TYPE Original Research
PUBLISHED 10 July 2024
DOI 10.3389/feart.2024.1381192

OPEN ACCESS

EDITED BY
Abani Kumar Patra,
Tufts University, United States

REVIEWED BY
Satya Narayana,
Chaitanya Bharathi Institute of Technology, India
Yongfeng Zhang,
Soochow University, China
Yang Zhou,
Argonne National Laboratory, United States
Huapeng Yu,
Chengdu University, China

*CORRESPONDENCE
Qiliang Bao,
control@ioe.ac.cn
Zhenming Peng,
zmpeng@uestc.edu.cn

RECEIVED 03 February 2024
ACCEPTED 31 May 2024
PUBLISHED 10 July 2024

CITATION
Rao J, Wu T, Li H, Zhang J, Bao Q and Peng Z (2024), Remote sensing object detection with feature-associated convolutional neural networks.
Front. Earth Sci. 12:1381192.
doi: 10.3389/feart.2024.1381192

Remote sensing object detection with feature-associated convolutional neural networks

Jianghao Rao[1,2,3], Tao Wu[1], Hongyun Li[4], Jianlin Zhang[1], Qiliang Bao[3]* and Zhenming Peng[2]*

[1]Laboratory of Photoelectric Detection and Signal Processing, Institute of Optics and Electronics, Chinese Academy of Sciences (CAS), Chengdu, China, [2]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China, [3]Key Laboratory of Optical Engineering, Institute of Optics and Electronics, Chinese Academy of Sciences (CAS), Chengdu, China, [4]Chengdu Hikvision Digital Technology Co., Ltd., Chengdu, China

Neural networks have become integral to remote sensing data processing. Among neural networks, convolutional neural networks (CNNs) in deep learning offer numerous advanced algorithms for object detection in remote sensing imagery, which is pivotal in military and civilian contexts. CNNs excel in extracting features from training samples. However, traditional CNN models often lack specific signal assumptions tailored to remote sensing data at the feature level. In this paper, we propose a novel approach aimed at effectively representing and correlating information within CNNs for remote sensing object detection. We introduce object tokens and incorporate global information features in embedding layers, facilitating the comprehensive utilization of features across multiple hierarchical levels. Consideration of feature maps from images as two-dimensional signals, matrix image signal processing is employed to correlate features for diverse representations within the CNN framework. Moreover, hierarchical feature signals are effectively represented and associated during end-to-end network training. Experiments on various datasets demonstrate that the CNN model incorporating feature representation and association outperforms CNN models lacking these elements in object detection from remote sensing images. Additionally, integrating image signal processing enhances efficiency in end-to-end network training. Various signal processing approaches increase the process ability of the network, and the methodology could be transferred to other specific and well-defined task.

KEYWORDS

object detection, remote sensing imagery, convolutional neural networks, feature mining, dynamic association

1 Introduction

The advancement of various technologies, including telecommunications, control systems, sensors, and manufacturing, has led to the emergence of unmanned aerial vehicles (UAVs) in the market. Sensors of various types have been equipped on UAVs (Linchant et al., 2015), encompassing radar, radio-frequency sensors, Lidar, and cameras ranging from simple visible light to advanced systems such as multispectral, hyperspectral, or thermal infrared cameras (Jennifer et al., 2022; Panthi and Iungo, 2023; Tian et al., 2024). From remote sensing data collected by those sensors, automatic detection and identification of objects have been widely used in the operation and monitoring of those UVAs. These data imageries by visible-light cameras are of great significance for their abundant information, accessible high resolution, and low costs.

Nowadays, UAVs are being designed to meet the requirements of communication, connectivity, speed, and flight time (Mohsan et al., 2023) in specific scenarios. An important trend is to make UVAs smarter and more intelligent. It is evident that vision tasks play a crucial role in the journey toward machine intelligence. Remote sensing-based approaches demonstrate their significance in the vision tasks for the operation of unmanned aerial vehicles (Aldea and Le Hégarat-Mascle, 2015; Jingyu et al., 2022; Kulkarni et al., 2023). Despite the longstanding existence of this domain, numerous issues and aspects persist, necessitating further study and discussion, especially in light of advancements in machine learning. In analyzing aerial vehicle imagery for vision tasks, the learning mechanism of convolutional neural networks is paramount. These approaches leverage image processing, endowing aerial vehicles with the capability to identify desired targets (Tellaeche et al., 2011; Wu et al., 2021; Al-Badri et al., 2022). Additionally, the availability of pretrained deep neural networks facilitates the study of the underlying mechanisms involved in vision tasks.

Numerous efforts have focused on feature extraction primarily through manual feature engineering (Lv et al., 2019; Lei et al., 2021; Ma and Filippi, 2021). In addition to these established methods, some studies have advanced this approach further. For instance, Xiao et al. (2014) employed the elliptic Fourier transform (EFT), enhancing invariance compared with the histogram of oriented gradients (HOG) features.

Local features constitute the most commonly employed foundation for capturing the attributes of items, for instance, the scale-invariant feature transform (Han et al., 2015), HOG (Shao et al., 2012), and saliency (Han et al., 2014; Zhang et al., 2015). Although these methods exhibit a degree of adaptability and certain invariance, they fall short when confronted with intricate scenarios and heightened performance demands. This limited invariant capability proves insufficient for practical feature extraction requirements. Consequently, for aerial vehicles, the imperative to bolster feature extraction capabilities when addressing diverse objects in remote sensing imagery becomes increasingly evident.

