MSAAM Loss is an extension of the basic cross-entropy loss for semantic segmentation tasks. The second term is a variance term, intended to encourage the model to allocate different weights across the three disparate dimensions—scale, space, and channel—to enhance the model's diversity and robustness. Finally, we merge the network's predicted output with the MSAAM attention map to obtain the final integrated prediction. Such a design helps the model better capture the importance across different scales, spaces, and channels, while also encouraging greater attention to the variances among these dimensions.

## 4. Experiments

### 4.1. Datasets

MSAAM is proposed to improve the semantic segmentation for autonomous driving cars in street scenes, we empirically verify it on CamVid dataset and Cityscapes dataset in this section. CamVid contains 367 training images, 101 validation images, and 233 test images. The resolution of images in this dataset is $960\times720$ which will be downsampled to $480\times360$ for accelerating the training stage of SS models. Cityscapes is comprised of a large, diverse set of high-resolution ($2048\times1{,}024$) images recorded in streets, where 5,000 of these images have high quality pixel-level labels of 19 classes and results $9.43\times10\text{^}9$ labeled pixels in total. Following the standard setting of Cityscapes, the 5,000 images are split into 2,975 training and 500 validation images with publicly available annotation, as well as 1,525 test images with annotations withheld and comparison to other methods is performed via a dedicated evaluation server.

### 4.2. Experimental setup

#### 4.2.1. Implementation details

We adopt DeepLabv3+ with ResNet101 backbone for our segmentation architecture with the output stride set to 8. MSAAM is incorporated at the end of the encoder. We train our segmentation networks on Cityscapes. We use the same pre-trained network for all experiments.

To avoid over-fitting, common data augmentations are used as preprocessing, including random flipping horizontally, random

scaling in the range of [0.5, 2], random brightness jittering within the range of [−10, 10], and random crop of 512 × 512 image patches. For training, we use the Adam optimizer (Kahneman, 1973) with an initial learning rate of 0.0003 and weight decay of 0.00001. The learning rate is scheduled by multiplying the initial learning rate with $\left(1-\frac{epoch}{maxEpoches}\right)^{0.9}$. All models are trained for 80 epochs with minibatch size of 8.

### 4.2.2. Evaluation metrics

For quantitative evaluation, mean of class-wise Intersection over Union (mIoU) are used. We also use the class accuracy (ClassAcc) to evaluate the performance of compared methods on different datasets. We compare the performance by the area under receiver operating characteristics (AUROC) and average precision (AP). In addition, we measure the false positive rate at a true positive rate of 95% (FPR95) since the rate of false positives in high-recall areas is crucial for safety-critical applications.

### 4.2.3. Baselines

In Cityscapes dataset, we pick up 19 the most frequently occurred classes from the original 35 classes based on the official evaluation metrics (Raymond et al., 1992), and their importance groupings from trivial to important are.

Group 1 = {Sky, Building, Vegetation, Terrain, Wall};

Group 2 = {Pole, Road, Sidewalk, Fence};

Group 3 = {Traffic sign, Traffic light, Car, Truck, Bus, Train, Motorcycle, Person, Rider, Bicycle};

We compare our method with important approaches including Synboost, SML, Max logits, Entropy, MSP, Energy, SynthCP, Meta-OoD (Broadbent, 1958; Treisman and Gelade, 1980; Desimone and Duncan, 1995; Hubmann et al., 2018; Lis et al., 2019; Doshi and Yilmaz, 2020; Xia et al., 2020; Blum et al., 2021; Vojir et al., 2021) on test sets of CamVid and on validation sets of Cityscapes. Note that Synboost and SynthCP requires additional training of extra network and utilizing OoD data. Energy and Meta-OoD requires additional training of extra component or network. SML, Max logits, Entropy and MSP do not require additional training or utilize external datasets.

## 4.3. Evaluation results

In this section, we compare the performances of important approaches with MSAAM under the above experimental settings. The experimental results of compared methods on the investigated classes of the two datasets are shown in Tables 1–4, respectively. A more comprehensive set of quantitative analysis metrics is shown in Table 5.

From the results shown in Tables 1, 2, we find that by embedding our MSAAM to the adopted deep models, the performance of the

TABLE 1 The comparison results (%) of various methods on the Groups 1 and 2 of Camvid Dataset.

| Models | Group 1 | | | Group 2 | | | |
|---|---|---|---|---|---|---|---|
| | Sky | Building | Tree | Column | Road | Sidewalk | Fence |
| Synboost | 97.06 | 71.61 | 77.84 | 34.31 | 93.41 | 90.35 | 53.57 |
| SML | 93.77 | 86.75 | 83.29 | 21.59 | 98.28 | 86.38 | 31.38 |
| Max logits | 94.21 | 71.6 | **90.88** | 48.92 | 93.17 | 88.78 | 45.19 |
| Entropy | 89.98 | **88.92** | 84.58 | 9.71 | 94.56 | 81.27 | 19.86 |
| MSP | 93.38 | 87.45 | 83.87 | 17.23 | 90.24 | 88.76 | 43.33 |
| Energy | 85.12 | 86.4 | 71.77 | 20.23 | **98.66** | 75.03 | 25.56 |
| SynthCP | 94.44 | 78.71 | 88.09 | 42.28 | 98.29 | 94.57 | 44.84 |
| Meta-OoD | **97.87** | 86.28 | 81.18 | 30.04 | 98.66 | 86.04 | 32.74 |
| MSAAM | 96.82 | 75.16 | 82.81 | **60.36** | 92.11 | **95.19** | **62.02** |

The bold values mean highlighting the best results in the data comparison.

TABLE 2 The comparison results (%) of various methods on the Group 3 of Camvid Dataset.

| Models | Group 3 | | | | | |
|---|---|---|---|---|---|---|
| | Sign | Car | Pedestrian | Bicyclist | ClassAvg | mIoU |
| Synboost | 50.49 | 82.92 | 67.21 | 33.11 | 71.21 | 51.19 |
| SML | 40.79 | 80.28 | 59.93 | 15.19 | 64.21 | 51.08 |
| Max logits | 26.58 | 79.38 | 39.43 | 42.29 | 67.88 | 52.34 |
| Entropy | 0.72 | 75.37 | 25.09 | 0.48 | 52.32 | 45.35 |
| MSP | 32.33 | 83.53 | 36.08 | 23.45 | 58.91 | 47.71 |
| Energy | 29.39 | 80.82 | 48.08 | 28.25 | 60.11 | 48.51 |
| SynthCP | 43.37 | 76.01 | 66.39 | 52.05 | 72.51 | 55.31 |
| Meta-OoD | 19.58 | 76.56 | 37.65 | 36.08 | 63.07 | 53.21 |
| MSAAM | **67.57** | **91.63** | **78.17** | **62.51** | **74.81** | **55.87** |

The bold values mean highlighting the best results in the data comparison.

TABLE 3 The comparison results (%) of various methods on the Groups 1 and 2 of Cityscapes Dataset.

| Models | Group 1 | | | | | Group 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sky | Building | Vegetation | Terrain | Wall | Pole | Road | Sidewalk | Fence |
| Synboost | 95.57 | 94.27 | 94.73 | **77.53** | 57.85 | **74.28** | 94.89 | 84.84 | 64.16 |
| SML | 99.21 | 92.21 | **97.64** | 66.35 | 35.54 | 49.66 | 98.36 | 82.78 | 59.97 |
| Max logits | 98.58 | 85.37 | 95.73 | 52.48 | 43.38 | 59.65 | 93.39 | 86.97 | 36.92 |
| Entropy | 94.17 | 92.95 | 93.06 | 61.71 | 12.45 | 40.11 | 96.48 | 81.23 | 43.69 |
| MSP | 92.19 | 81.63 | 93.44 | 64.77 | 32.95 | 30.43 | 97.81 | 80.23 | 35.33 |
| Energy | 93.56 | **95.02** | 90.73 | 41.24 | 16.85 | 28.79 | 98.61 | 77.03 | 25.84 |
| SynthCP | **99.52** | 90.52 | 90.79 | 76.21 | **68.52** | 70.03 | 96.80 | 87.28 | **64.95** |
| Meta-OoD | 94.67 | 93.04 | 93.72 | 75.85 | 58.48 | 67.48 | **99.62** | 92.61 | 59.81 |
| MSAAM | 93.55 | 86.86 | 91.26 | 67.14 | 54.47 | 70.73 | 94.66 | **94.48** | 62.03 |

The bold values mean highlighting the best results in the data comparison.

TABLE 4 The comparison results (%) of various methods on the Group 3 of Cityscapes Dataset.

| Models | Group 3 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Traffic Sign | Traffic Light | Car | Truck | Bus | Train | Motorcycle | Person | Rider | Bicycle | ClassAvg | mIoU |
| Synboost | 75.96 | 71.18 | 98.92 | 68.10 | 73.87 | 61.07 | 42.50 | 87.29 | 57.79 | 81.82 | 74.66 | 58.20 |
| SML | 62.75 | 27.42 | 91.60 | 0.00 | 62.93 | 0.00 | 0.00 | 83.05 | 0.00 | 63.91 | 58.52 | 44.84 |
| Max logits | 55.08 | 21.27 | 96.42 | 44.86 | 41.29 | 16.94 | 3.14 | 67.28 | 39.47 | 66.89 | 59.72 | 42.58 |
| Entropy | 15.03 | 7.57 | 90.01 | 13.20 | 1.04 | 52.52 | 2.55 | 62.68 | 0.00 | 50.58 | 45.80 | 38.92 |
| MSP | 45.98 | 14.01 | 91.50 | 1.34 | 29.85 | 1.02 | 0.52 | 67.59 | 3.57 | 61.25 | 48.52 | 40.20 |
| Energy | 42.59 | 11.60 | 93.85 | 2.25 | 3.51 | 11.83 | 0.29 | 61.65 | 0.10 | 57.02 | 46.28 | 37.76 |
| SynthCP | 83.64 | 77.40 | 95.90 | 77.59 | 87.49 | 78.30 | 56.92 | 85.37 | 66.96 | 85.38 | 75.69 | 67.89 |
| Meta-OoD | 74.72 | 67.08 | 96.56 | 72.26 | 82.57 | 72.02 | 53.00 | 87.59 | 64.57 | 81.22 | 79.99 | 69.34 |
| MSAAM | **89.55** | **81.61** | **99.36** | **88.85** | **89.52** | **85.82** | **57.41** | **89.11** | **70.11** | **89.64** | **81.13** | **71.46** |

The bold values mean highlighting the best results in the data comparison.

TABLE 5 The comparison results of various methods on AUROC, AP, and $FPR_{95}$.

| Models | AUROC↑ | AP↑ | FPR95↓ |
|---|---|---|---|
| Synboost | 92.48 | 47.88 | 49.04 |
| SML | 96.77 | 50.09 | 17.37 |
| Max logits | 93.75 | 28.07 | 29.86 |
| Entropy | 90.39 | 21.93 | 34.75 |
| MSP | 88.26 | 14.85 | 33.97 |
| Energy | 92.61 | 30.30 | 38.37 |
| SynthCP | 89.34 | 22.26 | 32.72 |
| Meta-OoD | 97.38 | 67.41 | 13.76 |
| MSAAM | **98.79** | **68.46** | **5.72** |

The bold values mean highlighting the best results in the data comparison.

investigated important classes like sign/symbol, pedestrian, and bicyclist can be significantly improved when compared with the results of other approaches. Not surprisingly, the performance on unimportant classes such as sky, building, and tree weakly drop because they are the target of the attention mechanism. The performance gain of MSAAM over the second approach are 17.08, 8.1, 11.04, 10.46 on sign, car, pedestrian, bicyclist, respectively. Meanwhile, MSAAM achieve better performance than other approaches of ClassAvg and mIoU values.

From the results in Tables 3, 4, we observe that the important classes in Group 3 are segmented with very high performance by MSAAM. The performance gain of MSAAM on ClassAvg and mIoU are 1.14 and 2.12. For some unimportant classes in Group 1 and 2, the performances of the MSAAM-based model are inferior to the other models. However, they will not have a large impact on safe-driving as explained above.

To further evaluate the experimental results through quantitative analysis, we conducted a data analysis on the three metrics, AUROC, AP, $FPR_{95}$ presented in Table 5. From the results, we observe that embedding our MSAAM to the adopted deep models, the performance achieved the best results compared to all other models. The performance gain of MSAAM on AUROC, AP and mIoU are 1.41, 1.05, 8.04, respectively.

## 4.4. Auxiliary hierarchical representation

To qualitatively analyze the experimental results, we design an algorithm to extract the weights from multiple attention modules. It then simplifies the attention pixels into rectangular

blocks for visualization. This algorithm is named auxiliary hierarchical representation.

The original image dimensions are $H \times W \times C$. The attention weights $\alpha_{sc}$, $\alpha_{sp}$ and $\alpha_c$ are extracted from the GRU model. In the Scale Attention Auxiliary Hierarchical Representation, the weight $\alpha_{sc}$ and the scale attention output $f^{sc}(F_{in})$ are utilized to compute an $H \times W$ scale weight matrix. In this matrix, the weight of each pixel $(i, j)$ is the weighted sum of $\alpha_{sc} \cdot f_{ij}^{sc}$ across all scales, defined as follows:

$$ScAH_{ij} = \sum_{s} \alpha_{sc,s} \cdot f_{s,ij}^{sc} \tag{8}$$

here, ScAH stands for Scale Attention Highlight.

In the case of Spatial Attention Auxiliary Hierarchical Representation, the weight $\alpha_{sp}$ and the spatial attention output $f^{sp}(F)$ are employed to calculate an $H \times W$ spatial weight matrix, defined as:

$$SpAH_{ij} = \alpha_{sp} \cdot f_{ij}^{sp} \tag{9}$$

here, SpAH stands for Spatial Attention Highlight.

For Channel Attention Auxiliary Hierarchical Representation, the weight $\alpha_c$ and the channel attention output $f^{c}(F_{in})$ are used to compute an $H \times W$ channel weight matrix. Here, the weight of each pixel $(i, j)$ is the weighted sum of $\alpha_c \cdot f_{ij}^{c}$ across all channels, defined as follows:

$$CAH_{ij} = \sum_{c} \alpha_{c,c} \cdot f_{c,ij}^{c} \tag{10}$$

in this context, CAH represents Channel Attention Highlight.

Upon the completion of the hierarchical model construction, the model undergoes normalization and color mapping to facilitate the high-contrast highlighting of attention regions. For an optimized visual experience, a simplified treatment is generally applied to the regions of attention.

After auxiliary hierarchical modeling is accomplished for all three attention mechanisms—scale, spatial, and channel—their respective weights are combined to create a rectangular attention visualization model, providing a more straightforward and interactive way to represent attention intervals.

Initially, the weights are amalgamated by integrating the weight matrices of Scale, Spatial, and Channel into a new weight matrix termed as Combined Attention Highlight, abbreviated as CoAH. The combination is formalized as:

$$CoAH_{ij} = \alpha_{sc} \cdot ScAH_{ij} + \alpha_{sp} \cdot SpAH_{ij} + \alpha_c \cdot CAH_{ij} \tag{11}$$

here, $\alpha_{sc}$, $\alpha_{sp}$, and $\alpha_c$ are normalized weights retrieved from the GRU model.

Subsequently, a simplified rectangular model is established. A simplification algorithm, such as a greedy algorithm or another optimization technique, is employed to identify a rectangular region with the highest average attention weight. Assuming the rectangular region is defined by the top-left corner $(x_1, y_1)$ and the bottom-right corner $(x_2, y_2)$, the average weight for this area is computed as follows:

$$AW = \frac{1}{(x_2 - x_1 + 1) \times (y_2 - y_1 + 1)} \sum_{i=x_1}^{x_2} \sum_{j=y_1}^{y_2} CoAH_{ij} \tag{12}$$

in this equation, AW stands for Average Weight.

The visualization of the auxiliary hierarchical representation based on the MSAAM attention mechanism is shown in Figure 3. Scale attention captures objects of the focused category at different sizes. Subsequently, spatial attention tends to prioritize obscured targets, while channel attention is inclined toward targets with significant appearance variations. Both spatial and channel attentions assist scale attention in optimizing the areas and objects of focus, culminating in an integrated attention model. Auxiliary hierarchical representation is for the purpose of visualizing this process.

## 4.5. Ablation study

We integrated the MSAAM into the models that do not require additional training or utilize external datasets. These models include SML, Max logits, Entropy and MSP. From the results in Table 6, we observe that all performance metrics of every model improved. The experimental outcomes underscore the versatility and effectiveness of MSAAM.

## 4.6. Comparison on effectiveness

To demonstrate the effectiveness of MSAAM on Cityscapes dataset, Figure 4 shows some representative segmentation results of the SML, Max logits, Entropy and MSAAM. We find that the interested regions segmented by the MSAAM are highly compact, and the shapes of the segmented objects are also more close to that of the ground truth. Therefore, MSAAM is effective in emphasizing the small but critical targets, and thus is useful for semantic segmentation tasks.

## 4.7. Comparison on computational cost

To demonstrate that our method requires a negligible amount of computation cost, we report GFLOPs (i.e., the number of floating-point operations used for computation) and the inference time. As shown in Table 7, our method requires only a minimal amount of computation cost regarding both GFLOPs and the inference time compared to the other approaches.

# 5. Conclusion

In this paper, we present the Multi-Scale Adaptive Attention Mechanism (MSAAM), a specialized framework tailored for enhancing semantic segmentation in automotive environments. The

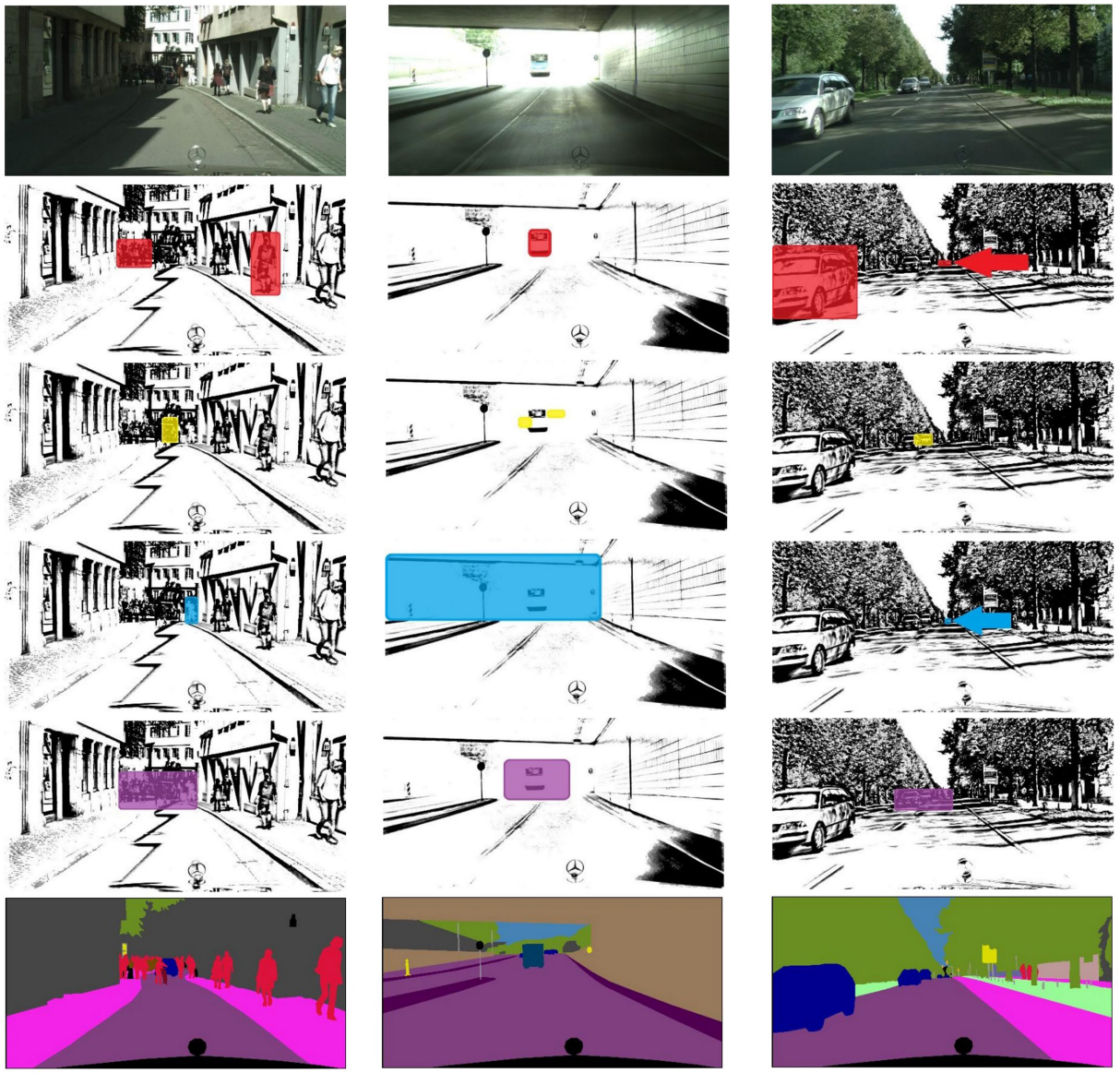


FIGURE 3
Visualization of the working process of the MSAAM attention mechanism.