Traditional handcrafted feature approaches (Xiao et al., 2015; Li et al., 2020) often falter in remote sensing object detection due to inherent signal assumptions. Local features represent the foundational attributes of images, whereas object detection, conversely, pertains to higher-level semantic analysis. Cheng and Han (2016) provide an overview of recent advancements in object detection within remote sensing imagery. In contrast to local features, part-based models (Huang et al., 2010; Li et al., 2012; Cheng et al., 2014; Zhang et al., 2014; Cheng et al., 2015) have gained traction as popular mid-level features. Moreover, some studies have embraced semantic models for extracting semantic information rather than relying on superficial features (Sun et al., 2012; Cheng et al., 2013; Yao et al., 2016).

Considering that traditional image processing approaches rely heavily on feature extraction operators and are not robust enough, we conduct a detailed study of convolutional neural networks. In object detection from aerial vehicle imagery, we focus on the latent mechanism of feature representation and association, aiming at broadening the use of image processing for intelligent aerial vehicles, which may be publicly available and developed for various uses after further studies. Additionally, neural networks mitigate the need for stringent data signal assumptions and can be regarded as a versatile signal model. Through training optimization, these networks adeptly capture the intricate input–output relationships.

In this paper, we delve into the inner functions of feature representation and association in object detection from remote sensing imagery, with the aim of expanding the scope of image processing for intelligent aerial vehicles. These advancements are intended to be publicly accessible and adaptable for diverse applications and future research. The main contributions are summarized as follows:

(1) We represent object tokens using embedding structures, whereas global information features are encoded using multi-head attention. Both methodologies draw inspiration from natural language processing, enabling the extraction of task-oriented features across various hierarchical maps.
(2) Image signal processing is employed to associate features from various perspective representations. This operation involves correlating target features and global features of remote sensing data in a pertinent and iterative manner.
(3) Various hierarchical feature signals are represented and associated to ensure adaptability to convolutional neural network (CNN) models during training and inference stages. This signal processing enhances the learning capabilities of convolutional networks in remote sensing object detection.

2 Related work

Some recent works (Kussul et al., 2017; Long et al., 2017; Li et al., 2022; Qingyun and Zhaokui, 2022; Ahmed et al., 2023; Ayesha et al., 2023) have successfully applied deep models to remote sensing object detection. Based on various CNN structures, these models achieved some performance improvements on these tasks. From data-driven point of view, the quality of data determines the performance of the model, and the effect of signal processing on object detection is ignored. However, for a specific task, when patterns of the task are already known as exactly true, any approach based on data-fitting to describe a task can only be a suboptimal solution. Remote sensing object detection is a typical well-defined task, which is relatively fixed in imaging scope and types of targets. For such a well-defined task, common CNN models are not targeted on how they perform signal processing (Cheng et al., 2015) using neural network feature maps.

3 Proposed models

In this section, we first introduce the overall pipeline used for the processing and describe the processing mechanisms. After this, signal processing to feature maps in the CNN model is analyzed in detail.

3.1 The proposed pipeline of remote sensing object detection

The overall architecture of our detector is shown in Figure 1.

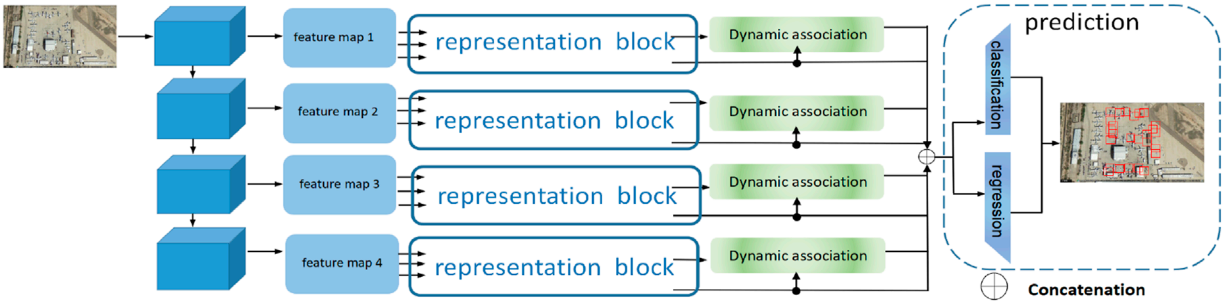


FIGURE 1
Pipeline and analysis of our proposed approach.

The first part extracts features for later use, just like the classical and popular CNN detectors in which convolutional layers are used as a backbone. The backbone usually is for deep mining and abstraction in multi-feature semantically. In the part of feature extraction, we use the residual structure of the famous and efficient backbone named ResNet50. To problems in the gradient descent when the model is trained, not only the skip connections are contained in the backbone but also features of varying depth and richness are provided as the features are pooled for later use.

The second part in the figure explains the representation block. The purpose of this block is to let the CNN focus on the features of interest and extract the important task-oriented features in various depths dynamically and fully.

The dynamic association part is designed to associate the ROI (region of interest) features with the attentive global features, which aims to get information about the remote sensing object. Finally, the processed information in various depths is added with the attentive feature maps from multi-head attention and then sent to the prediction network, and the outputs of classes and locations are obtained.

3.2 Representation block

Beyond the feature extraction of the backbone, we create two embedding layers per feature map in the top–down structure. The function of embedding layers in this paper is shown as Figure 2. One embedding layer is to get the tokens of proposal boxes, which can acquire the location of the target determined by the learnable weights, the other is to learn the tokens for the features per proposal box. These represent the various features of the objects in images.