TABLE 6 Comparison of metric gains after embedding our MSAAM to models that do not require additional training or utilize external datasets.

| | AUROC↑ | AP↑ | FPR95↓ | mIoU |
|---|---|---|---|---|
| SML + MSAAM | +0.63 | +1.66 | +2.70 | +1.49 |
| Max logits + MSAAM | +0.10 | +6.90 | +2.47 | +0.42 |
| Entropy + MSAAM | +1.54 | +7.21 | +1.85 | +1.65 |
| MSP + MSAAM | +1.45 | +2.34 | +1.63 | +0.19 |

attention mechanism uniquely harmonizes three critical dimensions—scale, spatial context, and channel features—while adaptively balancing their respective contributions. By integrating these multi-faceted channels, MSAAM excels in addressing complex scene attributes such as scale discrepancies, object occlusions, and diverse visual appearances. Notably, the architecture of this attention mechanism is highly modular, enabling seamless incorporation into a wide array of Convolutional Neural Network (CNN) models. As a result, it serves as a versatile, plug-and-play component that augments pixel-level semantic segmentation performance without significantly inflating the parameter count or complicating the training regimen.

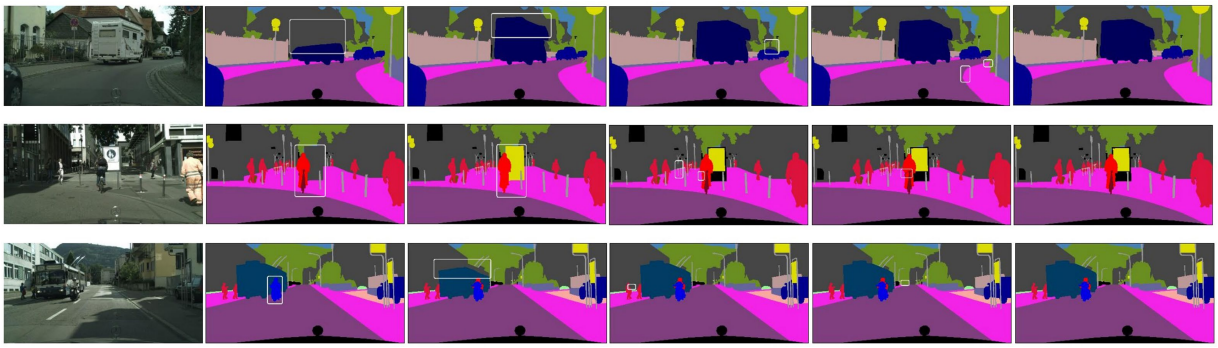


FIGURE 4
Comparison on effectiveness.

TABLE 7 Comparison on computational cost.

| Models | GFLOPs | Infer. Time (ms) |
|---|---|---|
| Synboost | 4762.15 | 165.27 |
| SML | 2139.86 | 61.41 |
| Max logits | 2169.32 | 66.45 |
| Entropy | 2631.33 | 72.88 |
| MSP | 2431.59 | 77.12 |
| Energy | 2201.09 | 70.15 |
| SynthCP | 4551.11 | 146.91 |
| Meta-OoD | 4776.81 | 150.84 |
| MSAAM | 2117.01 | 61.06 |

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.cityscapes-dataset.com/.

## Author contributions

DL: Writing – original draft. DZ: Writing – review & editing. LW: Writing – review & editing. JW: Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. End-to-End Autonomous Driving Research Based on Visual Perception Multi-Task Learning in Small Sample Scenarios (KJ2021A0978), Research on Decision-making Mechanisms of Unmanned Water Quality Monitoring Boats Based on Multi-Sensor Fusion Technology (KJ2021A0982).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Arnsten, A. F. T., and Li, B. M. (2005). Neurobiology of executive functions: catecholamine influences on prefrontal cortical functions. *Biol. Psychiatry* 57, 1377–1384. doi: 10.1016/j.biopsych.2004.08.019

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2015).

Bevandic´, P., Krešo, I., Oršic´, M., and Šegvic´, S. Simultaneous semantic segmentation and outlier detection in presence of domain shift Proceeding German Conference Pattern Recognition (2019). 33–47.

Blum, H., Sarlin, P.-E., Nieto, J., Siegwart, R., and Cadena, C. (2021). The fishyscapes benchmark: measuring blind spots in semantic segmentation. *Int. J. Comput. Vis.* 129, 3119–3135. doi: 10.1007/s11263-021-01511-6

Broadbent, D. E. *Perception and communication*. London: Pergamon Press (1958).

Chan, R., Rottmann, M., and Gottschalk, H. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation Proceeding IEEE/CVF International Conference Computing Vision (2021). 5108–5117.

Cichy, R. M., and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends Cogn. Sci.* 23, 305–317. doi: 10.1016/j.tics.2019.01.009

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Creusot, C., and Munawar, A. Real-time small obstacle detection on highways using compressive RBM road reconstruction. Proceeding IEEE Intelligent Vehicle Symposium (2015). 162–167.

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi: 10.1146/annurev.ne.18.030195.001205

Di Biase, G., Blum, H., Siegwart, R., and Cadena, C. Pixel-wise anomaly detection in complex driving scenes. Proceedings IEEE/CVF Conference Computing Vision Pattern Recognition, (2021). 16913–16922.

Ding, Y., Ma, Z., Wen, S., Xie, J., Chang, D., Si, Z., et al. (2021). Ap-cnn: weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. *IEEE Trans. Image Process.* 30, 2826–2836. doi: 10.1109/TIP.2021.3055617

Doshi, K., and Yilmaz, Y. Fast unsupervised anomaly detection in traffic videos. Proceedings IEEE/CVF Conference Computing Vision Pattern Recognition Work-shops (2020). 624–625.

Fu, J., Zheng, H., and Mei, T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4438–4446 (2017).

Ge, W., Lin, X., and Yu, Y. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3034–3043 (2019).

He, B., Lakshminarayanan, B., and The, Y. W. (2020). Bayesian deep ensembles via the neural tangent kernel. *Adv. Neural Inf. Process. Syst.* 33, 1010–1022.

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings Proceedings 39th Intelligent Conference Machine Learning (2022). 8759–8773.

Hendrycks, D., and Gimpel, K. A baseline for detecting misclas- sified and out-of-distribution examples in neural networks. arXiv:1610.02136 (2016).

Hu, Z., Xing, Y., Gu, W., Cao, D., and Lv, C. (2023). Driver anomaly quantification for intelligent vehicles: a contrastive learning approach with representation clustering. *IEEE Trans. Intell. Veh.* 8, 37–47. doi: 10.1109/TIV.2022.3163458

Hubmann, C., Schulz, J., Becker, M., Althoff, D., and Stiller, C. (2018). Automated driving in uncertain environments: planning with interaction and uncertain maneuver prediction. *IEEE Trans. Intell. Veh.* 3, 5–17. doi: 10.1109/TIV.2017.2788208

Jensen, O., and Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Front. Hum. Neurosci.* 4:186. doi: 10.3389/fnhum.2010.00186

Jiang, D., Yan, H., Chang, N., Li, T., Mao, R., Chi, D., et al. (2020). Convolutional neural network-based dosimetry evaluation of esophageal radiation treatment planning. *Med. Phys.* 47, 4735–4742. doi: 10.1002/mp.14434

Jin, Y., Ren, X., Chen, F., and Zhang, W. Robust monocular 3D lane detection with dual attention. Proceedings of IEEE International Conference Image Process (2021). 3348–3352.

Jung, S., Lee, J., Gwak, D., Choi, S., and Choo, J., Standardized max logits: a simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. Proceedings IEEE/CVF International Conference Computing Vision (2021). 15425–15434.

Kahneman, D. *Attention and effort*. Englewood Cliffs: Prentice-Hall (1973).

Kendall, A., Badrinarayanan, V., and Cipolla, R. Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv:1511.02680 (2015).

Kendall, A., and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision?. Proceedings 31st International Conference Advanced Neural Information Process System. (2017). 30, 5579–5584.

Li, T., Cheng, B., Ni, B., Liu, G., and Yan, S. (2016a). Multitask low-rank affinity graph for image segmentation and image annotation. *Trans. Intell. Syst. Technol.* 7, 1–18. doi: 10.1145/2856058

Li, T., Meng, Z., Ni, B., Shen, J., and Wang, M. (2016b). Robust geometric ℓp-norm feature pooling for image classification and action recognition. *Image Vis. Comput.* 55, 64–76.

Lis, K., Nakka, K., Fua, P., and Salzmann, M. Detecting the unexpected via image resynthesis. Proceedings IEEE/CVF International Conference Computing Vision (2019). 2152–2161.

Liu, S., Zhang, J., Li, T., Yan, H., and Liu, J. (2021). Technical note: a cascade 3D U-net for dose prediction in radiotherapy. *Med. Phys.* 48, 5574–5582. doi: 10.1002/mp.15034

Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. (2019). Multi-model attentional neural network for sentiment analysis. *Inf. Sci.* 497, 237–250.

Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Adv. Neural Inf. Proces. Syst.* 27, 2204–2212.

Nian, F., Li, T., Wu, X., Gao, Q., and Li, F. (2016). Efficient near-duplicate image detection with a local-based binary representation. *Multimed. Tools Appl.* 75, 2435–2452. doi: 10.1007/s11042-015-2472-1

Nobre, A. C., and van Ede, F. (2018). Anticipated moments: temporal structure in attention. *Nat. Rev. Neurosci.* 19, 34–48. doi: 10.1038/nrn.2017.141

Ohgushi, T., Horiguchi, K., and Yamanaka, M. Road obstacle detection method based on an autoencoder with semantic segmentation. Proceeding Asian Conference Computing Vision (2020). 223–238.

Rao, T., Li, X., Zhang, H., and Xu, M. (2019). Multi-level region-based convolutional neural network for image emotion classification. *Neuro Comput.* 333, 429–439. doi: 10.1016/j.neucom.2018.12.053

Raymond, J. E., Shapiro, K. L., and Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* 18, 849–860. doi: 10.1037/0096-1523.18.3.849

Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., and Carneiro, G. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. arXiv:2111.12264 (2021).

Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30, 5998–6008.

Vojir, T., Šipka, T., Aljundi, R., Chumerin, N., Reino, D. O., and Matas, J. Road anomaly detection by partial image reconstruction with segmentation coupling. Proceedings IEEE/CVF International Conference Computing Vision (2021). 15631–15640.

Xia, Y., Zhang, Y., Liu, F., Shen, W., and Yuille, A. L. Synthesize then compare: detecting failures and anomalies for semantic segmentation. Proceedings European Conference Computing Vision (2020). 145–161.

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., and Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 842–850 (2015).

Zhang, J., Li, Y., Li, T., Xun, L., and Shan, C. (2019). License plate localization in unconstrained scenes using a two-stage CNN-RNN. *IEEE Sensors J.* 19, 5256–5265. doi: 10.1109/JSEN.2019.2900257

Zhang, J., Liu, S., Yan, H., Li, T., Mao, R., and Liu, J. (2020). Predicting voxel-level dose prediction for esophageal radiotherapy using densely connected network with dilated convolutions. *Phys. Med. Biol.* 65:205013. doi: 10.1088/1361-6560/aba87b

Zheng, H., Fu, J., Zha, Z.-J., and Luo, J. Looking for the devil in the details: learning trilinear attention sampling network for fine- grained image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5012–5021 (2019).

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 06 November 2023
DOI 10.3389/fnins.2023.1293161



OPEN ACCESS

EDITED BY
Fudong Nian,
Hefei University, China

REVIEWED BY
Hu Guohua,
Hefei University, China
Jingjing Shi,
Hefei Normal University, China

*CORRESPONDENCE
Dong Sun
✉ sundong@ahu.edu.cn

RECEIVED 12 September 2023
ACCEPTED 09 October 2023
PUBLISHED 06 November 2023

CITATION
Ning W, Sun D, Gao Q, Lu Y and Zhu D (2023) Natural image restoration based on multi-scale group sparsity residual constraints. *Front. Neurosci.* 17:1293161. doi: 10.3389/fnins.2023.1293161



# Natural image restoration based on multi-scale group sparsity residual constraints

Wan Ning, Dong Sun*, Qingwei Gao, Yixiang Lu and De Zhu

Anhui Engineering Laboratory of Human-Robot Integration System and Intelligent Equipment, Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Electrical Engineering and Automation, Anhui University, Hefei, China

The Group Sparse Representation (GSR) model shows excellent potential in various image restoration tasks. In this study, we propose a novel Multi-Scale Group Sparse Residual Constraint Model (MS-GSRC) which can be applied to various inverse problems, including denoising, inpainting, and compressed sensing (CS). Our new method involves the following three steps: (1) finding similar patches with an overlapping scheme for the input degraded image using a multi-scale strategy, (2) performing a group sparse coding on these patches with low-rank constraints to get an initial representation vector, and (3) under the Bayesian maximum a posteriori (MAP) restoration framework, we adopt an alternating minimization scheme to solve the corresponding equation and reconstruct the target image finally. Simulation experiments demonstrate that our proposed model outperforms in terms of both objective image quality and subjective visual quality compared to several state-of-the-art methods.

KEYWORDS

**image restoration, group sparsity residual, low-rank regularization, multi-scale, non-local self-similarity (NSS)**

## 1. Introduction

Unsuitable equipment and other disturbances unavoidably contribute noise in the target images. Image denoising is a crucial area of image processing and has attracted much attention from scholars in related fields recently. Digital image denoising techniques have a wide range of uses, involving disciplines of medicine and industry, and also in spectral images for weather forecasting, remote sensing images, and so on. Taking image denoising as a basis, the method can be introduced to more image restoration problems and be useful in more fields (Buades et al., 2005; Osher et al., 2005; Elad and Aharon, 2006; Zoran and Weiss, 2011; Gu et al., 2014; Zhang et al., 2014b; Liu et al., 2017; Keshavarzian et al., 2019; Ou et al., 2020; Zha et al., 2020a, 2022; Jon et al., 2021). This task aims to generate a latent image $\boldsymbol{x}$ from the degraded version $\boldsymbol{y}$. The process modeling can be depicted as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

Where $\boldsymbol{H}$ is an irreversible linear operator in matrix form and $\boldsymbol{n}$ is the additive white Gaussian noise vector. By requiring $\boldsymbol{H}$, Eq.(1) can be converted to diverse image restoration problems. For example, Eq.(1) represents the image denoising problem if $\boldsymbol{H}$ is an identity (Elad and Aharon, 2006; Ou et al., 2020); Eq.(1) denotes the image inpainting problem if $\boldsymbol{H}$ is a mask (Liu et al., 2017; Zha et al., 2020a); and Eq.(1) stands for the image CS problem if $\boldsymbol{H}$ is an undersampled random projection matrix (Keshavarzian et al., 2019; Zha et al., 2022). We concentrate on image denoising, inpainting, and CS challenges in this article.

Given that the problem always ill-posed, it is common to use image priors to regularize the model so as to gain excellent restored images. Namely, the Maximum A Posteriori (MAP) approach allows for the image restoration problem to be formulated as a mathematical equation to address the minimization problem:

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 + \lambda \boldsymbol{R}(\boldsymbol{x}) \tag{2}$$

The former is the data-fidelity term and the latter is the image prior constraint term. The weights between these two terms are regulated by the parameter λ. After establishing the mathematical model, we conceived an optimization algorithm to address various image restoration problems. The method yields a reconstructed image that approximates a clean image after several iterations.

Numerous image prior models have been put forward in earlier studies, mainly classified into local smoothness (Rudin et al., 1992; Osher et al., 2005; Dey et al., 2006), non-local self-similarity (Fazel et al., 2001; Buades et al., 2005; Gu et al., 2014), and sparsity (Zhang et al., 2014b; Ou et al., 2020, 2022a). Yet, the curse of dimensionality makes it difficult to construct a global model for the entire image. Therefore, the approach of building patch priors has become popular in recent years for its efficiency and convenience.

Sparse representation is one of the most representative patch-based priors. Elad and Aharon (2006) proposed K-SVD (K-Singular Value Decomposition) which is a pioneering work in applying sparse coding to image denoising. NSS is another crucial prior information widely used. Buades et al. (2005) proposed the first model using NSS for image denoising. In addition, the high correlation between patches leading to the data matrix of a clean image is as often low-rank. Related studies mainly fall into two categories: low-rank matrix factorization (LRMF) (Srebro and Jaakkola, 2003; Buades et al., 2005) and the Nuclear Norm Minimization (NNM) (Fazel et al., 2001; Hu et al., 2012). NNM is the more popular one in most cases. Gu et al. presented the Weighted Nuclear Norm Minimisation model (WNNM) (Gu et al., 2014) which dramatically enhances the flexibility of NNM, and it remains among most widespread image denoising methods. Apart from this, RRC (Zha et al., 2019), which makes use of low-rank residuals for modeling, has also achieved good quality in various image restoration problems.

Some studies have combined image sparsity and self-similarity to modeling, and these algorithms have shown great potential in image restoration research. For instance, in the study by Dabov et al. (2007), BM3D applies NSS to cluster patches before collaborative filtering, which is a benchmark method in the current area of image denoising. Both NCSR (Dong et al., 2012b) and GSR (Zhang et al., 2014b) use the NSS property to aggregate image patches into groups, and then perform sparse coding on the self-similar groups. Mairal J et al. devised the LSSC (Mairal et al., 2009) to force all self-similar groups to be imposed with the same dictionary. Zha et al. (2017) designed an efficient GSRC model that converts the task of image denoising into one of minimizing group sparse residuals. In addition, Zha et al. (2020a) also proposed a GSRC-NLP model with a better image restoration result based on the above.

Another groundbreaking patch-based image recovery method is Expected Patch Log Likelihood (EPLL) (Zoran and Weiss, 2011) which restores images by learning a Gaussian mixture model(GMM). Later on, Zoran et al. introduces a multi-scale EPLL (Papyan and Elad, 2015) model, which can improve the performance of image restoration further. Subsequently, image denoising methods using external GMM priors have been widely used. Most of the relevant studies have combined external GMM with internal NSS for modeling, such as Xu et al. (2015) proposed PGPD, Chen et al. (2015) proposed PCLR, and Zha et al. (2020b) proposed SNSS.

In addition to the above methods, deep convolutional neural networks (CNNs) (Zhang et al., 2017; Zhang and Ghanem, 2018) is an emerging approach in recent years, but it requires learning in an external database before restoring damaged images.

It is not comprehensive to only consider the sparsity or low-rankness property of the image. Hence, with the aim of obtaining a higher-quality restored image, our study uses the low-rank property of similar groups as a constraint in combination with sparsity to design the model. Furthermore, based on the NSS property, we can not only find similar patches for a specified patch on a single scale image but also extend the search window to multi-scales. Finally, we propose a novel Multi-scale Group Sparsity Residual Constraint (MS-GSRC) model with the following innovations:

1. We propose a novel MS-GSRC model that provides a simple yet effective approach for image restoration: find neighbor patches with an overlapping scheme for the input degraded image using a multi-scale strategy and perform a group sparse coding on these similar patches with a low-rank constraint.
2. An alternating minimization mechanism with an automatically tuned parameter scheme is applied to our proposed model, which guarantees a closed-form solution at each step.
3. Our proposed MS-GSRC model is validated on three tasks: denoising, inpainting, and compressed sensing. The model performs competitive in both objective image quality and subjective visual quality compared to several state-of-the-art image restoration methods.

The remainder of this article is as follows: In Section 2, after the brief overview of the GSRC framework and LR methods, we introduce a novel MS-GSRC model. Section 3 adopts an alternating minimization scheme with self-adjustable parameters to resolve our proposed algorithm. Section 4 lists extensive experimental results that prove the feasibility of our model. Conclusion is presented in Section 5.

## 2. Models

In this part, we briefly review some relevant knowledge and present our new model.

### 2.1. Group-based sparse representation

Principles of the GSR model can be described as follows: divide the image into many overlapping patches, find self-similarity groups for each image patch using the NSS property, perform sparse coding for each self-similarity group, and finally reconstruct the image (Dong et al., 2012b; Zha et al., 2020a; Ou et al., 2022a).

Specifically, the image $\boldsymbol{x} \in \mathbb{R}^M$ is divided into $m$ overlapping patches $\{\boldsymbol{x_i}\}_{i=1}^m$, where $\boldsymbol{x_i} \in \mathbb{R}^{n\times n}$. Next, for each overlapping patch $\boldsymbol{x_i}$, we use the K-Nearest Neighbor classification (KNN) algorithm (Keller et al., 1985; Xie et al., 2016) to select $k$ neighbor patches from a $W \times W$ search window to form the group $\boldsymbol{K_i}$. Subsequently, stack all $\boldsymbol{K_i}$ into a data matrix $\boldsymbol{X_i} \in \mathbb{R}^{n\times k}$; this matrix contains each element of $\boldsymbol{K_i}$ as its column, i.e., $\boldsymbol{X_i} = \{x_{i,1}, x_{i,2} \ldots x_{i,k}\}$, where $\{\boldsymbol{x_{i,j}}\}_{j=1}^k$ denotes the k-th similar patch of the k-th group. Each similarity group $\boldsymbol{X_i}$ is represented sparsely as $\hat{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{D}\hat{\boldsymbol{B}}_{\boldsymbol{i}}$, where $\boldsymbol{D_i}$ denotes the dictionary.