In this block, the embedding layers play important roles when the whole model is trained. In the beginning, features in the features pool are analyzed by these. The tokens in embedding layers represent attentive features to our input images. Following these, ROI pooling and multi-head attention provide the global features and the target features perspectively for later use. This block is shown as Figure 3.

3.3 Association block

The previous block generates the ROI features and attentive feature maps. Based on these, the association module generates new feature maps, which correspond to different kinds of objects in various inputs. The internal structure of this module is shown in Figure 4. The input features consist of ROI features (features of interest) and attention features provided by the attentive feature maps.

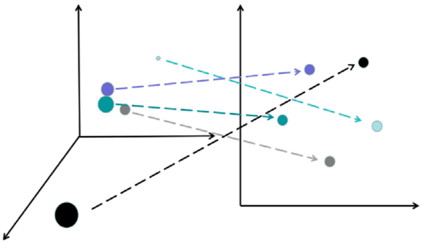


FIGURE 2
Embedding layer reduces the data dimension and makes the distance of feature attributes more reasonable.

The association operation is essentially made up of multiplication and addition with nonlinear transformation. In order to have the ability to adjust dynamically based on the characteristics of inputs and targets, the attention features containing the highly important targets and the information generate parameters that can be used as a basis for the dynamic attention of targets.

To make the whole dynamic mechanism have the properties of validity and accuracy, this module contains three main operations:

3.3.1 Coefficient matrix (CM)

We designed trainable linear layers and a series of subsequent operations to get parameter tensors. For input feature maps, we make it a vector in order, $F_{input} = \{x_1, x_2 \cdots x_{w*h}\}$; we define:

$$Linear\left(F_{input}, W^*\right) = F_{input} \cdot W \tag{1}$$

W represents the parameter matrix $(w_{i,j})$, $0 < = i < = w * h$, $0 < = j < = c$, and then, we get vector F_1.

In this way, given input feature maps F_{input} and $F_{input} \in R^{H_1, W_1}$ are transferred to $F_1 \in R^{H_1 * W_1}$, $F_1 \in R^{1,C,H_1 * W_1}$, to channel $C = c$, $I(x,y) = F_1^{1,c,x*y}$; we define factor η, and then, we get the new region:

$$I(x,y) = \begin{cases} F_1^{1,c,x*y*\eta}, & if\ (x * y)//\eta = 0 \\ 0, & Otherwise \end{cases} \tag{2}$$

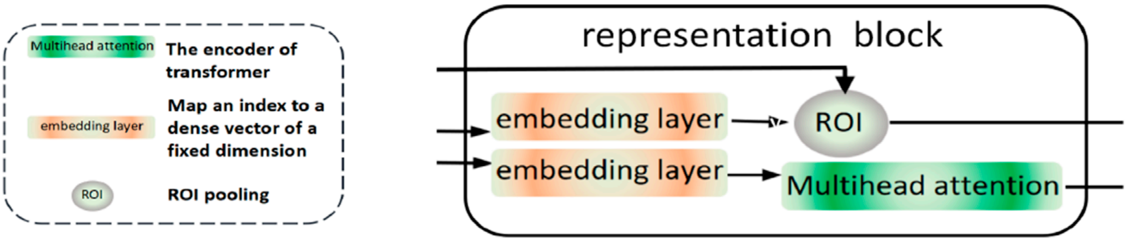


FIGURE 3
Diagram of the representation block.

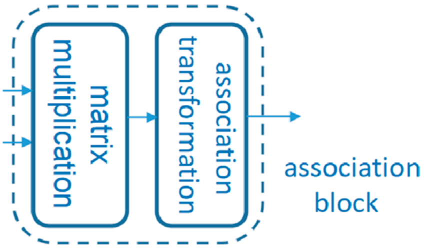


FIGURE 4
Diagram of the association block.

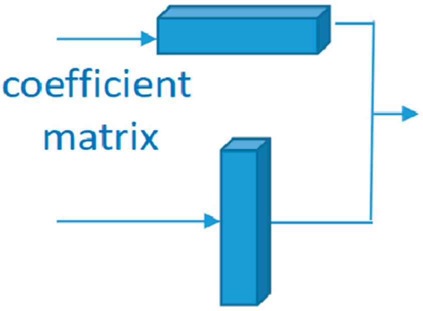


FIGURE 5
Diagram of matrix multiplication.

To ensure that the size of the tensor is effective and seamless in different situations, the parameter generation inside the dotted box is not the same as those outside. We can visually see the difference between the two from the diagram below. The structure of parameter generation inside the dotted box is shown in Figure 4. The structure of parameter generation outside the dotted box is shown in Figure 5.

We select the non-zero elements in order from $I(x,y)$ and get tensor $F_{ouput} \in R^{c,h*w}$. If the feature maps are not suitable for the following layers, we can reorganize this to get the coefficient matrix with the various shapes we want.

The whole process can be summarized as follows:

$$P = CM(F). \tag{3}$$

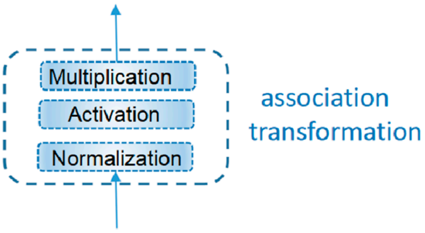


FIGURE 6
Diagram of association transformation.

3.3.2 Association transformation (AT)

As shown in the dotted box of the figure, nonlinear transformation is essential and effective. “Norm” represents the normalization in the neural network; the module generates a feature map set $B = \{F_1, F_2 \cdots F_m\}$, $B \in R^{H,W,C}$.

In set B, there are m feature maps, where m is the number of channels, x_j^i is the i-th pixel value of feature map F_j, and first μ_b is calculated.

$$\mu_B = \frac{1}{w*h*m} \sum_{j=1}^{m} \sum_{i=1}^{w*h} x_j^i. \tag{4}$$

w and h represent the width and height of feature maps, correspondingly.

$$\sigma_B^2 = \frac{1}{w*h*m} \sum_{j=1}^{m} \sum_{i=1}^{w*h} \left(x_j^i - \mu_B\right)^2 \tag{5}$$

$$X_j^i = \frac{x_j^i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{6}$$

X_j^i represents the i-th pixel value of feature map f_{outj}, in order to have a better value distribution, we scale this as

$$X_{out_j}{}^{i} = \gamma X_j^i + \beta. \tag{7}$$

γ and β are parameters to be learned.

For two sets $B_1 = \{F_{11}, F_{12} \cdots F_{1m}\}$, $B_2 = \{F_{21}, F_{22} \cdots F_{2m}\}$, we can get two groups of outputs in this way, $B_{out1} = \{F_{out11}, F_{out12} \cdots F_{out1m}\}, B_{out2} = \{F_{out21}, F_{out22} \cdots F_{out2m}\}$.

Matrix multiplication is used:

$$F_{out\,m=i} = relu\left(F_{out1\,m=i}\right) \cdot relu\left(F_{out2\,m=i}\right) \tag{8}$$

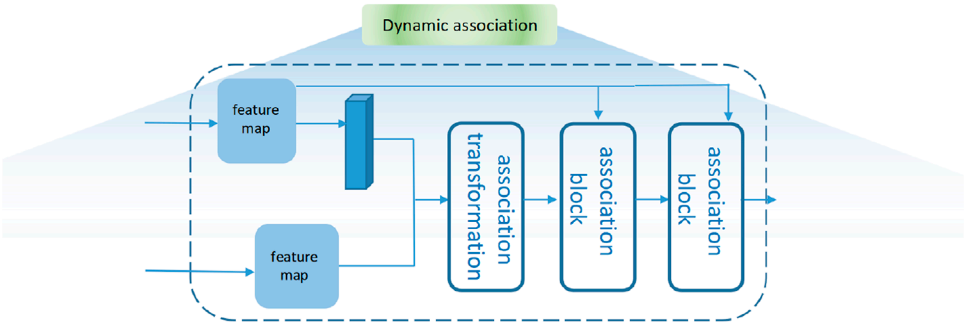
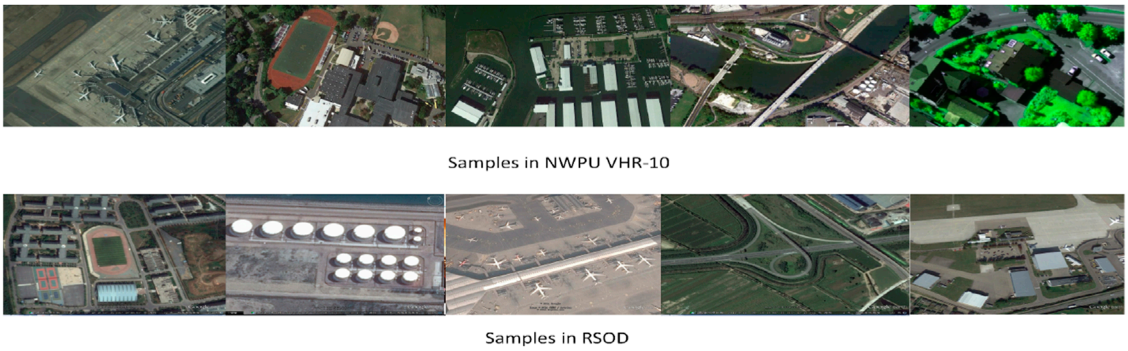


FIGURE 7
Diagram of dynamic association.


Samples in NWPU VHR-10

Samples in RSOD

FIGURE 8
Samples in datasets.

The function *relu*() mentioned above is defined as follows:

$$relu(x) = Maximum(x, 0) \tag{9}$$

This part is shown in Figure 6. The whole process can be noted as follows:

$$B_{out} = AT(B_1, B_2). \tag{10}$$

3.3.3 Feature association (FA)

When we get features of interests $F_1 \in R^{h_1,w_1}$ and attention features $F_2 \in R^{h_2,w_2}$, the network generates parameter tensors:

$$B_{out1} = CM(F_2). \tag{11}$$

We reshape B_{out1} to $B_1 \in R^{w_1,w_3}$

Then, F_1 and B_{out} are processed as follows:

$$F_3 = AT(F_1, B_1), F_3 \in R^{h_1,w_3}. \tag{12}$$

We repeat the first step to generate another coefficient matrix:

$$B_{out2} = CM(F_2) \tag{13}$$

Then, reshape B_{out2} to $B_2 \in R^{w_3,w_1}$:

$$F_4 = AT(F_3, B_2), F_4 \in R^{h_1,w_1} \tag{14}$$

Notably, the input F_1 and output F_4 are tensors of the same size, and it is a plug-and-play-in section. Therefore, it can be added directly later to get a deeper dynamic association module. The experiments prove that two of these works are better than one.