Nevertheless, solving the 0-norm minimization problem is NP-hard, so for the ease of making the solution, The sparse code $\hat{\boldsymbol{B}}_{\boldsymbol{i}}$ is obtained from the following equation (Zhang et al., 2014b):

$$\hat{\boldsymbol{B}}_{\boldsymbol{i}} = \min_{\boldsymbol{B_i}} \left( \frac{1}{2} \|\boldsymbol{X_i} - \boldsymbol{D_i}\boldsymbol{B_i}\|_F^2 + \lambda \|\boldsymbol{B_i}\|_1 \right) \quad \forall i \tag{3}$$

It is well-known that clean images $\boldsymbol{x}$ are unavailable in image restoration problems. Thus, we replace $\boldsymbol{x}$ with degenerate images $\boldsymbol{y} \in \mathbb{R}^{M\times M}$. Eq.(3) can be transformed into the problem of recovering the group sparse code $\boldsymbol{A}_i$ from $\boldsymbol{Y_i}$:

$$\hat{\boldsymbol{A}}_{\boldsymbol{i}} = \min_{\boldsymbol{A_i}} \left( \frac{1}{2} \|\boldsymbol{Y_i} - \boldsymbol{D_i}\boldsymbol{A_i}\|_F^2 + \lambda \|\boldsymbol{A_i}\|_1 \right) \quad \forall i \tag{4}$$

The restored $\boldsymbol{X_i}$ is obtained by $\hat{\boldsymbol{X}}_{\boldsymbol{i}} = \boldsymbol{D_i}\hat{\boldsymbol{A}}_{\boldsymbol{i}}$, and the final complete image $\boldsymbol{X}$ can be gained by simple averaging $\{\boldsymbol{X_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{m}}$.

## 2.2. Group sparsity residual constraint

After observing the GSR model, it is clear that the closer the computed $\boldsymbol{A}$ approximates to $\boldsymbol{B}$, the better the quality of the final restoration image. Consequently, the following definition of the group sparsity residual constraint (GSRC) (Zha et al., 2017) is given: $\boldsymbol{R} = \boldsymbol{A} - \boldsymbol{B}$. Then, Eq.(4) for solving the group sparse coefficients $\boldsymbol{A_i}$ can be converted into:

$$\hat{\boldsymbol{A}}_{\boldsymbol{i}} = \min_{\boldsymbol{A_i}} \left( \frac{1}{2} \|\boldsymbol{Y_i} - \boldsymbol{D_i}\boldsymbol{A_i}\|_F^2 + \lambda \|\boldsymbol{A_i} - \boldsymbol{B_i}\|_1 \right) \quad \forall i \tag{5}$$

This model uses BM3D to restore the the degenerate observation $\boldsymbol{y}$ to the image $\boldsymbol{z}$. Moreover, $\boldsymbol{z}$ can be viewed as a good approximation of the target $\boldsymbol{x}$ considering BM3D has an excellent denoising performance. Thus, the group sparsity coefficients $\boldsymbol{B_i}$ can be obtained from $\boldsymbol{z}$. In the study by Zha et al. (2020a), GSRC-NLP uses NLP before constraining the input image.

## 2.3. Low-rank approximation

According to Gu et al. (2014), Zha et al. (2019), and Zha et al. (2020b), it can be found that NNM is a popular low-rank approximations methods. For $\boldsymbol{X}$, define the i-th singular value as $\sigma_i(\boldsymbol{x})$, and the nuclear norm as $\|\boldsymbol{X}\|_* = \Sigma_i \sigma_i(\boldsymbol{x})$. The specific solution for $\boldsymbol{X}$ is:

$$\hat{\boldsymbol{X}} = \min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{X}\|_F^2 + \|\boldsymbol{X}\|_* \quad \forall i \tag{6}$$

Equation (6) yields a simple solution: $\hat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{S}_{\boldsymbol{\tau}}\boldsymbol{V}^{\boldsymbol{T}}$, where $\hat{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}$ is the SVD for $\boldsymbol{Y}$ and $\boldsymbol{S}_{\boldsymbol{\tau}}(\Sigma)$ is a soft-thresholding (Cai et al., 2010) function. Namely, $\boldsymbol{S}_{\boldsymbol{\tau}}(\boldsymbol{\Sigma})_{\boldsymbol{ii}} = max(\boldsymbol{\Sigma}_{\boldsymbol{ii}} - \boldsymbol{\tau}, \boldsymbol{0})$, where $\boldsymbol{\Sigma}_{\boldsymbol{ii}}$ is the diagonal element of $\boldsymbol{\Sigma}$.

## 2.4. Multi-scale GSRC

The established GSRC model has performed well in image denoising, but it requires additional pre-processing of degraded images for obtaining the group sparsity coefficients $\boldsymbol{B}$. Thus, we combine group sparsity and low-rank property to build a model. Furthermore, the GSRC model only focuses on a single scale. However, it is evident that NSS can appear not only on the original scale of an image but also on a coarse scale, so we can find neighbor patches for the original image patch at multi-scales (Yair and Michaeli, 2018; Ou et al., 2022a,b). The specific steps of our proposed new Multi-Scale Group Sparse Residual Constraint (MS-GSRC) model are as follows:

(a) First, we use KNN to find a specified number of similar patches from both the original scale and scaled-down version for the overlapping patches of the input image.

(b) Then, these similar patches are stacked separately into groups.

(c) Next, the low-rank constraint is imposed on each group to obtain good group sparsity coefficients $\boldsymbol{B_i}$.

(d) After estimating the group sparsity coefficients $\boldsymbol{A_i}$ by using the group sparsity residuals $\boldsymbol{R_i}$, each group was recovered in sequence.

(e) Finally, we select the patch belonging to the original image from each group, and aggregate the complete image by simple averaging.

We propose the following constraint function:

$$\begin{aligned} \hat{\boldsymbol{x}} = \min_{\boldsymbol{x}} \frac{1}{2\boldsymbol{\sigma}_{\boldsymbol{n}}^2} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|_2^2 + \frac{1}{2\mu} \sum_{i=1}^{m} \|\boldsymbol{R_i}\boldsymbol{x}^{\boldsymbol{MS}} \\ -\boldsymbol{D_i}\boldsymbol{A_i}\|_F^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{A_i} - \boldsymbol{B_i}\|_1 \quad \forall i \end{aligned} \tag{7}$$

$\boldsymbol{R_i}\boldsymbol{x}^{\boldsymbol{MS}}$ is a multi-scale similarity group, which is a matrix with $\boldsymbol{k}$ nearest neighbor patches matched for each original image patch. These similar patches are derived from both the original and coarse scales of the image. The window size is $W \times W$ in the original scale, and it is $\chi W \times \chi W$ in the other scale images, where $\boldsymbol{\chi}$ indicates the scale factor ($0 < \boldsymbol{\chi} < 1$). $\boldsymbol{\chi}$ will be set to different values in different experiments.

For image denoising, for example, the flowchart of MS-GSRC model is shown in Figure 1.

# 3. Algorithm for image restoration

This section is a detailed analysis of our proposed MS-GSRC model. The solution of this algorithm is obtained using an alternating minimization method whose parameter is self-adjustment.

First, we divide Eq.(7) into three sub-problems:

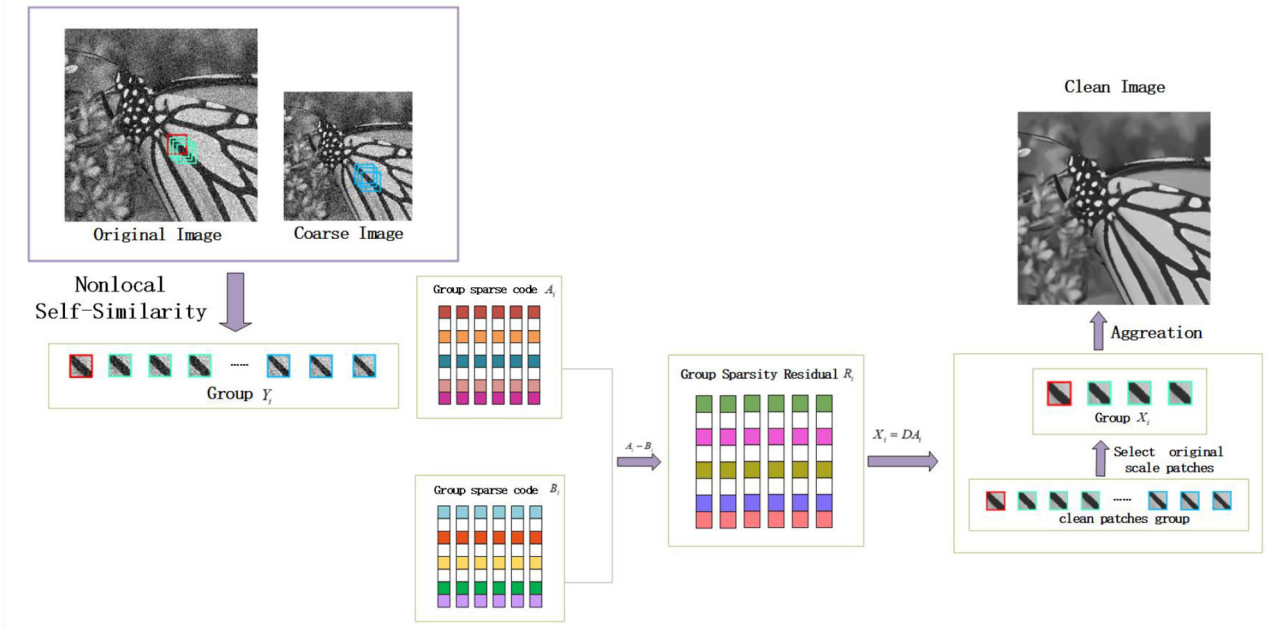



FIGURE 1
Flowchart of the proposed MS-GSRC for image denoising.

**Require:** The observation $y$ and the degradation operator $H$.
Initialize $\hat{x}^0 = y$ and parameters $m$, $n$, $W$, $\gamma$, $\zeta$, $\alpha$, $\beta$, $\chi$, $\varepsilon$,$\epsilon$ and $Iter$.
**for** t=1:Iter **do**
  Update parameters by Eq.(21-23);
  divide $x^{(t)}$ into patches $\{x_i\}_{i=1}^{m}$.
  **for** each $x_i$ **do**
    Construct multi-scale group $R_i x^{MS}$;
  **end for**
  **for** each group $R_i x^{MS}$ **do**
    Construct dictionary $D_i$ by $P_i$ by using PCA;
    Compute $A_i$ by Eq.(9);
    Compute $B_i$ by Eq.(12);
  **end for**
  ADMM:
  Initialize:$c = 0$ and $s = \hat{x}$.
  Compute $s^{(t+1)}$ by Eq.(18);
  Compute $c^{(t+1)}$ by Eq.(17);
  **if** H is an unnstructured random projection matrix **then**
    Construct $x^{(t+1)}$ by Eq.(20);
  **else**
    Construct $x^{(t+1)}$ by Eq.(19);
  **end if**
**end for**
Output:Restored image $\hat{x}$.

Algorithm 1. The MS-GSRC algorithm for image restoration.

## 3.1. $A_i$ sub-problem

Given $x$ and $D_i$, we get a sub-problem of $A_i$:

$$\begin{aligned}\hat{A}_i &= \min_{A_i} \sum_{i=1}^{n} \frac{1}{2\mu} \left\| R_i x^{MS} - D_i A_i \right\|_F^2 + \lambda \left\| A_i - B_i \right\|_1 \\ &= \min_{A_i} \sum_{i=1}^{m} \left\| P_i - A_i \right\|_F^2 + 2\lambda\mu \left\| A_i - B_i \right\|_1 \\ &= \min_{\alpha_i} \sum_{i=1}^{m} \left\| p_i - \alpha_i \right\|_F^2 + 2\lambda\mu \left\| \alpha_i - \beta_i \right\|_1\end{aligned} \tag{8}$$

where $P_i = D_i^{-1} R_i x^{MS}$, $\alpha_i$, $\beta_i$, $p_i$ stand for the vector representations of $A_i$, $B_i$, and $P_i$, respectively. $D_i$ is a dictionary, A crucial step for solving the $A_i$ problem is to design an efficient $D_i$. The restored image is prone to visual artifacts (Lu et al., 2013) if learning the over-complete dictionary. To reduce this terrible phenomenon, we choose to adopt principal component analysis (PCA) (Abdi and Williams, 2010) for learning the dictionary $D_i$ in this study because PCA is more robust and adjustable.

Equation (8) can be deduced as a closed-form solution:

$$\hat{A}_i = Soft\left(\tilde{p}_i - \tilde{\beta}_i, 2\lambda\mu\right) + \tilde{\beta}_i \quad \forall i \tag{9}$$

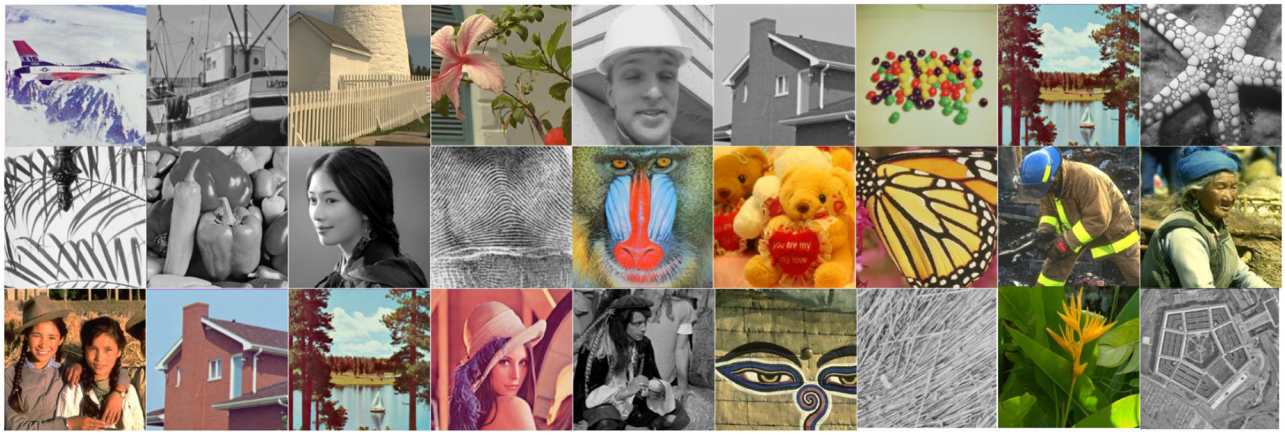


FIGURE 2
The 27 widely tested images for experiences.

$\boldsymbol{Soft}(\cdot)$ represents the soft-thresholding operator.

Since $\boldsymbol{x}$ is an unknown target image, it is impossible to gain the group sparse coefficients $\boldsymbol{B_i}$ directly. Consequently, we must utilize methods to gain an approximation value.

The introduction of low-rank constraints into the model is a practical approach. After applying LR constraints to the $\boldsymbol{Y_i}$ group, we can obtain a matrix $\boldsymbol{S_i}$. The clean group sparsity coefficients $\boldsymbol{B_i}$ can be computed from $\boldsymbol{S_i}$. It is easy to derive the following equation:

$$\|\boldsymbol{Y_i} - \boldsymbol{S_i}\|_F^2 = \|\boldsymbol{A_i} - \boldsymbol{B_i}\|_F^2 \tag{10}$$

So,we can obtain

$$\hat{\boldsymbol{S}}_i = \arg\min_{\boldsymbol{S_i}} \frac{1}{2} \|\boldsymbol{Y_i} - \boldsymbol{S_i}\|_F^2 + \theta \|\boldsymbol{S_i}\|_*$$
$$\Leftrightarrow \hat{\boldsymbol{B}}_i = \arg\min_{\boldsymbol{B_i}} = \frac{1}{2} \|\boldsymbol{A_i} - \boldsymbol{B_i}\|_F^2 + \theta \|\boldsymbol{B_i}\|_*$$

where $\|\boldsymbol{B_i}\|_* = \Sigma_j \delta_{i,j}$, $\{\delta_{i,j}\}_{j=1}^s$ are singular value of matrix. Apparently, we are able to get a closed-form solution for $\boldsymbol{B_i}$:

$$\hat{\boldsymbol{B}}_i = \boldsymbol{U_i soft}(\boldsymbol{\Delta_i}, \theta)\, \boldsymbol{V_i^T} \quad \forall i \tag{11}$$

where $\boldsymbol{A_i} = \boldsymbol{U_i \Delta_i V_i^T}$ is the SVD for $\boldsymbol{A_i}$ and $\boldsymbol{\Delta_i}$ is the diagonal element of the singular matrix.

## 3.2. $\boldsymbol{x}$ sub-problem

Given $\boldsymbol{A_i}$ and $\boldsymbol{D_i}$, subproblem of $\boldsymbol{x}$ in Eq.(7) turns into:

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x}} \frac{1}{2\boldsymbol{\sigma_n^2}} \left\|\boldsymbol{y} - \boldsymbol{Hx}\right\|_2^2 + \frac{1}{2\mu} \sum_{i=1}^{m} \left\|\boldsymbol{R_i x^{MS}} - \boldsymbol{D_i A_i}\right\|_F^2 \tag{12}$$

Clearly, Eq.(13) is a quadratic optimization equation. We adopt Alternate Directional Multiplication Method (ADMM) (Boyd et al., 2011) to simplify the optimization process.

First, we bring in an auxiliary variable $\boldsymbol{s} = \boldsymbol{x^{MS}}$, and Eq.(13) can be converted into an equivalence equation:

$$\langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{s}} \rangle = \min_{\boldsymbol{x,s}} \frac{1}{2\boldsymbol{\sigma_n^2}} \left\|\boldsymbol{y} - \boldsymbol{Hx}\right\|_2^2 + \frac{1}{2\mu} \sum_{i=1}^{m} \|\boldsymbol{R_i s} - \boldsymbol{D_i A_i}\|_F^2 \tag{13}$$

By observing Eq.(14), it is plain that this equation has three unknown variables requiring solutions. Thus, we decompose Eq.(14) into three iterative processes. In the t-th iteration:

$$\hat{\boldsymbol{x}}^{\boldsymbol{t+1}} = \min_{\boldsymbol{x}} \tfrac{1}{2\boldsymbol{\sigma_n^2}} \|\boldsymbol{y} - \boldsymbol{Hx}\|_2^2 + \tfrac{1}{2\zeta} \left\|\boldsymbol{x} - \boldsymbol{s^t} - \boldsymbol{c^t}\right\|_2^2 \tag{14}$$

$$\hat{\boldsymbol{s}}^{\boldsymbol{t+1}} = \min_{\boldsymbol{s}} \tfrac{1}{2\mu} \textstyle\sum_{i=1}^{m} \|\boldsymbol{R_i s} - \boldsymbol{D_i A_i}\|_F^2 + \tfrac{1}{2\zeta} \left\|\boldsymbol{x^{t+1}} - \boldsymbol{s} - \boldsymbol{c^t}\right\|_2^2 \tag{15}$$

$$\boldsymbol{c^{t+1}} = \boldsymbol{c^t} - \left(\boldsymbol{x^{t+1}} - \boldsymbol{s^{t+1}}\right) \tag{16}$$

The parameter $\boldsymbol{c}$ indicates the Lagrangian multiplier. To make the derivation process look more concise, we omit $\boldsymbol{t}$ in the following formulation expression.

***Update s*** : Given $\boldsymbol{D_i A_i}$, $\boldsymbol{x}$, and $\boldsymbol{c}$, $\boldsymbol{s}$ can be represented as a closed-form solution by Eq.(16), namely:

$$\hat{\boldsymbol{s}} = \left(\mu \boldsymbol{I} + \zeta \sum_{i=1}^{m} \boldsymbol{R_i^T R_i}\right)^{-1} \left(\mu \boldsymbol{x} - \mu \boldsymbol{c} + \zeta \sum_{i=1}^{m} \boldsymbol{R_i^T D_i A_i}\right) \tag{17}$$

Since $\boldsymbol{I}$ is a matrix of identities and $\boldsymbol{R_i^T R_i}$ represents a diagonal matrix, $\left(\mu \boldsymbol{I} + \zeta \sum_{i=1}^{m} \boldsymbol{R_i^T R_i}\right)$ is positive. So the above formula is valid.