The whole process can be noted as follows:

$$F_4 = FA(F_2) \tag{15}$$

3.4 Dynamic association

From every embedding layer in the representation block, we get a group of feature maps with various sizes, $B = \{F_1, F_2 \cdots F_m\}$, $F_i \in R^{h_i,w_i}$. Processed by ROI pooling, we get a series of feature maps with the same size from B called B_1, $B_1 = \{f_1^1, f_2^1 \cdots f_m^1\}$, $f_i^1 \in R^{h_1,w_1}$.

To the classification and regression of remote sensing data, another group of attentive feature maps is generated by multi-head attention, and we name them as B_2, $B_2 = \{f_1^2, f_2^2 \cdots f_m^2\}$, $f_i^2 \in R^{h_2,w_2}$.

For $f_i^2 \in R^{h_2,w_2}$, we can get coefficient matrix, P_i:

$$P_i = CM(f_i^2) \tag{16}$$

Then, the features of interest are processed based on the coefficient matrix:

$$f_i^3 = AT(f_i^1, P_i) \tag{17}$$

TABLE 1 Image resolution of RSOD.

Category	Oil tank	Aircraft	Overpass	Playground
Resolution (m)	0.3 ~ 1	0.5 ~ 2	1.25 ~ 3	0.4 ~ 1

TABLE 2 Ablation study.

Method	Representation block	Association block	Dynamic association	mAP
1				65.4
2	✓			65.8
3	✓	✓		66.0
4	✓	✓	✓	68.6

TABLE 3 Comparison with popular deep learning detection models on RSOD.

Method	AP	AP50	APs	APm	APl
Sparse-RCNN (Sun et al., 2021)	64.6	94.7	37.2	69.6	69.8
Yolo-v8 (Reis et al., 2023)	55.0	90.1	48.8	68.8	58.1
Yolox (Song et al., 2022)	53.6	91.1	47.2	65.8	57.0
Faster-rcnn (Ren et al., 2015)	67.9	95.2	**52.3**	**71.5**	71.8
Retinanet (Lin et al., 2020)	65.2	95.8	46.1	70.7	69.4
DETR (Carion et al., 2020)	57.4	90.7	8.4	57.1	63.9
ours	**68.6**	**96.6**	42.0	62.2	**74.4**

The highest performance results are shown in bold black.

TABLE 4 Comparison with popular deep learning detection models on NWPU VHR-10.

Method	AP	AP50	APs	APm	APl
Sparse-RCNN	48.8	75.0	26.8	51.9	52.3
Yolo-v8	55.4	85.3	37.4	56.8	61.1
Yolox	53.2	78.7	31.3	55.2	58.8
Faster-rcnn	54.5	81.7	**44.7**	58.9	53.0
Retinanet	54.5	90.4	20.5	56.2	49.5
DETR	46.5	87.1	16.8	47.2	44.9
Ours	**62.7**	**91.9**	35.0	**62.1**	**61.1**

The highest performance results are shown in bold black.

Different features processed for the probable objects are inter-correlated by the dynamic association module. Dynamic association is shown in Figure 7. The approach of inter-attention makes these features fully distinguished and compared. We do the operation below:

$$f_i^4 = \left\{FA\left(f_i^3\right)\right\}_{N_1} \tag{18}$$

$\{\cdot\}_{N_1}$ means that the operation is repeated N_1 times. The usual way is to set fixed anchor boxes to class and regress. Instead, the approach pays attention to the probable information at the feature extraction level and compares dynamically from the global attentive feature maps.

$$f_i^5 = f_i^4 + f_i^2 \tag{19}$$

So $B_5 = \{f_1^5, f_2^5 \cdots f_m^5\}$; for the feature map of the $i-$th channel, elements in f_i^5 are extracted in order to get a 1D vector and then classifier and regression layer, which is followed by the calculation of the final results. The loss function is calculated as follows:

$$L = \lambda_{cls} \cdot L_{cls} + \lambda_{L1} \cdot L_{L1} + \lambda_{giou} \cdot L_{giou} \tag{20}$$

Here, L_{cls} is the focal loss (Chen, 2009) of predicted classifications and ground truth category labels, and L_{L1} and $Lgiou$ are L1 loss and generalized IoU (intersection over union) loss (Tianditu, 2016) between normalized center coordinates and height and width of predicted boxes and ground truth box, respectively.

TABLE 5 Performances of various categories on RSOD.

Category	Aircraft	Playground	Oil tank	Overpass
AP	62.0	88.3	78.0	46.0

TABLE 6 Performances of various categories on NWPU VHR-10-1.

Category	Airplane	Ship	Storage tank	Baseball diamond	Tennis court
AP	70.8	64.0	43.2	71.6	64.9

TABLE 7 Performances of various categories on NWPU VHR-10-2.

Category	Basketball court	Ground track field	Harbor	Bridge	Vehicle
AP	77.3	85.0	42.4	42.8	64.4

4 Experiments

In this section, we select various datasets with practical significance. Through the ablation study, we verify the innovation and effectiveness of our method. Through a series of evaluation metrics, we compare our proposed method with the current mainstream and high-performance neural network detectors.

4.1 Datasets

All datasets were divided into training sets and test sets according to the ratio of 4:1. After models were fully trained on training sets, the untrained test sets were used to verify the performances of models. Samples in datasets are shown in Figure 8.

RSOD (Tianditu, 2016): It collected 2326 images downloaded from Google Earth and Tianditu (Han et al., 2023) and labeled the objects in these images with four categories: oil tank, aircraft, overpass, and playground. The image resolution for each class, representing image size and clarity under different imaging conditions, is listed in Table 1. The sensors involved are panchromatic and multispectral due to the various sources of the image data sets. In this way, the diversity of the data set poses comprehensive performance challenges.