***Update x*** : Given $\boldsymbol{s}$ and $\boldsymbol{c}$, Eq.(15) provides a solution to the variable $\boldsymbol{x}$:

$$\hat{\boldsymbol{x}} = \left(\zeta \boldsymbol{H^T H} + \sigma_n^2 \boldsymbol{I}\right)^{-1} \left(\zeta \boldsymbol{H^T y} + \sigma_n^2 \boldsymbol{s} + \sigma_n^2 \boldsymbol{c}\right) \tag{18}$$

Notably, since $\boldsymbol{H}$ is an unstructured random projection matrix, the cost required to solve $\boldsymbol{x}$ using Eq.(19) directly is too high in CS

TABLE 1 PSNR (dB) and SSIM comparison of different methods for image denoising.

| Image | Airplane | Flower | Foreman | J.Bean | Lake | Leaves | Lena | Lin | Monarch | Starfish | Pentagon | Peppers | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ = 15 | | | | | | | | | | | | | |
| BM3D | 32.14 | 31.57 | 35.68 | 35.70 | 30.45 | 31.72 | 33.04 | 34.23 | 31.86 | 31.15 | 29.68 | 31.80 | 32.42 |
| | 0.9230 | 0.9074 | 0.9178 | 0.9693 | 0.9063 | 0.9648 | 0.9209 | 0.9243 | 0.9353 | 0.8958 | 0.8716 | 0.8764 | 0.9177 |
| PGPD | 32.31 | 31.85 | 35.51 | 35.65 | 30.67 | 32.02 | 33.13 | 34.16 | 32.23 | 31.31 | 29.72 | 31.78 | 32.53 |
| | 0.9193 | 0.9076 | 0.9140 | 0.9582 | 0.9086 | 0.9671 | 0.9185 | 0.9110 | 0.9362 | 0.9024 | 0.8724 | 0.8725 | 0.9156 |
| WNNM | 32.47 | 32.04 | **35.88** | 36.56 | 30.83 | 32.83 | **33.34** | **34.47** | 32.72 | **31.83** | **30.06** | **32.03** | 32.92 |
| | 0.9252 | 0.9132 | 0.9234 | 0.9735 | 0.9129 | 0.9735 | 0.9248 | **0.9227** | 0.9424 | **0.9081** | 0.8810 | **0.8770** | 0.9231 |
| NCSR | **32.95** | 31.77 | 35.52 | **37.89** | **31.21** | 32.16 | 33.04 | 34.27 | 32.31 | 31.46 | 29.93 | 31.86 | 32.86 |
| | 0.9201 | 0.9082 | 0.9189 | **0.9782** | 0.8965 | 0.9694 | 0.9192 | 0.9190 | 0.9401 | 0.9042 | 0.8779 | 0.8725 | 0.9187 |
| RRC | 32.38 | 31.81 | 35.71 | 36.16 | 30.70 | 32.55 | 33.23 | 34.31 | 32.61 | 31.50 | 29.72 | 31.85 | 32.71 |
| | 0.9248 | 0.9076 | 0.9225 | 0.9734 | 0.9091 | 0.9719 | 0.9233 | 0.9197 | 0.9435 | 0.8988 | 0.8693 | 0.8720 | 0.9197 |
| LGSR | 32.47 | 32.02 | **35.88** | 36.40 | 30.84 | 32.73 | 33.32 | 34.43 | 32.69 | 31.67 | 30.05 | 32.00 | 32.87 |
| | 0.9255 | 0.9127 | **0.9244** | 0.9751 | 0.9130 | 0.9732 | **0.9249** | 0.9222 | 0.9435 | 0.9040 | 0.8818 | 0.8753 | 0.9230 |
| GSRC-NLP | 32.37 | 31.97 | 35.84 | 36.10 | 30.76 | 32.61 | 33.23 | 34.33 | 32.59 | 31.55 | 29.96 | 31.96 | 32.77 |
| | 0.9240 | 0.9107 | 0.9235 | 0.9720 | 0.9107 | 0.9729 | 0.9234 | 0.9183 | 0.9422 | 0.9007 | 0.8764 | 0.8748 | 0.9208 |
| OURS | 32.56 | **32.10** | 35.80 | 36.69 | 30.93 | **32.99** | 33.33 | 34.46 | **32.78** | **31.83** | 30.03 | **32.03** | **32.96** |
| | **0.9266** | **0.9141** | 0.9229 | 0.9741 | **0.9162** | **0.9745** | 0.9245 | 0.9203 | **0.9436** | 0.9072 | **0.8860** | 0.8764 | **0.9239** |
| $\sigma$ = 30 | | | | | | | | | | | | | |
| BM3D | 28.49 | 27.97 | 32.75 | 31.97 | 26.74 | 27.81 | 29.46 | 30.95 | 28.36 | 27.65 | 26.41 | 28.66 | 28.94 |
| | 0.8642 | 0.8204 | 0.8779 | 0.9371 | 0.8256 | 0.9254 | 0.8590 | 0.8701 | 0.8808 | 0.8217 | 0.7492 | 0.8167 | 0.8540 |
| PGPD | 28.63 | 28.11 | 32.83 | 31.99 | 26.90 | 27.99 | 29.60 | 30.96 | 28.49 | 27.67 | 26.31 | 28.70 | 29.02 |
| | 0.8646 | 0.8213 | 0.8818 | 0.9317 | 0.8294 | 0.9300 | 0.8622 | 0.8606 | 0.8853 | 0.8277 | 0.7400 | 0.8164 | 0.8542 |
| WNNM | 28.75 | 28.34 | **33.23** | 32.50 | 27.02 | 28.61 | 29.72 | 31.07 | 28.91 | 28.07 | **26.66** | **28.84** | 29.31 |
| | 0.8698 | 0.8318 | 0.8892 | 0.9438 | 0.8355 | 0.9389 | 0.8670 | 0.8643 | 0.8926 | 0.8357 | **0.7615** | 0.8201 | 0.8625 |
| NCSR | 28.40 | 27.58 | 32.66 | 32.85 | 26.65 | 28.24 | 29.35 | 30.71 | 28.59 | 27.77 | 26.37 | 28.64 | 28.99 |
| | 0.8473 | 0.7704 | 0.8853 | 0.9468 | 0.7902 | 0.9377 | 0.8583 | 0.8669 | 0.8890 | 0.8304 | 0.7492 | 0.8153 | 0.8489 |
| RRC | 28.63 | 28.12 | 33.27 | 32.33 | 26.89 | 28.35 | 29.67 | 30.96 | 28.79 | 27.95 | 26.33 | 28.67 | 29.16 |
| | 0.8716 | 0.8240 | 0.8952 | 0.9482 | 0.8323 | 0.9366 | 0.8672 | **0.8703** | 0.8954 | 0.8304 | 0.7374 | 0.8184 | 0.8606 |
| LGSR | 28.76 | 28.30 | 33.36 | 32.32 | 27.05 | 28.48 | 29.78 | 30.96 | 28.87 | 28.02 | 26.58 | 28.77 | 29.27 |
| | 0.8749 | 0.8316 | **0.8960** | **0.9491** | 0.8378 | 0.9386 | **0.8718** | 0.8663 | 0.8952 | 0.8348 | 0.7541 | 0.8204 | 0.8642 |
| GSRC-NLP | 28.68 | 28.21 | 33.15 | 32.28 | 26.89 | 28.56 | 29.66 | 30.92 | 28.80 | 28.02 | 26.41 | 28.71 | 29.19 |
| | 0.8726 | 0.8262 | 0.8941 | 0.9482 | 0.8303 | 0.9401 | 0.8682 | 0.8647 | 0.8939 | 0.8313 | 0.7383 | 0.8186 | 0.8605 |
| OURS | **28.85** | **28.38** | 33.09 | **32.63** | **27.10** | **28.90** | **29.79** | **31.13** | **28.97** | **28.23** | 26.50 | 28.82 | **29.37** |
| | **0.8767** | **0.8332** | 0.8912 | 0.9470 | **0.8418** | **0.9431** | 0.8692 | 0.8680 | **0.8957** | **0.8405** | 0.7551 | **0.8218** | **0.8653** |
| $\sigma$ = 50 | | | | | | | | | | | | | |
| BM3D | 25.76 | 25.49 | 30.36 | 29.26 | 24.29 | 24.68 | 26.90 | 28.71 | 25.82 | 25.04 | 24.21 | 26.17 | 26.39 |
| | 0.7967 | 0.7311 | 0.8396 | 0.9038 | 0.7381 | 0.8639 | 0.7938 | 0.8200 | 0.8197 | 0.7377 | 0.6282 | 0.7548 | 0.7856 |
| PGPD | 25.98 | 25.63 | 30.45 | 29.20 | 24.49 | 25.03 | 27.15 | 28.79 | 26.00 | 25.11 | 24.17 | 26.31 | 26.53 |
| | 0.8059 | 0.7324 | 0.8410 | 0.8934 | 0.7483 | 0.8794 | 0.7990 | 0.8118 | 0.8269 | 0.7457 | 0.6206 | 0.7578 | 0.7885 |
| WNNM | 26.18 | 25.93 | 30.98 | 29.63 | 24.56 | 25.47 | 27.27 | 28.74 | 26.32 | 25.43 | 24.47 | 26.41 | 26.78 |
| | 0.8133 | 0.7502 | 0.8548 | 0.9098 | 0.7567 | 0.8926 | 0.8074 | 0.8138 | 0.8350 | 0.7596 | 0.6418 | 0.7630 | 0.7998 |

*(Continued)*

TABLE 1 (Continued)

| Image | Airplane | Flower | Foreman | J.Bean | Lake | Leaves | Lena | Lin | Monarch | Starfish | Pentagon | Peppers | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NCSR | 25.63 | 25.31 | 30.41 | 29.24 | 24.15 | 24.94 | 26.94 | 28.23 | 25.73 | 25.06 | 23.92 | 26.04 | 26.30 |
| | 0.8066 | 0.7217 | 0.8559 | 0.9134 | 0.7420 | 0.8787 | 0.8009 | 0.8171 | 0.8252 | 0.7440 | 0.6058 | 0.7567 | 0.7890 |
| RRC | 26.13 | 25.72 | 30.87 | 29.38 | 24.48 | 25.30 | 27.17 | 28.51 | 26.22 | 25.34 | 24.21 | 26.23 | 26.63 |
| | 0.8171 | 0.7413 | 0.8611 | 0.9125 | 0.7571 | 0.8910 | 0.8073 | 0.8140 | 0.8361 | 0.7589 | 0.6162 | 0.7643 | 0.7981 |
| LGSR | 26.15 | 25.92 | 31.03 | 29.40 | 24.59 | 25.39 | 27.27 | 28.56 | 26.24 | 25.40 | 24.47 | 26.37 | 26.73 |
| | **0.8212** | **0.7544** | **0.8637** | 0.9141 | 0.7629 | 0.8930 | **0.8140** | 0.8171 | 0.8364 | 0.7616 | 0.6423 | **0.7655** | 0.8039 |
| GSRC-NLP | 26.17 | 25.76 | 30.77 | 29.58 | 24.44 | 25.66 | 27.06 | 28.60 | 26.25 | 25.36 | 24.24 | 26.32 | 26.69 |
| | 0.8201 | 0.7416 | 0.8610 | **0.9166** | 0.7492 | 0.8991 | 0.8014 | 0.8153 | 0.8297 | 0.7540 | 0.6125 | 0.7633 | 0.7970 |
| OURS | **26.23** | **26.02** | **31.08** | **29.67** | **24.64** | **25.79** | **27.34** | **28.82** | **26.39** | **25.59** | **24.49** | **26.44** | **26.87** |
| | 0.8209 | 0.7530 | 0.8605 | 0.9067 | **0.7631** | **0.8991** | 0.8110 | **0.8188** | **0.8369** | **0.7663** | **0.6473** | **0.7665** | **0.8042** |
| $\sigma$ = 75 | | | | | | | | | | | | | |
| BM3D | 23.99 | 23.82 | 28.07 | 27.22 | 22.63 | 22.49 | 25.17 | 26.96 | 23.91 | 23.27 | 22.59 | 24.43 | 24.55 |
| | 0.7331 | 0.6515 | 0.7880 | 0.8613 | 0.6636 | 0.8021 | 0.7310 | 0.7704 | 0.7557 | 0.6619 | 0.5240 | 0.6973 | 0.7200 |
| PGPD | 24.15 | 23.82 | 28.39 | 27.07 | 22.76 | 22.61 | 25.30 | 27.05 | 24.00 | 23.23 | 22.55 | 24.46 | 24.62 |
| | 0.7492 | 0.6468 | 0.7965 | 0.8503 | 0.6760 | 0.8121 | 0.7356 | 0.7669 | 0.7642 | 0.6638 | 0.5145 | 0.7026 | 0.7232 |
| WNNM | 24.25 | 24.07 | 28.95 | 27.42 | 22.76 | 23.06 | 25.52 | 26.91 | 24.31 | 22.84 | 24.45 | 23.47 | 24.84 |
| | 0.7601 | 0.6697 | 0.8133 | 0.8707 | 0.6850 | 0.8351 | 0.7514 | 0.7717 | 0.7754 | 0.5412 | 0.7035 | 0.6801 | 0.7381 |
| NCSR | 23.76 | 23.50 | 28.18 | 27.15 | 22.48 | 22.60 | 25.02 | 26.22 | 23.67 | 23.18 | 22.10 | 24.19 | 24.34 |
| | 0.7547 | 0.6409 | 0.8171 | 0.8792 | 0.6743 | 0.8234 | 0.7415 | 0.7730 | 0.7648 | 0.6685 | 0.4881 | 0.7073 | 0.7277 |
| RRC | 24.10 | 23.77 | 28.83 | 27.17 | 22.64 | 22.91 | 25.33 | 26.86 | 24.24 | 23.32 | 22.56 | 24.35 | 24.67 |
| | 0.7638 | 0.6499 | 0.8259 | 0.8749 | 0.6822 | 0.8377 | 0.7498 | 0.7729 | 0.7782 | 0.6741 | 0.5028 | 0.7172 | 0.7358 |
| LGSR | 24.25 | 24.14 | 29.10 | 27.37 | 22.74 | 23.09 | **25.55** | 26.97 | 24.31 | 23.43 | **22.91** | 24.56 | 24.87 |
| | 0.7709 | **0.6772** | **0.8296** | 0.8828 | 0.6836 | 0.8410 | **0.7577** | **0.7839** | 0.7794 | 0.6805 | **0.5354** | 0.7190 | 0.7451 |
| GSRC-NLP | 24.13 | 23.88 | 28.76 | 27.29 | 22.61 | 23.33 | 25.32 | 26.84 | 24.35 | 23.32 | 22.65 | 24.45 | 24.74 |
| | 0.7671 | 0.6614 | 0.8251 | 0.8796 | 0.6772 | 0.8512 | 0.7480 | 0.7806 | 0.7779 | 0.6712 | 0.5146 | 0.7179 | 0.7393 |
| OURS | **24.32** | **24.19** | **29.11** | **27.62** | **22.80** | **23.49** | 25.51 | **27.24** | **24.48** | **23.56** | 22.65 | **24.64** | **24.97** |
| | **0.7721** | 0.6677 | 0.8273 | **0.8851** | **0.6916** | **0.8514** | 0.7545 | 0.7873 | **0.7807** | **0.6859** | 0.5269 | **0.7231** | **0.7461** |

The data marked in red represent the best values.

problem. Hence, after setting step size $\gamma$ and gradient direction $q$, we employ the gradient descent method (Ruder, 2016): $\hat{\boldsymbol{x}} = \boldsymbol{x} - \gamma \boldsymbol{q}$ to rewrite Eq.(19) as:

$$\hat{\boldsymbol{x}} = \boldsymbol{x} - \gamma \left( \frac{1}{\sigma_n^2} \left( \boldsymbol{H}^T \boldsymbol{H} \boldsymbol{x} - \boldsymbol{H}^T \boldsymbol{y} \right) + \frac{1}{\zeta} (\boldsymbol{x} - \boldsymbol{s} - \boldsymbol{c}) \right) \tag{19}$$

In addition, it is recommended to compute $\boldsymbol{H}^T\boldsymbol{H}$ and $\boldsymbol{H}^T\boldsymbol{y}$ in advance to further enhance the algorithm efficiency.

## 3.3. Parameter settings

In the model we proposed above, there are four parameters $(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\zeta})$ requiring setting. Here, we set a strategy for the parameters that can be automatically adjusted in each iteration, which allows us to achieve more robust and accurate experimental results.

The noise standard deviation $\boldsymbol{\sigma_n}$ is automatically updated in each iteration (Osher et al., 2005):

$$\boldsymbol{\sigma}_{\boldsymbol{e}}^{\boldsymbol{t}} = \omega \sqrt{\boldsymbol{\sigma}_{\boldsymbol{n}}^2 - \left\| \boldsymbol{y} - \hat{\boldsymbol{x}}^{(\boldsymbol{t})} \right\|_2^2} \tag{20}$$

Where $\omega$ represents a scaling factor, it is evident from Gu et al. (2014) and Chen et al. (2015) that this approach to regularize $\boldsymbol{\sigma_e}$ has been implemented in diverse models and has exhibited positive performance.

After setting $\boldsymbol{\sigma_e}$, the value of $\mu$ is tuned to change in proportion to $\boldsymbol{\sigma_e^2}$ (Zha et al., 2022):

$$\mu = \rho \left( \boldsymbol{\sigma}_{\boldsymbol{e}}^{\boldsymbol{2}} \right)^t \tag{21}$$

where $\rho$ denotes a constant.

Moreover, the regularization parameters $\lambda$ and $\theta$ represent the constraint penalties on sparsity and LR, respectively. Inspired by Dong et al. (2012a), they are adjusted in each iteration as follows:

$$\lambda^{(t)} = \frac{2\sqrt{2}\alpha\left(\sigma_e^t\right)^2}{\boldsymbol{m_i} + \varepsilon} \qquad \theta^{(t)} = \frac{2\sqrt{2}\beta\left(\sigma_e^t\right)^2}{\boldsymbol{n_i} + \epsilon} \tag{22}$$

where $\boldsymbol{m_i}$ is the estimated standard variance of $\boldsymbol{R_i}$ and $\boldsymbol{n_i}$ stands for the estimated standard variance of $\boldsymbol{\Delta_i}$. The $\varepsilon$ and $\epsilon$ are two small constants to avoid zero divisors. $\alpha$ and $\beta$ are set to two constants. Finally, parameter $\zeta$ is also set to a fixed constant.

The detailed procedure of the MS-GSRC algorithm is presented in Algorithm 1.

## 4. Experiences

In this chapter, extensive trial are conducted on image denoising, inpainting, and CS to verify that our proposed MS-GSRC model possesses better image restoration capabilities compared to some classical methods. To obtain intuitive comparison results, we set on two metrics: peak signal-to-noise ratio (PSNR) and structural self-similarity (SSIM) (Wang et al., 2004).

PSNR is commonly used to measure signal distortion. This parameter is calculated based on the gray scale values of the image pixels. Although sometimes the value of PSNR is not consistent with competent human perception, it remains an important reference evaluation metric. SSIM is a metric intended for assessing similarity between two images, which is an intuitive human standard for evaluating image quality.

If the degraded image is in color, we mainly recover the luminance channel due to the fact that variations in the luminance of color images are more easily perceived by the human eye.

The codes for all comparison algorithms used in this study are obtained from the original author's homepage and uses the given default parameters directly. For reasons of limited space, only a few images frequently used for testing are detailed list in Figure 2. In all tables, the data marked in red represent the best values.

### 4.1. Image denoising

First, we verify the performance of our MS-GSRC model on the image denoising task. The corresponding parameters are set as follows. We set the search window $W \times W$ to $30 \times 30$, the patch size $\sqrt{m} \times \sqrt{m}$ to $6 \times 6, 7 \times 7, 9 \times 9$ for $\sigma \leq 15$, $15 < \sigma \leq 30$, and $30 < \sigma \leq 75$, with the number of neighbor patches $k$ to 70, 110, 120 for $\sigma \leq 30$, $30 < \sigma \leq 50$, $50 < \sigma \leq 75$, respectively. The parameters $(\alpha, \beta, \omega, \zeta)$ are set to $(0.03, 1.75, 0.81, 0.085)$, $(0.015, 1.8, 0.86, 0.07)$, $(0.05, 2.2, 0.81, 0.12)$, $(0.006, 2, 0.86, 0.05)$ for $\sigma \leq 15$, $15 < \sigma \leq 30$, $30 < \sigma \leq 50$, $50 < \sigma \leq 75$. In addition, we set the multi-scale to [1,0.8], [1,0.85], and [1,0.9] for $\sigma \leq 15$, $15 < \sigma \leq 50$, and $50 < \sigma \leq 75$, separately.

Our MS-GSRC method is compared with several recently proposed popular denoising methods and classical traditional denoising methods, including BM3D (Dabov et al., 2007), PGPD (Xu et al., 2015), WNNM (Gu et al., 2014), NCSR (Dong et al., 2012b), RRC (Zha et al., 2019), LGSR (Zha et al., 2022) and GSRC-NLP (Zha et al., 2020a). Of all the comparison methods, BM3D is a frequently adopted benchmarking method, NCSR, PGPD, and GSRC-NLP all use GSR as a prior, and WNNM and RRC exploit low-rankness knowledge. And LGSR combines GSR and LR. Besides, both GSRC-NLP and our proposed model use the GSRC framework. Taking 12 frequently used images as an example, Table 1 lists the PSNR and SSIM results for various denoising methods at different noise levels. It is observed that our proposed MS-GSRC method produced superior performance. Specifically, the average PSNR and SSIM we achieve are improved by (0.47 dB, 0.0149) compared to BM3D, (0.38 dB, 0.0107) compared to PGPD, (0.07 dB, 0.0032) compared to WWNM, (0.42 dB, 0.0149) compared to NCSR, (0.25 dB, 0.0066) compared to RRC, (0.1 dB, 0.0005) compared to LGSR, and (0.19 dB, 0.0054) compared to GSRC-NLP.

TABLE 2 PSNR (dB) and SSIM comparison of different methods for image denoising on BSD68 dataset.

| $\sigma$ | 15 | 30 | 50 | 75 | Average |
|---|---|---|---|---|---|
| BM3D | 31.08 | 27.76 | 25.62 | 24.21 | 27.17 |
| | 0.8722 | 0.7732 | 0.6869 | 0.6221 | 0.7386 |
| PGPD | 31.14 | 27.81 | 25.75 | 24.30 | 27.25 |
| | 0.8697 | 0.7698 | 0.6873 | 0.6214 | 0.7370 |
| WNNM | 31.32 | 27.97 | 25.86 | **24.39** | 27.39 |
| | 0.8766 | 0.7802 | 0.6983 | 0.6348 | 0.7475 |
| NCSR | 31.18 | 27.78 | 25.57 | 24.04 | 27.14 |
| | 0.8769 | 0.7771 | 0.6858 | 0.6209 | 0.7402 |
| RRC | 31.07 | 27.74 | 25.67 | 24.18 | 27.17 |
| | 0.8644 | 0.7643 | 0.6840 | 0.6117 | 0.7311 |
| LGSR | 31.37 | 27.99 | 25.86 | 24.35 | 27.39 |
| | 0.8817 | 0.7862 | 0.7025 | 0.6347 | 0.7512 |
| GSRC-NLP | 31.15 | 27.74 | 25.66 | 24.15 | 27.18 |
| | 0.8681 | 0.7646 | 0.6835 | 0.6217 | 0.7345 |
| OURS | **31.38** | **28.01** | **25.88** | 24.38 | **27.41** |
| | **0.8827** | **0.7889** | **0.7042** | **0.6400** | **0.7539** |

The data marked in red represent the best values.

We also utilize the BSD68 dataset (Wang et al., 2004) to assess the denoising ability of all compared approaches. We can observe from Table 2 that the average PSNR gains obtained by our proposed MS-GSRC method in comparison to the BM3D, PGPD, WNNM, NCSR, RRC, GSRC-NLP, and LGSR methods are 0.24 dB, 0.16 dB, 0.02 dB, 0.27 dB, 0.24 dB, 0.23 dB, and 0.03 dB. Meanwhile, on average, the proposed MS-GSRC achieve an SSIM improvement of 0.0153 on BM3D, 0.0169 on PGPD, 0.0064 on WNNM, 0.0137 on NCSR, 0.0228 on RRC, 0.0027 on LGSR, and 0.0194 on GSRC-NLP. Evidently, our proposed MS-GSRC method yields better PSNR and SSIM in almost all noise cases. Our method is only 0.01 dB lower than WWNM in PSNR, but 0.0052 higher than in SSIM at $\sigma = 75$. Beyond objective metrics, the subjective perception of the human body is also a crucial criterion for assessing the quality of an image. Consequently, we present the visual contrast between the two images of starfish and 223,061

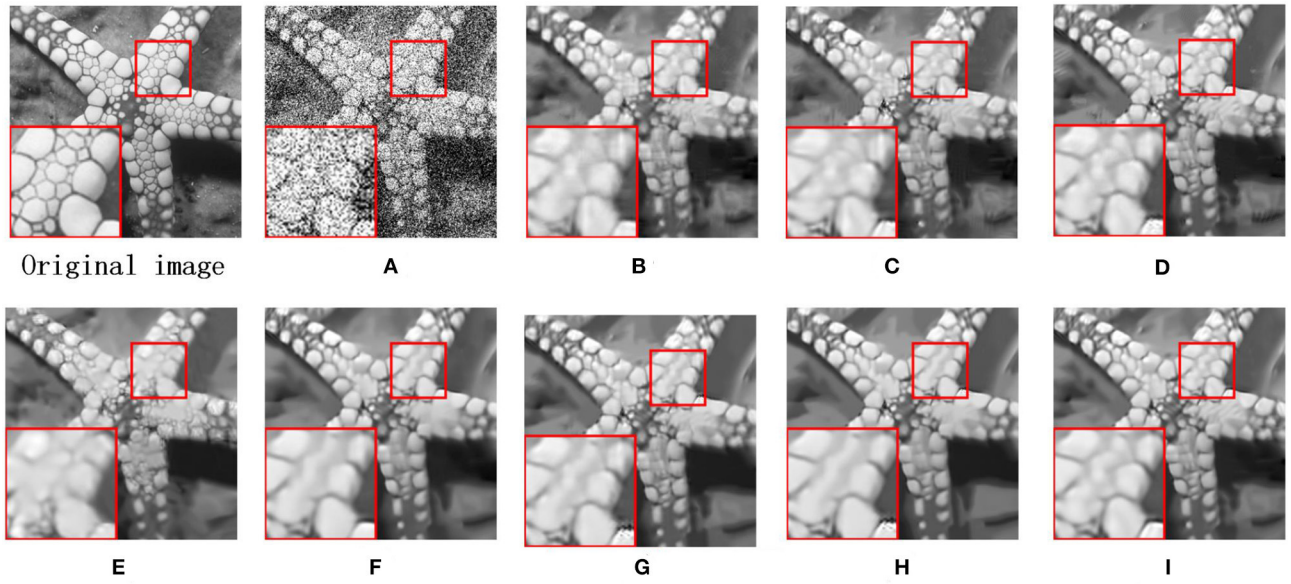



FIGURE 3
Denosing results on image starfish ($\sigma$ = 75). **(A)** Noise image. **(B)** BM3D (PSNR = 23.27 dB and SSIM = 0.6619). **(C)** PGPD (PSNR = 23.23 dB and SSIM = 0.6638). **(D)** WNNM (PSNR = 22.84 dB and SSIM = 0.5412). **(E)** NSRC (PSNR = 23.18 dB and SSIM = 0.6685). **(F)** RRC (PSNR = 23.32 dB and SSIM = 0.6741). **(G)** LGSR (PSNR = 23.43 dB and SSIM = 0.6805). **(H)** GSRC-NLP (PSNR = 23.32 dB and SSIM = 0.6712). **(I)** OURS (PSNR = 23.56 dB and SSIM = 0.6859).

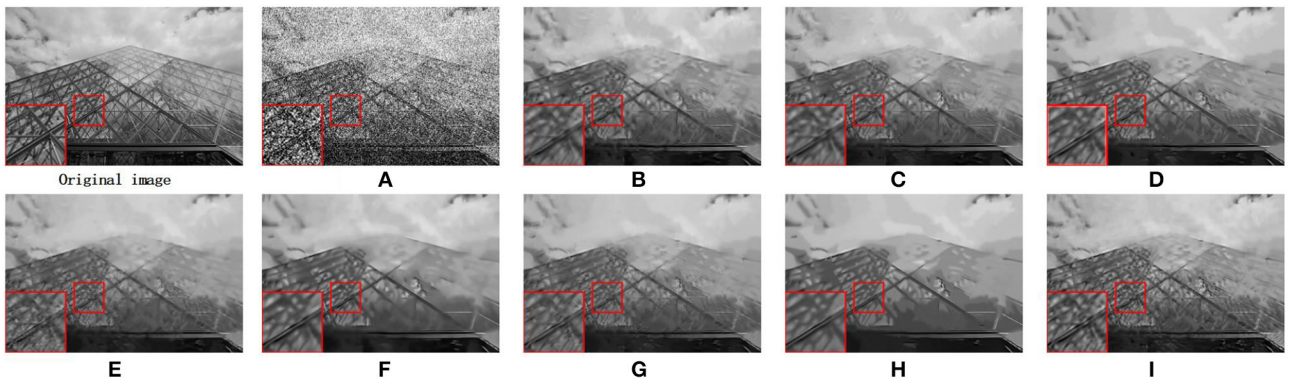



FIGURE 4
Denosing results on image 223061 ($\sigma$ = 75). **(A)** Noise image. **(B)** BM3D (PSNR = 22.27 dB and SSIM = 0.5470). **(C)** PGPD (PSNR = 22.30 dB and SSIM = 0.5420). **(D)** WNNM (PSNR = 22.51 dB and SSIM = 0.5690). **(E)** NSRC (PSNR = 22.15 dB and SSIM = 0.5383). **(F)** RRC (PSNR = 22.22 dB and SSIM = 0.5351). **(G)** LGSR (PSNR = 22.32 dB and SSIM = 0.5545). **(H)** GSRC-NLP (PSNR = 22.13 dB and SSIM = 0.5313). **(I)** OURS (PSNR = 22.42 dB and SSIM = 0.5761).

restored by different methods in Figures 3, 4, respectively. Figure 3 indicates that BM3D, PGPD, WNNM, and RRC are likely to over-smooth the restored image, whereas NCSR, GSRC-NLP, and LGSR can lead to the appearance of some undesired visual artifacts. As can be seen in Figure 4, although the image restored by WNNM has a higher PSNR, the image restored by our MS-GSRC method has a higher SSIM value and presents a better visual effect. PGPD, NCSR, RRC, and GSRC-NLP are susceptible to loss of detail in the restored images, while BM3D, WNNM, and LGSR may result in undesirable artifacts.

## 4.2. Image inpainting

Next, we verify the superiority of the MS-GSRC model on inpainting. We likewise compare the proposed MS-GSRC method with many classical or recently popular methods, such as SAIST (Afonso et al., 2010), TSLRA (Guo et al., 2017), GSR (Zhang et al., 2014b), JSM (Zhang et al., 2014c), JPG-SR (Zha et al., 2018b), LGSR (Zha et al., 2022), and IDBP (Tirer and Giryes, 2018). Among these, SAIST is one of the earliest proposed methods for image restoration, GSR, JPG-SR, LGSR, TSLRA,

TABLE 3 PSNR (dB) and SSIM comparison of different methods SAIST, TSLRA, GSR, JSM, JPG-SR, LGSR, IDBP, and OURS for image inpainting.

| Images | Bahoon | Bear | House | Lake | Leaves | Lena | Lily | Pepper | Nanna | Butterfly | Gilrs | Fireman | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pixels missing = 80% | | | | | | | | | | | | | |
| SALSA | 23.15 | 27.29 | 26.63 | 22.20 | 19.78 | 25.96 | 24.31 | 25.55 | 21.96 | 19.95 | 21.80 | 22.17 | 23.40 |
| | 0.5815 | 0.7952 | 0.8421 | 0.7420 | 0.7749 | 0.8294 | 0.7485 | 0.8633 | 0.7288 | 0.7883 | 0.7078 | 0.6812 | 0.7569 |
| JSM | 25.21 | 29.35 | 34.28 | 25.57 | 26.17 | 30.50 | 27.92 | 30.26 | 25.16 | 25.38 | 25.07 | 25.25 | 27.51 |
| | 0.6577 | 0.8378 | 0.9102 | 0.8302 | 0.9209 | 0.8991 | 0.8410 | 0.9214 | 0.8196 | 0.9011 | 0.8015 | 0.7664 | 0.8423 |
| GSR | 24.58 | 30.28 | **35.57** | 25.67 | 27.46 | 31.42 | 28.87 | 31.10 | 25.23 | 26.03 | 25.50 | 25.46 | 28.10 |
| | 0.6893 | 0.8650 | 0.9313 | 0.8560 | **0.9452** | 0.9250 | 0.8820 | 0.9393 | 0.8531 | 0.9223 | 0.8386 | 0.8041 | 0.8709 |
| TSLRA | **25.44** | 29.34 | 31.30 | 25.31 | 25.09 | 30.09 | 27.96 | 28.39 | 25.32 | 24.91 | 24.99 | 25.44 | 26.96 |
| | 0.6714 | 0.8401 | 0.9106 | 0.8103 | 0.8934 | 0.8904 | 0.8400 | 0.9087 | 0.8163 | 0.8835 | 0.7974 | 0.7759 | 0.8365 |
| JPG-SR | 24.99 | 30.15 | 34.92 | 25.93 | 27.42 | 31.46 | 28.97 | 31.23 | 25.66 | 26.29 | 25.60 | 25.48 | 28.18 |
| | 0.6904 | 0.8562 | 0.9148 | 0.8508 | 0.9409 | 0.9193 | 0.8767 | 0.9326 | 0.8500 | 0.9214 | 0.8373 | 0.7977 | 0.8657 |
| LGSR | 25.24 | 30.55 | 35.83 | 26.33 | 27.48 | 31.69 | 29.07 | 31.75 | 25.91 | 26.53 | 25.81 | 25.79 | 28.50 |
| | 0.6989 | 0.8678 | **0.9333** | 0.8611 | 0.9419 | 0.9251 | 0.8813 | 0.9383 | 0.8541 | 0.9244 | 0.8423 | 0.8078 | 0.8730 |
| IDBP | 25.03 | 30.06 | 33.69 | 25.84 | 26.48 | 30.29 | 28.10 | 30.89 | 25.42 | 25.60 | 25.48 | 25.46 | 27.70 |
| | 0.6695 | 0.8447 | 0.9060 | 0.8319 | 0.9233 | 0.8979 | 0.8486 | 0.9153 | 0.8214 | 0.9011 | 0.8146 | 0.7645 | 0.8449 |
| OURS | 25.32 | **30.62** | 35.55 | **26.38** | **27.60** | **31.91** | **29.20** | **31.97** | **26.06** | **26.78** | **25.97** | **26.04** | **28.62** |
| | **0.7006** | 0.8694 | **0.9246** | **0.8619** | **0.9436** | **0.9267** | **0.8840** | **0.9405** | **0.8564** | **0.9278** | **0.8461** | **0.8125** | **0.8745** |
| Pixels missing = 70% | | | | | | | | | | | | | |
| SALSA | 24.32 | 29.29 | 27.49 | 24.33 | 22.01 | 28.10 | 26.20 | 28.40 | 23.93 | 22.41 | 23.53 | 23.96 | 25.33 |
| | 0.6867 | 0.8542 | 0.8827 | 0.8325 | 0.8572 | 0.8864 | 0.8278 | 0.9159 | 0.8179 | 0.8669 | 0.7962 | 0.7703 | 0.8329 |
| JSM | 26.48 | 31.56 | 36.69 | 27.56 | 29.28 | 32.67 | 29.74 | 33.28 | 27.19 | 27.84 | 27.18 | 27.07 | 29.71 |
| | 0.7514 | 0.8895 | 0.9402 | 0.8854 | 0.9581 | 0.9351 | 0.8924 | 0.9535 | 0.8819 | 0.9374 | 0.8739 | 0.8385 | 0.8948 |
| GSR | 26.17 | 32.01 | 37.63 | 28.08 | 31.18 | 33.54 | 31.10 | 34.77 | 27.89 | 28.92 | 27.86 | 27.47 | 30.55 |
| | 0.7797 | 0.9043 | 0.9543 | 0.9057 | **0.9744** | 0.9507 | 0.9246 | **0.9633** | 0.9076 | 0.9506 | 0.9015 | 0.8681 | 0.9154 |
| TSLRA | 26.71 | 31.65 | 35.86 | 27.30 | 27.94 | 32.58 | 29.91 | 32.64 | 27.27 | 27.74 | 27.05 | 27.23 | 29.49 |
| | 0.7602 | 0.8917 | 0.9485 | 0.8770 | 0.9440 | 0.9355 | 0.8942 | 0.9494 | 0.8808 | 0.9342 | 0.8668 | 0.8412 | 0.8936 |
| JPG-SR | 26.38 | 32.21 | 37.41 | 28.04 | 30.89 | 33.58 | 31.12 | 34.49 | 27.95 | 29.18 | 27.91 | 27.54 | 30.56 |
| | 0.7774 | 0.8997 | 0.9445 | 0.9011 | 0.9707 | 0.9469 | 0.9197 | 0.9580 | 0.9036 | 0.9494 | 0.8982 | 0.8624 | 0.9110 |
| LGSR | 26.65 | 32.28 | **37.98** | 28.72 | 31.31 | 33.76 | 31.19 | 34.92 | 28.21 | 29.39 | 28.14 | 27.90 | 30.87 |
| | 0.7846 | 0.9065 | **0.9555** | 0.9097 | 0.9729 | 0.9507 | 0.9237 | 0.9619 | 0.9065 | 0.9523 | 0.9025 | 0.8707 | 0.9165 |
| IDBP | 26.39 | 31.74 | 36.48 | 27.92 | 29.23 | 32.58 | 30.08 | 33.36 | 27.16 | 28.25 | 27.49 | 27.37 | 29.84 |
| | 0.7582 | 0.8872 | 0.9293 | 0.8856 | 0.9549 | 0.9340 | 0.8974 | 0.9460 | 0.8767 | 0.9387 | 0.8766 | 0.8391 | 0.8936 |
| OURS | **26.76** | **32.38** | 37.97 | **28.77** | **31.57** | **33.88** | **31.53** | **35.11** | **28.42** | **29.62** | **28.35** | **28.13** | **31.04** |
| | **0.7867** | **0.9078** | 0.9541 | **0.9102** | **0.9744** | **0.9515** | **0.9284** | 0.9632 | **0.9090** | **0.9542** | **0.9053** | **0.8743** | **0.9183** |
| Pixels missing = 60% | | | | | | | | | | | | | |
| SALSA | 25.40 | 29.73 | 29.99 | 25.84 | 24.65 | 29.69 | 28.11 | 30.60 | 25.37 | 25.28 | 25.06 | 25.37 | 27.09 |
| | 0.7648 | 0.8880 | 0.9096 | 0.8772 | 0.9192 | 0.9203 | 0.8848 | 0.9443 | 0.8688 | 0.9186 | 0.8536 | 0.8349 | 0.8820 |
| JSM | 27.71 | 33.07 | 38.53 | 29.35 | 31.43 | 34.60 | 31.56 | 35.35 | 29.06 | 29.77 | 28.96 | 28.72 | 31.51 |
| | 0.8175 | 0.9182 | 0.9580 | 0.9213 | 0.9748 | 0.9559 | 0.9278 | 0.9678 | 0.9182 | 0.9567 | 0.9151 | 0.8871 | 0.9265 |
| GSR | 27.74 | 33.60 | 39.68 | 29.86 | 33.39 | 35.81 | 33.05 | 36.42 | 30.13 | 31.09 | 29.55 | 29.32 | 32.47 |

*(Continued)*

TABLE 3 (Continued)

| Images | Bahoon | Bear | House | Lake | Leaves | Lena | Lily | Pepper | Nanna | Butterfly | Gilrs | Fireman | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.8445 | 0.9298 | 0.9674 | 0.9366 | 0.9849 | 0.9668 | 0.9505 | 0.9739 | 0.9383 | 0.9667 | 0.9359 | 0.9086 | 0.9420 |
| TSLRA | 27.92 | 32.77 | 37.23 | 29.01 | 30.19 | 34.26 | 31.55 | 34.96 | 29.17 | 29.42 | 28.79 | 28.73 | 31.17 |
| | 0.8239 | 0.9195 | 0.9641 | 0.9156 | 0.9666 | 0.9555 | 0.9282 | 0.9654 | 0.9173 | 0.9531 | 0.9097 | 0.8860 | 0.9254 |
| JPG-SR | 27.92 | 33.61 | 39.22 | 30.13 | 33.26 | 35.73 | 33.10 | 36.40 | 30.21 | 31.30 | 29.84 | 29.46 | 32.52 |
| | 0.8404 | 0.9240 | 0.9594 | 0.9328 | 0.9829 | 0.9626 | 0.9464 | 0.9692 | 0.9350 | 0.9641 | 0.9326 | 0.9039 | 0.9378 |
| LGSR | 28.15 | 33.94 | 39.82 | 30.66 | 33.70 | 35.97 | 33.31 | 36.85 | 30.40 | 31.58 | 30.13 | 29.83 | 32.86 |
| | 0.8481 | 0.9320 | **0.9678** | 0.9396 | 0.9848 | 0.9665 | 0.9506 | 0.9732 | 0.9381 | 0.9673 | 0.9373 | 0.9119 | 0.9431 |
| IDBP | 27.71 | 33.53 | 38.18 | 29.76 | 31.55 | 34.35 | 31.85 | 35.27 | 29.22 | 29.71 | 29.24 | 28.98 | 31.61 |
| | 0.8226 | 0.9176 | 0.9487 | 0.9209 | 0.9728 | 0.9531 | 0.9301 | 0.9628 | 0.9184 | 0.9544 | 0.9153 | 0.8839 | 0.9251 |
| OURS | **28.23** | **34.10** | **39.90** | **30.71** | **34.11** | **36.08** | **33.54** | **36.99** | **30.55** | **31.84** | **30.43** | **30.05** | **33.04** |
| | **0.8481** | **0.9331** | 0.9676 | **0.9402** | **0.9861** | **0.9672** | **0.9528** | **0.9740** | **0.9396** | **0.9684** | **0.9395** | **0.9142** | **0.9442** |
| Pixels missing = 50% | | | | | | | | | | | | | |
| SALSA | 26.50 | 31.79 | 31.64 | 27.83 | 26.61 | 30.98 | 29.59 | 31.08 | 26.85 | 27.28 | 26.90 | 27.09 | 28.68 |
| | 0.8270 | 0.9226 | 0.9326 | 0.9176 | 0.9471 | 0.9436 | 0.9181 | 0.9595 | 0.9062 | 0.9452 | 0.8992 | 0.8826 | 0.9168 |
| JSM | 29.05 | 34.63 | 40.43 | 30.99 | 33.80 | 36.37 | 33.41 | 37.32 | 30.73 | 31.35 | 30.63 | 30.27 | 33.25 |
| | 0.8697 | 0.9415 | 0.9710 | 0.9447 | 0.9848 | 0.9705 | 0.9523 | 0.9773 | 0.9440 | 0.9692 | 0.9433 | 0.9196 | 0.9490 |
| GSR | 29.41 | 35.62 | 41.62 | 32.14 | 35.87 | 37.63 | 35.41 | 38.53 | 32.16 | 32.78 | 31.93 | 31.00 | 34.51 |
| | 0.8923 | 0.9509 | 0.9768 | 0.9575 | 0.9909 | 0.9779 | 0.9685 | **0.9817** | 0.9589 | 0.9759 | 0.9582 | 0.9353 | 0.9604 |
| TSLRA | 29.15 | 33.01 | 40.22 | 30.53 | 32.56 | 35.52 | 33.20 | 36.61 | 30.87 | 31.01 | 30.48 | 30.25 | 32.79 |
| | 0.8734 | 0.9407 | 0.9748 | 0.9409 | 0.9803 | 0.9702 | 0.9518 | 0.9758 | 0.9433 | 0.9672 | 0.9397 | 0.9186 | 0.9480 |
| JPG-SR | 29.49 | 35.53 | 40.85 | 31.89 | 35.83 | 37.39 | 35.21 | 38.19 | 32.27 | 32.89 | 32.02 | 30.96 | 34.38 |
| | 0.8887 | 0.9454 | 0.9704 | 0.9533 | 0.9896 | 0.9732 | 0.9647 | 0.9771 | 0.9558 | 0.9737 | 0.9556 | 0.9310 | 0.9565 |
| LGSR | 29.74 | 35.89 | 41.78 | 32.57 | 36.35 | 37.89 | 35.41 | 38.59 | 32.50 | 33.38 | 32.19 | 31.42 | 34.81 |
| | 0.8950 | 0.9524 | **0.9772** | 0.9592 | 0.9910 | 0.9775 | 0.9684 | 0.9810 | 0.9592 | 0.9771 | 0.9592 | 0.9379 | 0.9613 |
| IDBP | 29.14 | 34.85 | 40.20 | 31.51 | 34.05 | 36.36 | 33.66 | 37.60 | 30.86 | 31.99 | 31.11 | 30.53 | 33.49 |
| | 0.8726 | 0.9383 | 0.9653 | 0.9447 | 0.9836 | 0.9668 | 0.9523 | 0.9738 | 0.9420 | 0.9686 | 0.9427 | 0.9170 | 0.9473 |
| OURS | **29.80** | **35.99** | **41.80** | **32.58** | **36.60** | **38.07** | **35.63** | **38.89** | **32.53** | **33.44** | **32.38** | **31.58** | **34.94** |
| | **0.8950** | **0.9530** | 0.9769 | **0.9595** | **0.9915** | **0.9780** | **0.9694** | **0.9817** | **0.9597** | **0.9775** | **0.9604** | **0.9395** | **0.9618** |