NWPU VHR-10 (Cheng et al., 2016): It contains 800 high-resolution satellite images cropped from the Google Earth and Vaihingen datasets and was annotated by experts manually. The dataset was divided into 10 categories (airplanes, ships, storage tanks, baseball diamonds, tennis courts, basketball courts, ground track fields, harbors, bridges, and vehicles).

4.2 Evaluation metrics

The dataset was divided into a training set and a test set; we used mAP (mean average precision) to assess the overall performance of the test set. The confidence of the IoU (intersection over union) threshold for AP50 calculation is 0.5. The confidence of the IoU threshold for AP75 calculation is 0.75; when confidence is set from 0.5 to 0.95 and calculated once every 0.05 interval, we can get various results and calculate their average value to get the evaluation of AP.

AP was divided into APs, APl, and APm based on targets, with areas less than 32 square pixels, more than 96 square pixels, and in the middle of the two situations, respectively.

4.3 Parameter settings

The backbone of the network is ResNet-50. The optimizer is AdamW with a weight decay of 0.0001. Setting batch size to 16, we train models on 3080-Ti. The initial learning rate is set to 2.5×10^{-5}, divided by 10 at epochs 27 and 33, respectively. The whole schedule of training contains 36 epochs. The backbone is initialized with the pretrained weights on ImageNet, and other newly added layers are initialized with Xavier. Data augmentation includes random horizontal, scale jitter of resizing the input images.

4.4 Ablation study

To evaluate the effectiveness of the signal processing in the feature maps in CNNs, we conducted the ablation study on the processing pipeline and proved the effectiveness.

To show that the proposed processing varies from previous CNN models and prove its effectiveness in remote sensing object detection, ablation experiments on various datasets are conducted. In experiments, we compare famous CNN models with our proposed pipeline. Our pipeline and CNN models are fully trained in the same way. The results of the ablation study are shown in Table 2. There are various methods in the table. Method 1 refers

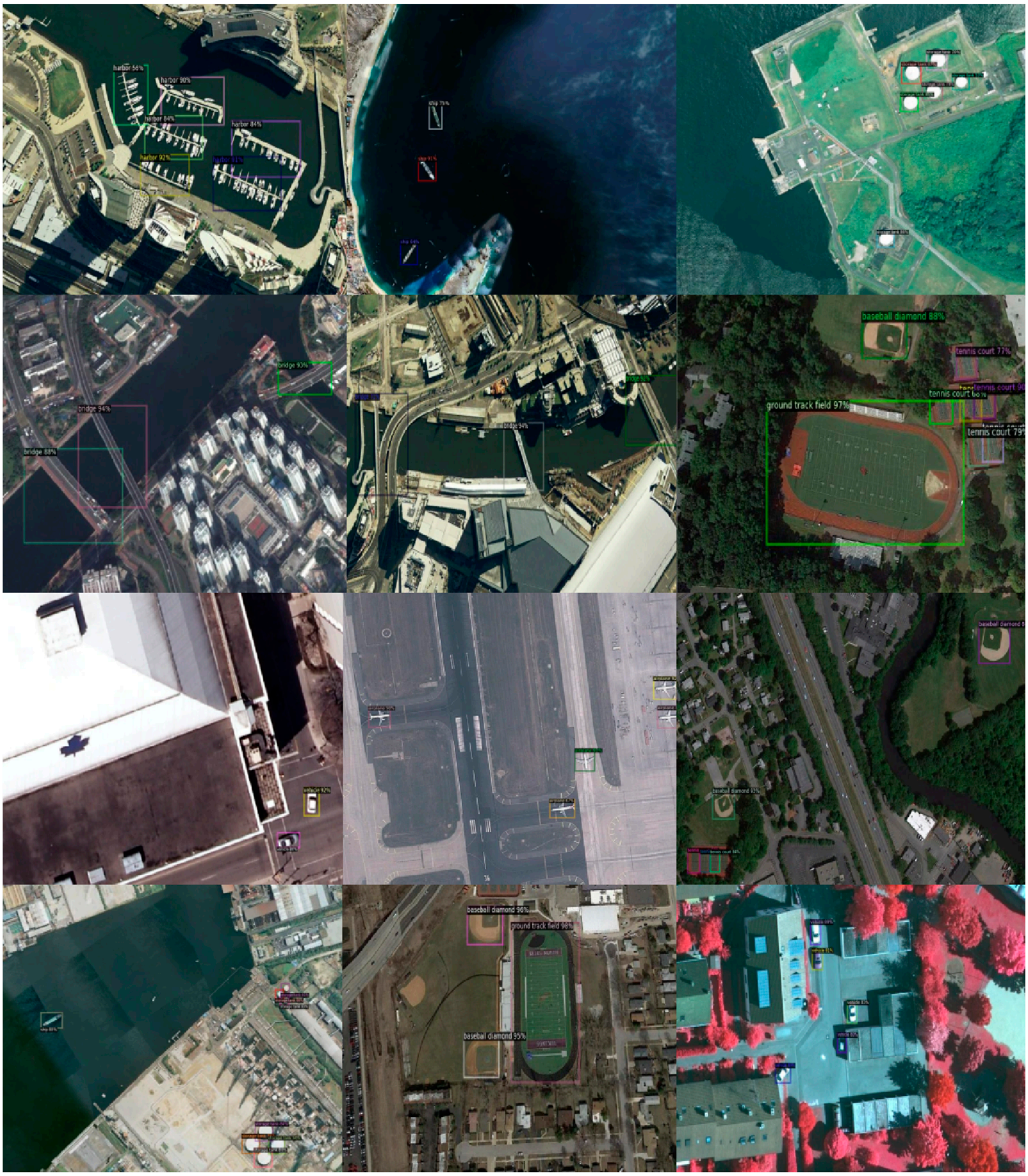


FIGURE 9
Detection results on NWPU VHR-10.