The data marked in red represent the best values.

and JSM use the NSS prior, and IDBP is a deep learning-based method. In simulation experiments, we test images by randomly generated masks that included missing pixels of 80%, 70%, 60%, and 50%. Following are the parameters that we set for the MS-GSRC model in different cases. We set the patch size to $7 \times 7$, the search window size to 25, and the non-local similar patches to 60. In addition, for all cases, we set the multi-scales to [1,0.85]. Moreover, we set $(0.0002, 0.0001, 1.5, 15)$ and $(0.0001, 0.0001, 1.5, 15)$ as parameters $(\omega, \zeta, \alpha, \beta)$ when the missing pixels are 0.8 and others, respectively. In addition, $\sigma = \sqrt{2}$ for all experiences.

Table 3 illustrates the PSNR and SSIM results for each method on the 12 frequently used test images. As observed in Table 3, our proposed method exceeds the comparison algorithm virtually often when it comes to image inpainting performance. The proposed MS-GSRC outperforms SAIST, JSM, GSR, TSLRA, JPG-SR, LGSR, and IDBP approaches in average PSNR performance, with gains of 5.8 dB, 1.43 dB, 0.51 dB, 1.82 dB, 0.51 dB, 0.13 dB, and 1.26 dB, respectively. Additionally, on average, the proposed MS-GSRC surpasses SAIST by 0.0776, JSM by 0.0216, GSR by 0.0025, TSLRA by 0.0238, JPG-SR by 0.007, LGSR by 0.0012, and IDBP by 0.022.

Similarly, two images are selected for detailed visual analysis. The image butterfly with a 80% loss of pixels restored by different methods are presented in Figure 5. Moreover, Figure 6 displays the outcomes of a visual comparison of image flowers with a 70% loss of pixels restored with different algorithms. By analyzing the visual comparison images, we can find that images restored using SAIST, JSM, TSLRA, IDBP GSR, and JPG-SR are susceptible to excessive smoothing, and images restored using LGSR tend to show excessive visual artifacts. The images restored using our proposed MS-GSRC model have significantly better restoration capabilities with regard to image detail and edges.

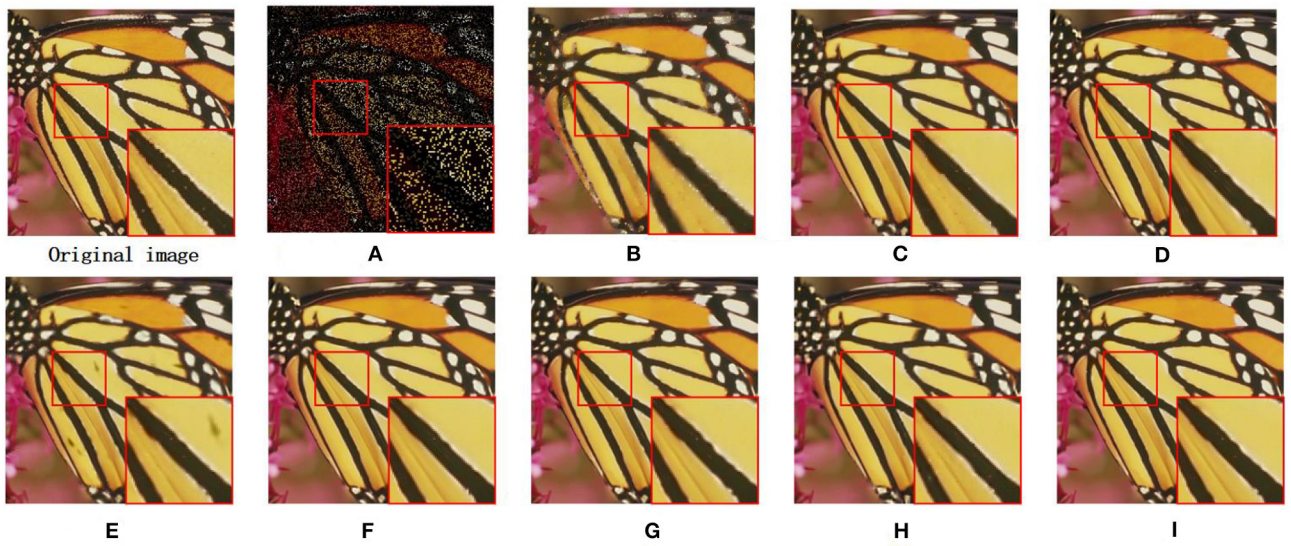



FIGURE 5
Inpainting results on image butterfly (missing ratio=80%). **(A)** Missing pixels image. **(B)** SAIST(PSNR = 19.95 dB and SSIM = 0.7883). **(C)** JSM (PSNR = 25.38 dB and SSIM = 0.9011). **(D)** GSR (PSNR = 26.03 dB and SSIM = 0.9223). **(E)** TSLRA (PSNR = 24.91 dB and SSIM = 0.8835). **(F)** JPG-SR (PSNR = 26.29 dB and SSIM = 0.9214). **(G)** LGSR (PSNR = 26.53 dB and SSIM = 0.9244). **(H)** IDBP (PSNR = 25.60 dB and SSIM = 0.9011). **(I)** OURS(PSNR = 26.78 dB and SSIM = 0.9278).

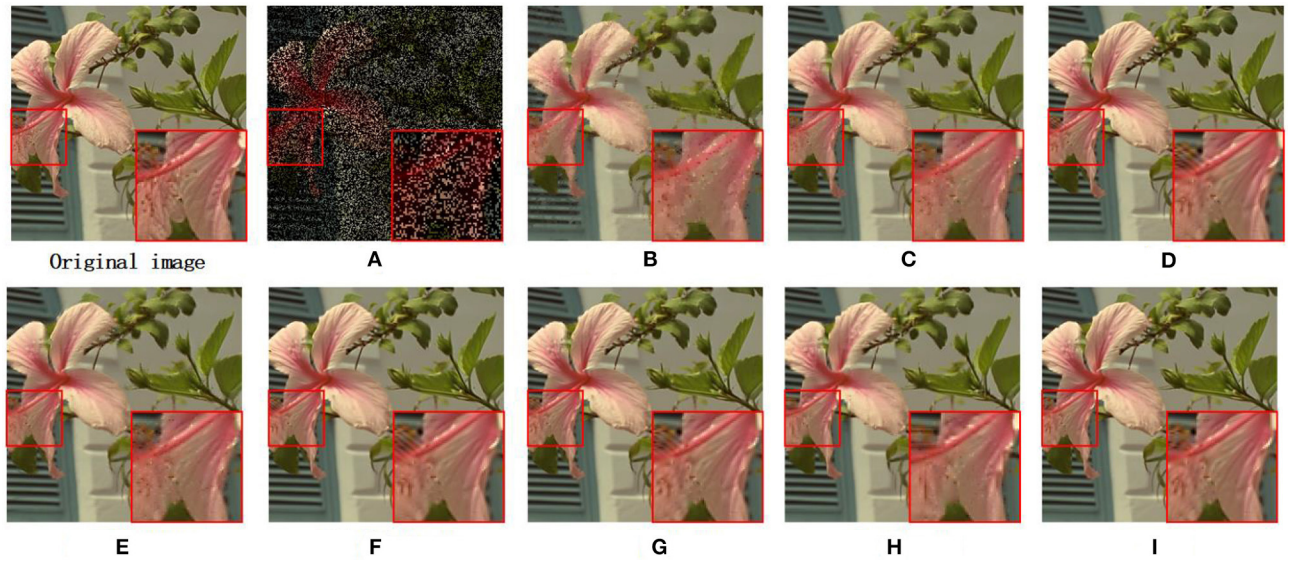



FIGURE 6
Inpainting results on image flowers (missing ratio = 70%). **(A)** Missing pixels image. **(B)** SAIST (PSNR = 27.69 dB and SSIM = 0.8422). **(C)** JSM (PSNR = 29.74 dB and SSIM =0.8924). **(D)** GSR (PSNR = 31.10 dB and SSIM = 0.9246). **(E)** TSLRA (PSNR = 29.91 dB and SSIM = 0.8942). **(F)** JPG-SR (PSNR = 31.12 dB and SSIM = 0.9197). **(G)** LGSR (PSNR = 31.19 dB and SSIM = 0.9237). **(H)** IDBP (PSNR = 30.08 dB and SSIM = 0.8974). **(I)** OURS (PSNR = 31.53 dB and SSIM = 0.9284).

## 4.3. Image compressed sensing

Finally, we validate the restoration capability of our proposed MS-GSRC model on the image compressed sensing problem. In this part of experiments, we use the Gaussian random projection matrix (Zhang et al., 2014b) to generate blocks of size 32 × 32 to test the CS restoration effects. The parameters set for the MS-GSRC model are as follows: For all cases, the patch size is set to be 8 × 8, the patch number to 80, the search window size to be 25, and the multi-scales to be [1,0.75]. In addition, $(0.004, 0.00002, 0.6, 25)$, $(0.0014, 0.00005, 0.9, 15)$, $(0.0015, 0.00001, 0.5, 10)$, and $(0.0015, 0.00001, 1.4, 6)$ are set for $(\zeta, \omega, \alpha, \beta)$ when subrate is 0.1N, 0.2N, 0.3N, and 0.4N.

BSC (Mun and Fowler, 2009), RCOS (Zhang et al., 2012), ALSB (Zhang et al., 2014a), GSR (Zhang et al., 2014b), ASNR (Zha et al., 2018a), and LGSR (Zha et al., 2022) are choosen as competing

methods. Among them, GSR performs a sparse representation on similar groups of images, ASNR is an image of the CS method that extends on the basis of NCSR, and LGSR combines sparsity and LR. Similarly, we selected 12 images frequently used in image restoration experiments as test images. Table 4 presents the average outcomes of PSNR and SSIM of the restored images using different method. To be concrete, the proposed MS-GSRC model over BCS, RCOS, ALSB, GSR, ASNR, and LGSR methods are 5.36 dB, 2.34 dB, 1.24 dB, 0.46 dB, 0.45 dB, and 0.14d B in PSNR and 0.1591, 0.0634, 0.0023, 0.0056, 0.0082, and 0.0029 in SSIM, respectively.

TABLE 4 PSNR (dB) and SSIM comparison of different methods for image CS on 12 test images.

| Subrate | 0.1 | 0.2 | 0.3 | 0.4 | Average |
|---|---|---|---|---|---|
| BCS | 23.60 | 26.26 | 28.19 | 29.88 | 26.98 |
| | 0.6308 | 0.7418 | 0.8117 | 0.8609 | 0.7445 |
| RCOS | 25.92 | 29.20 | 31.54 | 33.34 | 30.00 |
| | 0.7163 | 0.8298 | 0.8909 | 0.9236 | 0.8402 |
| ALSB | 26.66 | 30.19 | 32.67 | 34.87 | 31.10 |
| | 0.7778 | 0.8751 | 0.9209 | 0.9484 | 0.8806 |
| GSR | 27.00 | 30.96 | 33.66 | 35.89 | 31.88 |
| | 0.8002 | 0.8963 | 0.9367 | 0.9587 | 0.8980 |
| ASNR | 27.24 | 31.04 | 33.51 | 35.78 | 31.89 |
| | 0.7965 | 0.8953 | 0.9329 | 0.9568 | 0.8954 |
| LGSR | 27.51 | 31.34 | 33.89 | 36.07 | 32.20 |
| | 0.8062 | 0.8994 | 0.9379 | 0.9593 | 0.9007 |
| OURS | **27.91** | **31.40** | **33.94** | **36.09** | **32.34** |
| | **0.8150** | **0.9009** | **0.9387** | **0.9598** | **0.9036** |

The data marked in red represent the best values.

Due to the other competing algorithms used in this thesis, all use BCS to pre-process CS images, and here we use the BCS-processed images as corrupted images. Figure 7 shows the visual contrast of the image fence with 0.1 N CS measurements, and we can observe that RCOS and ALSB are less capable of restoring details, GSR and LGSR lead to over-smooth, and ASNR generates some redundant artifacts. Figure 8 illustrates the visual comparison of the image leaves measured with 0.1N CS. All comparison images have strong ringing phenomena and present terrible artifacts. In Figure 9, we have selected the image airplane processed with 0.2N CS for detailed analysis. It is obvious that the details of the images restored by ALSB and LGSR are seriously missing. The images restored by RCOS, GSR, and ASNR produced more artifacts. In the above three cases, our proposed MS-GSRC algorithm significantly outperforms other competing algorithms in recovering the image overall and some texture details.

## 5. Conclusion

In this study, we propose a novel model Multi-Scale Group Sparse Residual Constraint Model (MS-GSRC) for image restoration. This model introduces the low-rank property into the group sparse residual framework and finds similar patches for overlapping patches of the input image using a multi-scale strategy. Furthermore, under the MAP restoration framework, an alternating minimization method with adaptive tunable parameters is used to deliver a robust optimization solution for our

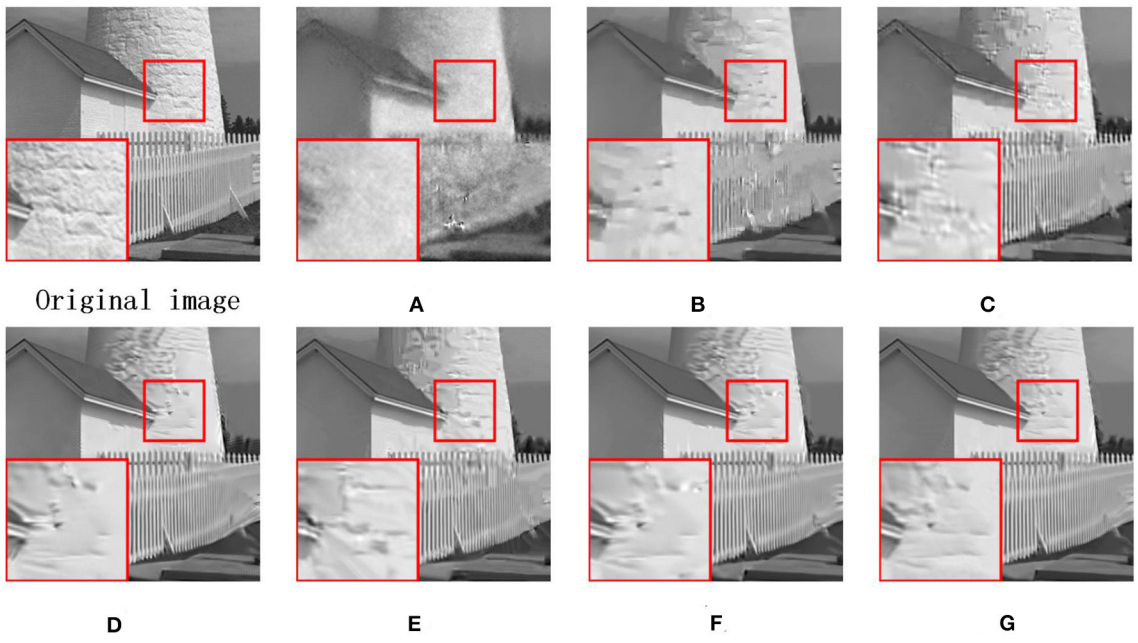



FIGURE 7
CS results on image fence (subrate = 0.1 N). **(A)** BCS (PSNR = 19.54 dB, SSIM = 0.5034). **(B)** RCOS (PSNR = 23.29 dB, SSIM = 0.6932). **(C)** ALSB (PSNR = 25.05 dB and SSIM = 0.7736). **(D)** GSR (PSNR = 26.06 dB and SSIM = 0.8047). **(E)** ASNR (PSNR = 26.01 dB and SSIM = 0.8006). **(F)** LGSR (PSNR = 26.58 dB and SSIM = 0.8107). **(G)** OURS (PSNR = 27.26 dB and SSIM = 0.8216).

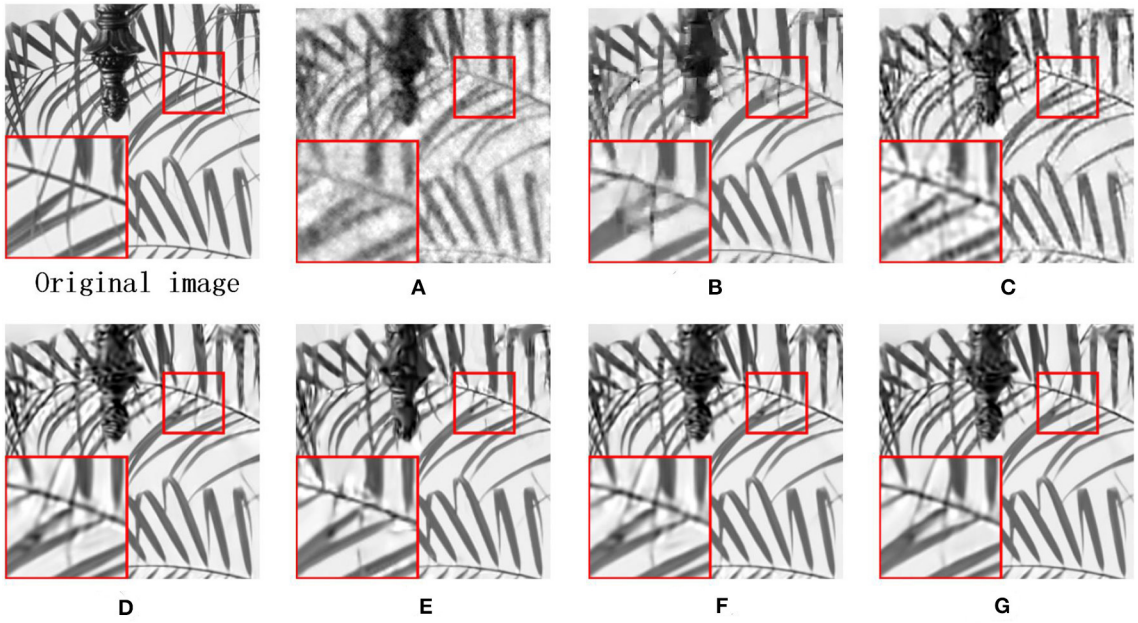



FIGURE 8
CS results on image leaves (subrate = 0.1 N). **(A)** BCS (PSNR = 18.37 dB, SSIM = 0.5767). **(B)** RCOS (PSNR = 22.17 dB, SSIM = 0.0.8323). **(C)** ALSB (PSNR = 21.52 dB and SSIM = 0.7939). **(D)** GSR (PSNR = 23.22 dB and SSIM = 0.8731). **(E)** ASNR (PSNR = 23.48 dB and SSIM = 0.8805). **(F)** LGSR (PSNR = 23.75 dB and SSIM = 0.8824). **(G)** OURS (PSNR = 24.57 dB and SSIM = 0.8992).

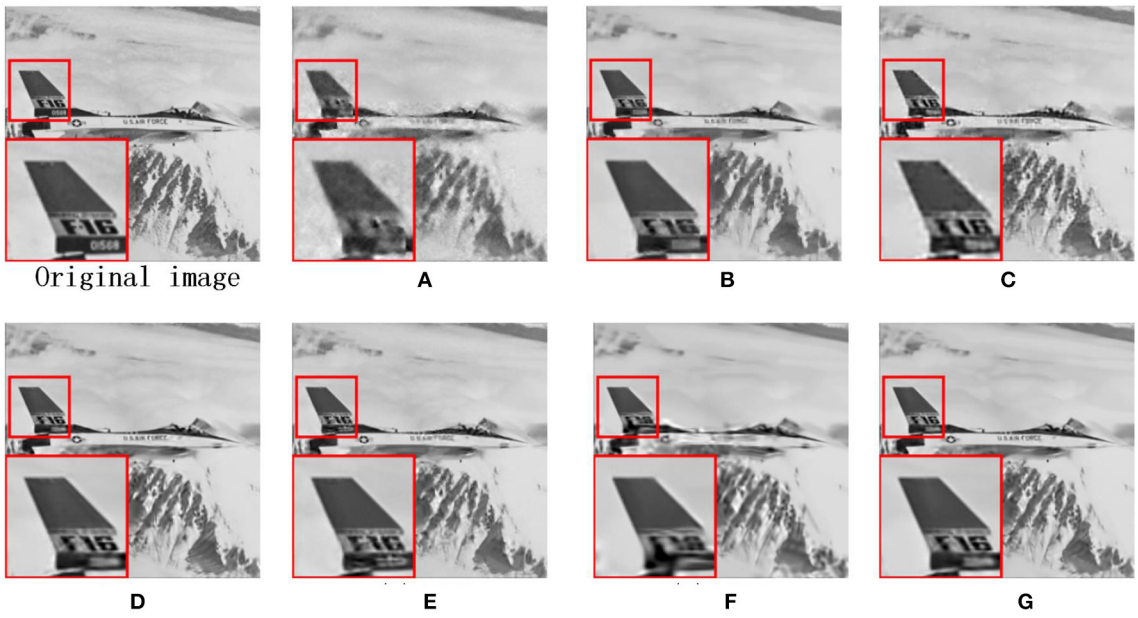



FIGURE 9
CS results on image airplane (subrate = 0.2 N). **(A)** BCS (PSNR = 25.87 dB, SSIM = 0.8111). **(B)** RCOS (PSNR = 28.22 dB, SSIM = 0.8854). **(C)** ALSB (PSNR = 28.39 dB and SSIM = 0.8942). **(D)** GSR (PSNR = 28.87 dB and SSIM = 0.9082). **(E)** ASNR (PSNR = 29.17 dB and SSIM = 0.9075). **(F)** LGSR (PSNR = 29.43 dB and SSIM = 0.9110). **(G)** OURS (PSNR = 29.59 dB and SSIM = 0.9120).

MS-GSRC method. We employ the MS-GSRC model to three image restoration problems, namely, denoising, inpainting, and compressed sensing. Extensive simulation trials show that our novel model performs superior to many classical methods in terms of both objective image quality and subjective visual quality.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Author contributions

WN: Writing — original draft. DS: Writing — review & editing. QG: Writing — review & editing. YL: Writing — review & editing. DZ: Writing — review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (NSFC) (62071001), the Nature Science Foundation of Anhui (2008085MF192, 2008085MF183, 2208085QF206, and 2308085QF224), the Key Science Program of Anhui Education Department (KJ2021A0013), and was also supported by the China Postdoctoral Science Foundation (2023M730009).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* 2, 433–459. doi: 10.1002/wics.101

Afonso, M. V., Bioucas-Dias, J. M., and Figueiredo, M. A. (2010). An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *IEEE Trans. Image Proc.* 20, 681–695. doi: 10.1109/TIP.2010.2076294

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Buades, A., Coll, B., and Morel, J.-M. (2005). "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. San Diego, CA: IEEE, 60-65.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optimizat.* 20, 1956–1982. doi: 10.1137/080738970

Chen, F., Zhang, L., and Yu, H. (2015). "External patch prior guided internal clustering for image denoising," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago, CA: IEEE), 603–611.

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.* 16, 2080–2095. doi: 10.1109/TIP.2007.901238

Dey, N., Blanc-Feraud, L., Zimmer, C., Roux, P., Kam, Z., Olivo-Marin, J.-C., et al. (2006). Richardson-lucy algorithm with total variation regularization for 3d confocal microscope deconvolution. *Microsc. Res. Tech.* 69, 260–266. doi: 10.1002/jemt.20294

Dong, W., Shi, G., and Li, X. (2012a). Nonlocal image restoration with bilateral variance estimation: a low-rank approach. *IEEE Trans. Image Process.* 22, 700–711. doi: 10.1109/TIP.2012.2221729

Dong, W., Zhang, L., Shi, G., and Li, X. (2012b). Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.* 22, 1620–1630. doi: 10.1109/TIP.2012.2235847

Elad, M., and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.* 15, 3736–3745. doi: 10.1109/TIP.2006.881969

Fazel, M., Hindi, H., and Boyd, S. P. (2001). "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*. Arlington, VA: IEEE, 4734–4739.

Gu, S., Zhang, L., Zuo, W., and Feng, X. (2014). "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2862–2869.

Guo, Q., Gao, S., Zhang, X., Yin, Y., and Zhang, C. (2017). Patch-based image inpainting via two-stage low rank approximation. *IEEE Trans. Vis. Comput. Graph.* 24, 2023–2036. doi: 10.1109/TVCG.2017.2702738

Hu, Y., Zhang, D., Ye, J., Li, X., and He, X. (2012). Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2117–2130. doi: 10.1109/TPAMI.2012.271

Jon, K., Sun, Y., Li, Q., Liu, J., Wang, X., and Zhu, W. (2021). Image restoration using overlapping group sparsity on hyper-laplacian prior of image gradient. *Neurocomputing* 420, 57–69. doi: 10.1016/j.neucom.2020.08.053

Keller, J. M., Gray, M. R., and Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. IEEE transactions on systems, man, and cybernetics. *IEEE Trans. Syst. Man. Cybern.* 4, 580–585. doi: 10.1109/TSMC.1985.6313426

Keshavarzian, R., Aghagolzadeh, A., and Rezaii, T. Y. (2019). Llp norm regularization based group sparse representation for image compressed sensing recovery. *Signal Proc.: Image Commun.* 78, 477–493. doi: 10.1016/j.image.2019.07.021

Liu, S., Wu, G., Liu, H., and Zhang, X. (2017). Image restoration approach using a joint sparse representation in 3d-transform domain. *Digital Signal Proc.* 60, 307–323. doi: 10.1016/j.dsp.2016.10.008

Lu, C., Shi, J., and Jia, J. (2013). "Online robust dictionary learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 415–422.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2009). "Non-local sparse models for image restoration," in *2009 IEEE 12th International Conference on Computer Vision*. Kyoto: IEEE, 2272–2279.

Mun, S., and Fowler, J. E. (2009). "Block compressed sensing of images using directional transforms," in *2009 16th IEEE International Conference on Image Processing (ICIP)*. Cairo: IEEE, 3021–3024.

Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W. (2005). An iterative regularization method for total variation-based image restoration. *Multiscale Model. Simul.* 4, 460–489. doi: 10.1137/040605412

Ou, Y., Luo, J., Li, B., and Swamy, M. S. (2020). Gray-level image denoising with an improved weighted sparse coding. *J. Vis. Commun. Image Represent.* 72, 102895. doi: 10.1016/j.jvcir.2020.102895

Ou, Y., Swamy, M., Luo, J., and Li, B. (2022a). Single image denoising via multi-scale weighted group sparse coding. *Signal Proc.* 200, 108650. doi: 10.1016/j.sigpro.2022.108650

Ou, Y., Zhang, B., and Li, B. (2022b). Multi-scale low-rank approximation method for image denoising. *Multimedia Tools Applicat.* 81, 20357–20371. doi: 10.1007/s11042-022-12083-z

Papyan, V., and Elad, M. (2015). Multi-scale patch-based image restoration. *IEEE Trans. Image Process.* 25, 249–261. doi: 10.1109/TIP.2015.2499698

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv.* abs/1609.04747. doi: 10.48550/arXiv.1609.04747

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenom.* 60, 259–268. doi: 10.1016/0167-2789(92)90242-F

Srebro, N., and Jaakkola, T. (2003). "Weighted low-rank approximations," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (Washington, DC: AAAI Press), 720–727.

Tirer, T., and Giryes, R. (2018). Image restoration by iterative denoising and backward projections. *IEEE Trans. Image Process.* 28, 1220–1234. doi: 10.1109/TIP.2018.2875569

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Xie, Y., Gu, S., Liu, Y., Zuo, W., Zhang, W., and Zhang, L. (2016). Weighted schatten *p*-norm minimization for image denoising and background subtraction. *IEEE Trans. Image Process.* 25, 4842–4857. doi: 10.1109/TIP.2016.2599290

Xu, J., Zhang, L., Zuo, W., Zhang, D., and Feng, X. (2015). "Patch group based nonlocal self-similarity prior learning for image denoising," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago, CA: IEEE), 244–252.

Yair, N., and Michaeli, T. (2018). "Multi-scale weighted nuclear norm image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3165–3174.

Zha, Z., Liu, X., Zhang, X., Chen, Y., Tang, L., Bai, Y., et al. (2018a). Compressed sensing image reconstruction via adaptive sparse nonlocal regularization. *Visual Comp.* 34, 117–137. doi: 10.1007/s00371-016-1318-9

Zha, Z., Liu, X., Zhou, Z., Huang, X., Shi, J., Shang, Z., et al. (2017). "Image denoising via group sparsity residual constraint," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, 1787–1791.

Zha, Z., Wen, B., Yuan, X., Zhou, J., Zhu, C., and Kot, A. C. (2022). Low-rankness guided group sparse representation for image restoration. *IEEE Trans. Neural. Netw. Learn Syst.* 34, 7593–7607. doi: 10.1109/TNNLS.2022.3144630

Zha, Z., Yuan, X., Wen, B., Zhou, J., Zhang, J., and Zhu, C. (2019). From rank estimation to rank approximation: Rank residual constraint for image restoration. *IEEE Trans. Image Process.* 29, 3254–3269. doi: 10.1109/TIP.2019.2958309

Zha, Z., Yuan, X., Wen, B., Zhou, J., and Zhu, C. (2018b). "Joint patch-group based sparse representation for image inpainting," in *Asian Conference on Machine Learning*, 145–160.

Zha, Z., Yuan, X., Wen, B., Zhou, J., and Zhu, C. (2020a). Group sparsity residual constraint with non-local priors for image restoration. *IEEE Trans. Image Process.* 29, 8960–8975. doi: 10.1109/TIP.2020.3021291

Zha, Z., Yuan, X., Zhou, J., Zhu, C., and Wen, B. (2020b). Image restoration via simultaneous nonlocal self-similarity priors. *IEEE Trans. Image Process.* 29, 8561–8576. doi: 10.1109/TIP.2020.3015545

Zhang, J., and Ghanem, B. (2018). "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 1828–1837.

Zhang, J., Zhao, C., Zhao, D., and Gao, W. (2014a). Image compressive sensing recovery using adaptively learned sparsifying basis via l0 minimization. *Signal Proc.* 103, 114–126. doi: 10.1016/j.sigpro.2013.09.025

Zhang, J., Zhao, D., and Gao, W. (2014b). Group-based sparse representation for image restoration. *IEEE Trans. Image Process.* 23, 3336–3351. doi: 10.1109/TIP.2014.2323127

Zhang, J., Zhao, D., Xiong, R., Ma, S., and Gao, W. (2014c). Image restoration using joint statistical modeling in a space-transform domain. *IEEE Trans. Circuits Syst. Video Technol.* IEEE. 24, 915–928. doi: 10.1109/TCSVT.2014.2302380

Zhang, J., Zhao, D., Zhao, C., Xiong, R., Ma, S., and Gao, W. (2012). Image compressive sensing recovery via collaborative sparsity. *IEEE J. Emerg. Sel. Topics Power Electron.* 2, 380–391. doi: 10.1109/JETCAS.2012.2220391

Zhang, K., Zuo, W., Gu, S., and Zhang, L. (2017). "Learning deep cnn denoiser prior for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 3929–3938.

Zoran, D., and Weiss, Y. (2011). "From learning models of natural image patches to whole image restoration," in *2011 International Conference on Computer Vision*. Barcelona: IEEE, 479–486.

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 13 November 2023
DOI 10.3389/fnins.2023.1288908



OPEN ACCESS

EDITED BY
Teng Li,
Anhui University, China

REVIEWED BY
Chengcheng Ren,
Anhui University, China
Yunliang Chen,
China University of Geosciences Wuhan, China
Mengchao Ma,
Hefei University of Technology, China

*CORRESPONDENCE
Zhifeng Chen
✉ zf982873139@163.com
Wei Fu
✉ lukeyoyo@tom.com

RECEIVED 05 September 2023
ACCEPTED 12 October 2023
PUBLISHED 13 November 2023

CITATION
Xu J, Chen Z and Fu W (2023) Research on product detection and recognition methods for intelligent vending machines.
*Front. Neurosci.* 17:1288908.
doi: 10.3389/fnins.2023.1288908



# Research on product detection and recognition methods for intelligent vending machines

Jianqiao Xu[1], Zhifeng Chen[2]* and Wei Fu[1]*

[1]Department of Information Security, Naval University of Engineering, Wuhan, China, [2]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China

With the continuous development of China's economy and the improvement of residents' living standards, it also brings increasing costs of labor and rent. In addition, the impact of the pandemic on the entity industry has brought opportunities for the development of new retail models. Based on the booming development of artificial intelligence, big data, and mobile payment in the new era, the new retail industry using artificial intelligence technology has shown outstanding performance in the market. Among them, intelligent vending machines have emerged in the new retail model. In order to provide users with a good shopping experience, the product detection speed and accuracy of intelligent vending machines must be high enough. We adopt Faster R-CNN, a mature object detection algorithm in deep learning, to solve the commodity settlement scenario of intelligent vending machines.

KEYWORDS

**deep learning, computer vision, object detection, ResNet, intelligent vending machines**

## 1. Introduction

In recent years, deep learning-based computer vision methods have received extensive research attention, especially the ResNet proposed by He et al. (2016), which addressed the degradation problem caused by increasing the number of layers in neural networks, and the Faster R-CNN proposed by Ren et al. (2017), which has made significant progress in object detection. These mature and efficient artificial intelligence algorithms have been widely used in the new retail industry, such as intelligent vending machines that use computer vision algorithms discussed in this article. Compared with traditional vending machines or physical retail stores, intelligent vending machines have lower costs, more flexible types of goods sold, and higher profits to the retail industry, thus standing out in the new retail market.

The object detection of retail product checkout in intelligent vending machines faces several challenges. One challenge is that it is difficult to predict user behavior, and the products in checkout images may be stacked, placed in abnormal ways, or obscured by obstacles (such as hands). The challenges mentioned above may result in the algorithm receiving insufficient information. Therefore, it is essential to ensure that the accuracy of product detection meets the requirements in such cases. Another challenge is the detection speed of the algorithm, which is crucial for improving the user experience. This article addresses these two issues by selecting a unique dataset for training on single-target commodities from multiple angles and perspectives and verifying it on multi-target items. Meanwhile, ResNet50 is chosen as the backbone neural network of Faster R-CNN to improve feature extraction for each product's angle and enhance the overall performance and prediction speed of the model. The Faster R-CNN based on ResNet50 used in this article achieves good accuracy and acceptable response speed in the intelligent vending machine product checkout scene.

## 2. Related work

Intelligent vending machines have advantages such as flexible stocking and low costs compared to traditional vending machines. Classic vending machines have complex manufacturing processes and high prices, are limited by the structure of the vending channel, and have a specific failure rate. At present, there are three different technical solutions for intelligent vending machines, namely gravity induction (Brolin et al., 2018), radio frequency identification (RFID) (Ramzan et al., 2017), and computer vision algorithms based on deep learning. The gravity solution is just an improvement method for traditional vending machines, and it does not entirely overcome the shortcomings of conventional vending machines. Due to technical limitations, RFID cannot perform well on metal goods, meaning that canned beverages are unsuitable for RFID vending machines. Moreover, because of the technical characteristics of RFID, each item must be manually labeled with an RFID tag before being placed in the intelligent vending machine for sale; this is an additional cost that cannot be ignored for the RFID technical solution.

The fundamental technology of intelligent vending machines based on computer vision is to identify products through images captured by the camera. Many works have achieved significant success in object detection (Li et al., 2016; Nian et al., 2016; Zhang et al., 2019; Ren et al., 2020, 2022), which can be applied to product recognition. Currently, deep learning-based object detection algorithms have become mainstream. These algorithms can be divided into two main types: region proposal-based methods and single-stage methods. Region proposal-based methods generate candidate regions and then classify and regress these regions to obtain the final detection results. These methods include RCNN (Girshick et al., 2014), Fast RCNN (Girshick, 2015), Faster RCNN (Ren et al., 2017), etc. Single-stage methods directly classify and regress the image without generating candidate regions. These methods include YOLO (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020), SSD (Liu et al., 2016), RetinaNet (Lin et al., 2017), etc. In addition, object detection faces many challenges, such as occlusion, scale variation, and illumination variation. To overcome these challenges, researchers have proposed many improved algorithms. For example, Mask RCNN (He et al., 2017) adds instance segmentation functionality to Faster RCNN, allowing the model to detect and segment objects simultaneously. CenterNet (Zhou et al., 2019) is a center point-based detection algorithm that can maintain high accuracy while improving detection speed.

For the datasets of retail product checkout, Goldman (Goldman et al., 2019) assembled a dataset consisting of images of supermarket shelves. It contains 110,712 product categories, averaging 147.2 instances per image. The dataset we used, Retail Product Checkout (RPC) proposed by Wei et al. (2019), is a large-scale retail dataset that includes 83,739 images with bounding box annotations for 200 categories of products. In the PRC dataset, training images only contain a single object. In contrast, testing images may contain multiple objects and are divided into three groups: easy, medium, and hard, making it an ideal dataset for our purposes.

## 3. Product detection and recognition methods for intelligent vending machines

This section applies the Faster R-CNN to the product settlement scenario of intelligent vending machines. Figure 1 illustrates the network architecture of the Faster R-CNN based on ResNet50, which can be summarized as the RPN network + Fast R-CNN. In this network, the candidate regions for Fast R-CNN are not selected by the Selective Search algorithm (Uijlings et al., 2013) but are provided by the RPN. Additionally, the Faster R-CNN used in this paper extracts features from the input image using ResNet50 rather than VGG16.

### 3.1. Input image preprocessing

The input image resolution of the dataset used in this paper ranges from 1,750×1,750 to 1,850×1,850. High-resolution images provide more detailed information but pose challenges for training due to the large number of parameters and calculations required by the deep neural network ResNet50 used in this paper. Modern deep-learning methods commonly use GPU acceleration for training. Still, training on personal computers with limited GPU and memory resources can easily lead to memory overflow and out-of-memory errors. For example, on my personal computer with 16GB RAM and 8GB GPU memory, when the batch size is set to 3, the memory usage is up to 95% when using the Dataloader to read data, and the GPU memory overflow occurs when preparing to start training after reading the data. When the batch size is set to 2, the training time for one epoch is as long as eight hours. Therefore, we attempted to reduce the resolution of all input images from 3×438×438 to 3×463×463 before training. And when calculating the bounding box loss, the predicted coordinates of the model's bounding boxes are multiplied by four before being compared to the coordinates in the labels. This can be done because there are generally no tiny targets in the checkout scenario, so the negative impact on the model is relatively small. Through experiments, this has been shown to improve the training speed.

### 3.2. ResNet50

As shown in Figure 2, the first layer of all ResNet consists of a 7×7 convolutional layer with a stride of 2, followed by a 3×3 max pooling layer with a stride of 2. After the convolutional layer, there is a 3×3 max pooling layer with a stride of 2. The max pooling layer downsamples the feature maps output from the convolutional layer, reducing the size of the feature maps while retaining the most salient features. After passing through the common convolutional and pooling layers, all ResNet structures are followed by four residual block layers. Specifically, implementing the residual block in ResNet involves adding a shortcut connection between two convolutional layers and adding the input directly to the output of the convolutional layers.