to the only use of ResNet and the prediction module in Figure 1, not considering other approaches or modules we proposed in this manuscript. Different from Method 1, Method 2 adopts the representation block we proposed based on Method 1. Furthermore, an association block is added to Method 2, which comes into being the Method 3. We can see that the representation block and association block prove the performance of CNN models slightly and consistently. Lastly, we combine Method 3 with the dynamic association and represent Method 4 in essence, which greatly improves the detection performance.

4.5 Comparison with other CNN models

Our pipeline focuses on feature signal representation and association, and we compared it with other popular CNN models.

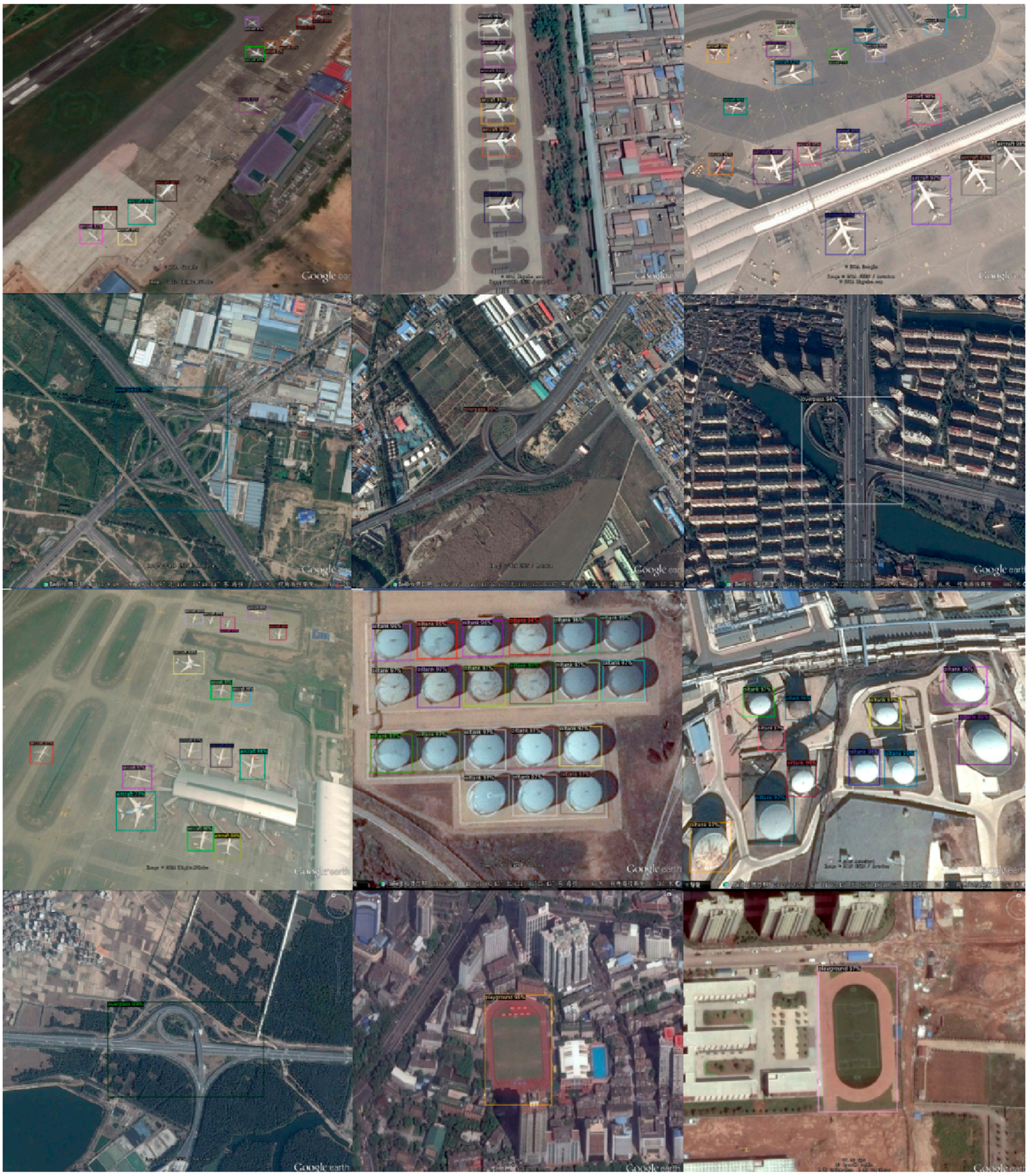


FIGURE 10
Detection results on RSOD.

All models are trained on these same datasets for detection. In Tables 3, 4, the performances of ours and others are shown in detail. The APs of each category in the two datasets are shown in Tables 5–7, the detection performances are shown in Figures 9 and 10. Given the correct category labels, objects of different scales in various backgrounds can be accurately detected.

Diverse detection models perform differently on targets of various sizes. Overall, our method takes better care of distinct targets of various sizes. Compared with other detection models, our approach achieved a better performance.

5 Conclusion

In this paper, the feature signal processing of remote sensing object detection within CNNs is discussed. In the feature level of neural networks, signal processing focused primarily on feature representation and association is studied. Within CNNs, the representation and association pipeline we proposed suits end-to-end network training effectively. Compared with several CNN models lacking this kind of processing on various datasets, our approach enhances the handling of features and outperforms other approaches on remote sensing object detection. However, the scope of the application and deeper reasons remain to be revealed after further exploration. A better signal processing analysis of features of CNNs, which may combine various signal processing approaches and increase the process ability of the network, could be the central idea in future works for a specific and well-defined task.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: https://map.tianditu.com/map/index.html.

Author contributions

JR: Conceptualization, data curation, and writing–original draft. TW: Conceptualization, data curation, methodology, project administration, and writing–original draft. HL: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, and writing–review and editing. JZ: Investigation, methodology, project administration, supervision, and writing–review and editing. QB: Investigation, methodology, project administration, and writing–review and editing. ZP: Conceptualization, investigation, methodology, supervision, validation, and writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

Author HL was employed by Chengdu Hikvision Digital 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

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

Ahmed, M. A., Althubiti, S. A., de Albuquerque, V. H. C., dos Reis, M. C., Shashidhar, C., Murthy, T. S., et al. (2023). Fuzzy wavelet neural network driven vehicle detection on remote sensing imagery. *Comput. Electr. Eng.* 109, 108765. doi:10.1016/j.compeleceng.2023.108765

Al-Badri, A. H., Ismail, N. A., Al-Dulaimi, K., Salman, G. A., Khan, A., Al-Sabaawi, A., et al. (2022). Classification of weed using machine learning techniques: a review—challenges, current and future potential techniques. *J. Plant Dis. Prot.* 129, 745–768. doi:10.1007/s41348-022-00612-9

Aldea, E., and Le Hégarat-Mascle, S. L. (2015). Robust crack detection for unmanned aerial vehicles inspection in an a-contrario decision framework. *J. Electron. Imaging* 24, 061119. doi:10.1117/1.jei.24.6.061119

Ayesha, B. E., Satyanarayana Murthy, T., Babu, P. R., and Kuchipudi, R. (2023). "Ship detection in remote sensing imagery for arbitrarily oriented object detection," in *ICT analysis and applications. ICT4SD 2023. Lecture notes in networks and systems, volume 782.* Editors S. Fong, N. Dey, and A. Joshi (Singapore: Springer), 457–466. doi:10.1007/978-981-99-6568-7_42

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., et al. (2020). End-to-End object detection with transformers. in *Computer Vision – ECCV 2020* (Glasgow, UK: Springer), 213–229. doi:10.1007/978-3-030-58452-8_13

Chen, C. H. (2009). "On the roles of advanced signal processing in remote sensing," in Proc. SPIE 7477, Image and Signal Processing for Remote Sensing XV, Berlin, Germany, 28 September 2009, 747709. doi:10.1117/12.836022

Cheng, G., Guo, L., Zhao, T., Han, J., Li, H., and Fang, J. (2013). Automatic land slide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. *Int. J. Remote Sens.* 34 (1), 45–59. doi:10.1080/01431161.2012.705443

Cheng, G., Han, J., Guo, L., Liu, Z., Bu, S., and Ren, J. (2015). Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* 53 (8), 4238–4249. doi:10.1109/tgrs.2015.2393857

Cheng, G., and Han, J. (2016). A survey on object detection in optical remote sensing images. *ISPRS J. Photogram. Remote Sens.* 117, 11–28. doi:10.1016/j.isprsjprs.2016.03.014

Cheng, G., Han, J., Zhou, P., and Guo, L. (2014). Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* 98, 119–132. doi:10.1016/j.isprsjprs.2014.10.002

Cheng, G., Zhou, P., and Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geoscience Remote Sens.* 54 (12), 7405–7415. doi:10.1109/tgrs.2016.2601622

Han, J., Zhang, D., Cheng, G., Guo, L., and Ren, J. (2015). Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* 53 (6), 3325–3337. doi:10.1109/tgrs.2014.2374218

Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., et al. (2014). Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *ISPRS J. Photogram. Remote Sens.* 89, 37–48. doi:10.1016/j.isprsjprs.2013.12.011

Han, L., Li, F., Yu, H., Xia, K., Xin, Q., and Zou, X. (2023). BiRPN-YOLOvX: a weighted bidirectional recursive feature pyramid algorithm for lung nodule detection. *J. Xray Sci. Technol.* 31, 301–317. doi:10.3233/XST-221310

Huang, Y., Zhang, L., Li, P., and Zhong, Y. (2010). High-resolution hyper-spectral image classification with parts-based feature and morphology profile in urban area. *Geo-Spatial Inf. Sci.* 13 (2), 111–122. doi:10.1007/s11806-010-0004-8

Jennifer, C., Narayanaswamy, B. E., Waluda, C. M., and Williamson, B. J. (2022). Aerial detection of beached marine plastic using a novel, hyperspectral short-wave infrared (SWIR) camera. *ICES J. Mar. Sci.* 79 (3), 648–660. doi:10.1093/icesjms/fsac006

Jingyu, H. E., Xiao, Y., Bogdan, C., Nazarian, S., and Bogdan, P. (2022). A design methodology for energy-aware processing in unmanned aerial vehicles. *ACM Trans. Des. Automation Electron. Syst.* 27 (1), 1–20. doi:10.1145/3470451

Kulkarni, N. N., Raisi, K., Valente, N. A., Benoit, J., Yu, T., and Sabato, A. (2023). Deep learning augmented infrared thermography for unmanned aerial vehicles structural health monitoring of roadways. *