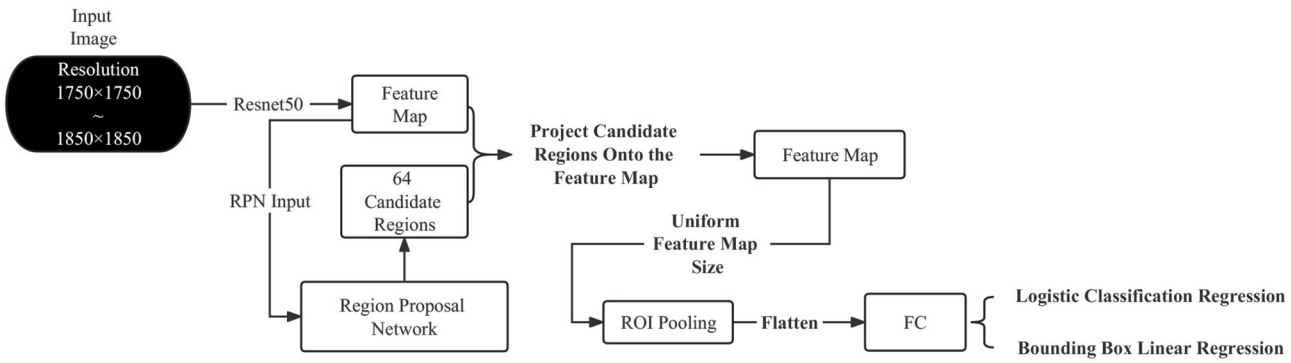



FIGURE 1
Network architecture of the faster R-CNN based on ResNet50.

| layer name | output size | 18-layer | 34-layer | 50-layer | 101-layer | 152-layer |
|---|---|---|---|---|---|---|
| conv1 | 112×112 | 7×7, 64, stride 2 | | | | |
| conv2_x | 56×56 | 3×3 max pool, stride 2 | | | | |
| | | [3×3, 64; 3×3, 64] ×2 | [3×3, 64; 3×3, 64] ×3 | [1×1, 64; 3×3, 64; 1×1, 256] ×3 | [1×1, 64; 3×3, 64; 1×1, 256] ×3 | [1×1, 64; 3×3, 64; 1×1, 256] ×3 |
| conv3_x | 28×28 | [3×3, 128; 3×3, 128] ×2 | [3×3, 128; 3×3, 128] ×4 | [1×1, 128; 3×3, 128; 1×1, 512] ×4 | [1×1, 128; 3×3, 128; 1×1, 512] ×4 | [1×1, 128; 3×3, 128; 1×1, 512] ×8 |
| conv4_x | 14×14 | [3×3, 256; 3×3, 256] ×2 | [3×3, 256; 3×3, 256] ×6 | [1×1, 256; 3×3, 256; 1×1, 1024] ×6 | [1×1, 256; 3×3, 256; 1×1, 1024] ×23 | [1×1, 256; 3×3, 256; 1×1, 1024] ×36 |
| conv5_x | 7×7 | [3×3, 512; 3×3, 512] ×2 | [3×3, 512; 3×3, 512] ×3 | [1×1, 512; 3×3, 512; 1×1, 2048] ×3 | [1×1, 512; 3×3, 512; 1×1, 2048] ×3 | [1×1, 512; 3×3, 512; 1×1, 2048] ×3 |
| | 1×1 | average pool, 1000-d fc, softmax | | | | |
| FLOPs | | $1.8\times10^9$ | $3.6\times10^9$ | $3.8\times10^9$ | $7.6\times10^9$ | $11.3\times10^9$ |

FIGURE 2
Network architecture of ResNet.

When VGG16 was used as the backbone neural network in the original Faster R-CNN paper, the number of parameters used for feature extraction was ~ 138 M, with a floating-point calculation of 30.8 G FlOPS. In contrast, ResNet50 only had about 23 M parameters and 8.2 G FlOPS floating point calculations. During training, ResNet50 had a much faster convergence rate than VGG16, making it both quick and efficient, significantly reducing training time. Additionally, ResNet50 has a larger receptive field in its feature map than VGG16 due to the multiple convolutional layers, which allows it to capture larger image contexts. A larger receptivefield is generally better in object detection tasks, as it can capture more overall features. When the receptive field is not large enough, it can cause the model to have bias errors, seriously affecting its performance. ResNet50 has a receptive lot of approximately 483, while VGG16's receptive field is only 212. Since the target pixels in the images used in this paper are mostly equal to or larger than 300×300, ResNet50 is better suited to this task than VGG16. The formula for calculating the receptive field is as follows:

$$RF_i = (RF_{i-1} - 1) * Stride_i + K_{sizei} \quad (1)$$

$RF_i$ refers to the receptive field of the i-th layer; $Stride_i$ is the stride of the i-th layer; $K_{sizei}$ is the size of the convolutional kernel used in the i-th layer.

## 3.3. Faster R-CNN

### 3.3.1. Region proposal network

In Faster R-CNN, the role of the region proposal network(RPN) is to generate region proposals, which are candidate regions that may contain objects. These region proposals are then fed into a subsequent classification network for object detection.

The RPN operates on a feature map and uses a convolutional neural network over the feature map, generating multiple anchor boxes of different sizes and aspect ratios, as shown in Figure 3.

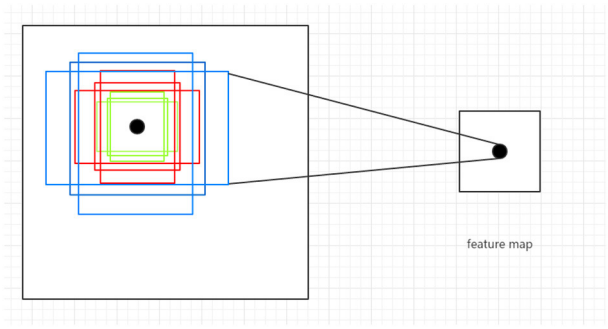



FIGURE 3
Illustration of anchor boxes.

There are three sizes of anchor boxes, which are 128, 256, and 512, and three aspect ratios, which are 1:2, 2:1, and 1:1. Based on the combinations of sizes and aspect ratios, nine different anchor boxes are generated at each point in the feature map, with their coordinates projected onto the original image as the center. For each anchor box, the RPN predicts whether it contains an object and the rough location of the object, thus generating region proposals. These region proposals can then be fed into a subsequent classification network for object detection, resulting in the final detection results. In the end, we divided the image into 9×14×14 anchor boxes (approximately 1.7k). Some of the anchor boxes we split may span across boundaries, but we ignore those that do. After removing the anchor boxes that span across boundaries, we sample 64 anchor boxes from the remaining ones, with an equal distribution of positive and negative samples, each accounting for 50%. If there are not enough positive samples to fill half of the selected samples, we can use negative samples to fill the remaining slots. Whether the IoU (Intersection over Union) value between each candidate box[1] and the ground-truth box exceeds a preset threshold is the criterion for determining positive and negative samples.

The loss function of RPN consists of two parts: classification loss and bounding box regression loss. In the classification loss function, we calculate a binary classification loss for each anchor box, representing the error of classifying it as foreground (containing an object) or background (not including an object). For each anchor box, the corresponding binary classification loss is:

$$L_{cls} = \begin{cases} -log(p) & if(y == 1) \\ -log(1-p) & else \end{cases} \quad (2)$$

Where $p$ represents the predicted probability of the anchor box being classified as foreground, $y$ represents the ture label.When $y == 1$, it represents true lable of the anchor box is foreground, and when $y == 0$, the true label is the background. In the bounding box regression loss function, we calculate a smooth L1 loss for each anchor box that is classified as foreground, which represents the difference between the predicted bounding box coordinates and the true bounding box coordinates. For each foreground anchor box, its corresponding L1 loss is:

$$smooth_{L1}(x) = \begin{cases} 0.5 * x^2 & if(x < 1) \\ |x| - 0.5 & else \end{cases} \quad (3)$$

$$L_{reg}(t^*, t) = smooth_{L1}(t_i^* - t_i) \quad (4)$$

Here, $t^*$ represents the true bounding box coordinate offset, $t$ represents the predicted bounding box coordinate offset, and $i$ represents the dimension of the coordinate axis. The $N$ represents the number of anchor boxes classified as foreground.After computing the loss functions for both components, we add them together to obtain the final RPN network loss function:

$$L_{RPN} = \frac{1}{N_{cls}} \sum_{i}^{N_{cls}} L_{cls} + \lambda \frac{1}{N_{reg}} \sum_{i}^{N_{reg}} p_i L_{reg}(t_i^*, t_i) \quad (5)$$

$p$ and $t$ denote the classification prediction and bounding box regression prediction of the RPN network, while and $t^*$ represent the true bounding box coordinate offsets. $N_{cls}$ and $N_{reg}$ correspond to the numbers of all and foreground anchor boxes, respectively. $\lambda$ is a hyperparameter that balances the classification loss and bounding box regression loss.

### 3.3.2. ROI pooling

Since the dimensions of the images are not the same, it means that the corresponding feature map sizes are also different. The purpose of ROI pooling is to unify the feature map sizes, making it easier for subsequent neural network processing. The implementation of ROI pooling involves dividing the feature map into 7×7 regions and performing max pooling within each region.The feature map image outputted by the ROI Pooling layer is a three-dimensional tensor of size 7×7×2048. We flatten it into a one-dimensional vector of size 1×100,352. Then, we concatenate these vectors in the order of their corresponding ROIs in the input image, forming a two-dimensional tensor of size 64×100,352 This two-dimensional tensor serves as the input to the fully connected layer for classification and regression tasks.

## 4. Experimental analysis

### 4.1. Retail product checkout dataset introduction

The dataset used in this article is a large-scale retail product checkout dataset publicly available on Kaggle (link: https://www.kaggle.com/datasets/diyer22/retail-product-checkout-dataset). This dataset provides rich image data of products during the checkout process and is currently the largest dataset regarding the number of images and product categories. It includes 200 common product categories in daily life, with a training set of 48,000 single-product images, a test set of 24,000 multi-target product images, and a validation set of 6,000 multi-target product images.

The training set consists of single-object images captured by four cameras placed at the top, 45 degrees upward, 30

1 The sampled 64 anchor boxes are referred to as candidate boxes.

degrees upward, and horizontally in a specified environment, covering 0–360 degrees, as shown in Figure 4. The validation and test sets are multi-object images. They are categorized into easy mode, medium mode, and hard mode based on the clutter level of the products in the images. The training set consists of single-object images captured by four cameras placed at the top, 45 degrees upward, 30 degrees upward, and horizontally, respectively, in a specified environment, covering 0–360 degrees. The validation and test sets are multi-object images and are categorized into easy mode, medium mode, and hard mode based on the clutter level of the products in the images.

The dataset validation is divided into three levels of difficulty based on the complexity of product arrangement, as shown in Figure 5.

## 4.2. Experimental parameter settings and experimental environment

The experimental environment is a personal computer with the following specifications: Processor: AMD R7-5800H; GPU: NVIDIA RTX 3070 8G; Memory: 16G. The editor used is Pycharm 2022.1; operating system: WIN11; CUDA version: 11.02; Pytorch version: 1.11.0. We used the Pytorch framework to construct our model. Before starting the training, we loaded the pre-trained parameters of ResNet50 into the model to speed up the training process. The optimizer we used is the stochastic gradient descent algorithm with a momentum value of 0.9 and set weight decay to prevent overfitting. Finally, we set a learning rate with dynamic decay. Since we trained on a personal computer with limited GPU memory, we set the batch size to 4.

## 4.3. Analysis of experimental results

In order to evaluate the model we trained, we used Pycocotools provided by the COCO official for evaluation. It provides 10 evaluation metrics including AP (Average Precision), AP (IOU = 0.5), AP (IOU = 0.75), AP (Small Area), AP (Medium Area), and AP (Large Area), AR (Average Recall), AR (Max = 1), AR (Max = 10), and AR (Max = 100). Among them, AP(IOU = 0.5) is the most commonly used metric. The experimental results are shown in Table 1. The above results indicate that using the Faster R-CNN algorithm for object detection on the Retail Product Checkout dataset can achieve good performance. The performance of the model varies under different AP metrics, with AP (IOU = 0.5) and AP (Large Area) performing well and AP (Small Area) and AP (Medium Area) performing poorly. This is because the environment of the intelligent vending machine is relatively fixed, and there are no small or medium-sized objects in the dataset, so APs and APm are close to 0. This can also be inferred from the fact that APl and AP values are always close.

## 4.4. Detection performance of the model under different difficulty levels

This section presents the model's prediction performance under different difficulty levels, and the detection of the goods is good. See Figures 6–8.

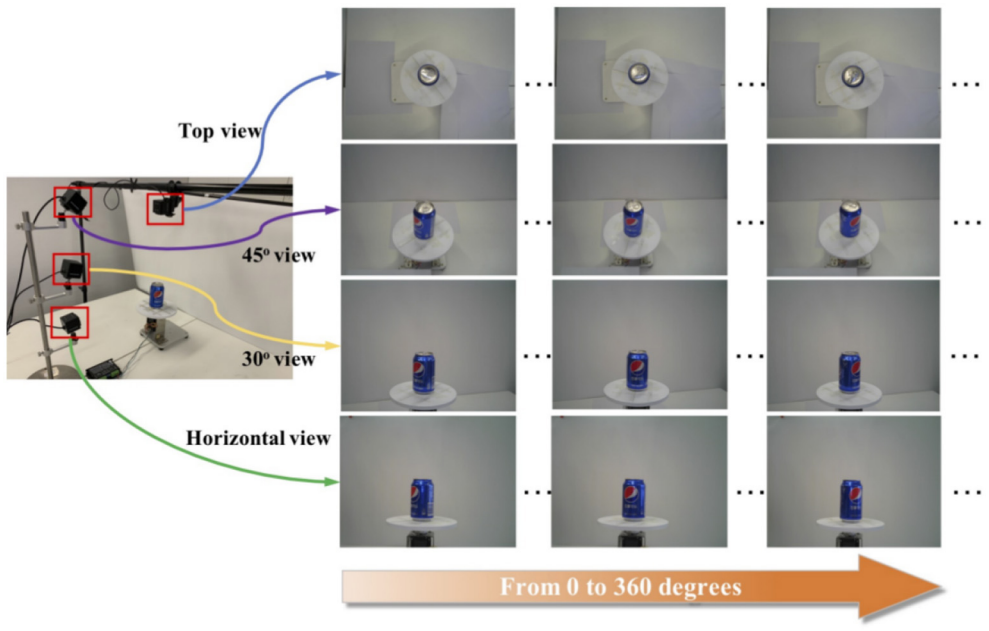



FIGURE 4
The collection form of the training set.


FIGURE 5
Three levels of validation difficulty. **(A)** Easy mode. **(B)** Medium mode. **(C)** Hard mode.

TABLE 1 **Evaluation results.**

| Epoch | Average precision | | | | | |
|---|---|---|---|---|---|---|
| | AP | AP (IOU = 0.5) | AP (IOU = 0.75) | AP (Small area) | AP (Medium area) | AP (Large area) |
| 20 | 0.539 | 0.6379 | 0.5596 | 0 | 0 | 0.5391 |
| 24 | 0.5794 | 0.6412 | 0.5784 | 0 | 0 | 0.5795 |
| 28 | 0.5818 | 0.6415 | 0.5807 | 0 | 0 | 0.5819 |
| 32 | 0.5888 | 0.6435 | 0.5825 | 0 | 0 | 0.5986 |
| 36 | 0.5962 | 0.6463 | 0.5875 | 0 | 0 | 0.5972 |
| 40 | 0.5921 | 0.6484 | 0.5866 | 0 | 0 | 0.5921 |

## 5. Conclusion

After years of development, object detection technology has made rapid progress, and there are now many mature and efficient object detection algorithms such as Faster R-CNN, YOLO, SSD, and others. In this paper, we successfully applied Faster R-CNN for object detection in the context of commodity settlement and achieved good results. Using object detection

FIGURE 6
Easy mode.

FIGURE 7
Medium mode.

FIGURE 8
Hard mode.

in computer vision as a commodity settlement recognition task for intelligent vending machines is reliable, low-cost, and efficient.

The Faster R-CNN object detection model based on ResNet50 constructed in this paper achieved good results on a large commodity dataset, with precision meeting the requirements on recognized targets and a very low probability of misclassification. However, there are still cases of missed detections in multi-object scenarios, which I believe can be improved through further training. At the same time, the model constructed in this paper has already met the recognition speed requirements for intelligent vending machines, but there is still room for improvement.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/datasets/diyer22/retail-product-checkout-dataset.

## Author contributions

JX: Conceptualization, Investigation, Software, Writing—original draft, Writing—review & editing. ZC: Writing—original draft, Writing—review & editing. WF: Writing—review & editing, Funding acquisition.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This paper was funded by the National Natural Science Foundation of China (Grant No. 62276273).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv*.

Brolin, A., Mithun, R., Gokulnath, V., and Harivishanth, M. (2018). "Design of automated medicine vending machine using mechatronics techniques," in *IOP Conference Series: Materials Science and Engineering*. Bristol: IOP Publishing, 012044.

Girshick, R. (2015). Fast r-cnn. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages* 1440–1448. doi: 10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE), 580–587. doi: 10.1109/CVPR.2014.81

Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., and Hassner, T. (2019). "Precise detection in densely packed scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE), 5227–5236.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE), 2961–2969.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Li, T., Meng, Z., Ni, B., Shen, J., and Wang, M. (2016). Robust geometric p-norm feature pooling for image classification and action recognition. *Image Vision Comp*. 55, 64–76. doi: 10.1016/j.imavis.2016.04.002

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE), 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). "SSD: Single shot multibox detector," in *Computer Vision-ECCV 2016: 14th European Conference*. Amsterdam: Springer, 21–37.

Nian, F., Li, T., Wu, X., Gao, Q., and Li, F. (2016). Efficient near-duplicate image detection with a local-based binary representation. *Multimedia Tools Appl.* 75, 2435–2452. doi: 10.1007/s11042-015-2472-1

Ramzan, A., Rehman, S., and Perwaiz, A. (2017). "RFID technology: Beyond cash-based methods in vending machine," in *2017 2nd International Conference on Control and Robotics Engineering (ICCRE)*. Bangkok: IEEE, 189–193.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 779–788.

Redmon, J., and Farhadi, A. (2017). "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263–7271.

Redmon, J., and Farhadi, A. (2018). Yolov3: an incremental improvement. *CoRR*. abs/1804.02767.

Ren, C., He, S., Luan, X., Liu, F., and Karimi, H. R. (2020). Finite-time l 2-gain asynchronous control for continuous-time positive hidden markov jump systems via t-s fuzzy model approach. *IEEE Trans. Cybernet*. 51, 77–87. doi: 10.1109/TCYB.2020.2996743

Ren, C., Park, J. H., and He, S. (2022). Positiveness and finite-time control of dual-switching poisson jump networked control systems with time-varying delays and packet drops. *IEEE Trans. Cont. Network Syst*. 9, 575–587. doi: 10.1109/TCNS.2022.3165075

Ren, S., He, K., Girshick, R., and Sun, J. (2017). "Faster r-cnn: towards real-time object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1137-1149.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. *Int. J. Comp. Vision* 104, 154–171. doi: 10.1007/s11263-013-0620-5

Wei, X.-S., Cui, Q., Yang, L., Wang, P., and Liu, L. (2019). Rpc: A large-scale retail product checkout dataset. *arXiv [Prerint]*. arXiv:1901.07249.

Zhang, J., Li, Y., Li, T., Xun, L., and Shan, C. (2019). License plate localization in unconstrained scenes using a two-stage cnn-rnn. *IEEE Sensors J*. 19, 5256–5265. doi: 10.1109/JSEN.2019.2900257

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv [Preprint]*. arXiv:1904.07850.

# Frontiers in Neuroscience

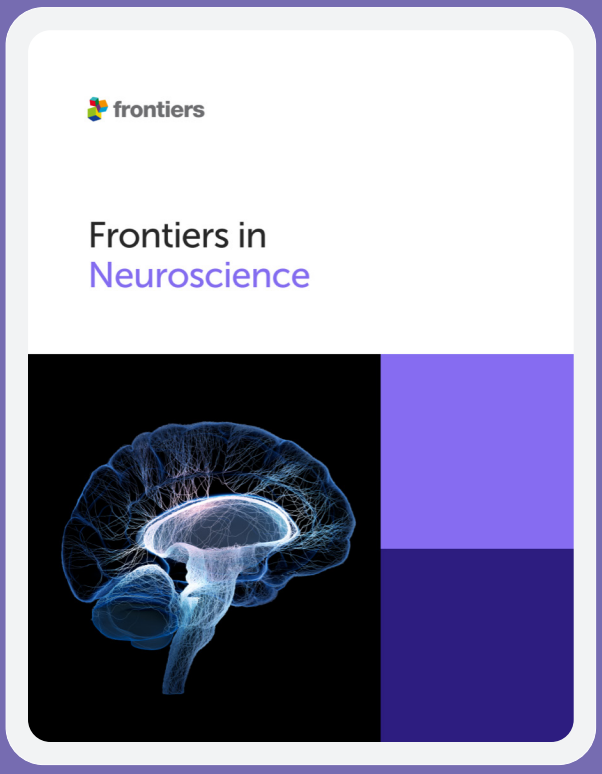
**Provides a holistic understanding of brain function from genes to behavior**

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

---

## Discover the latest Research Topics

See more →




**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact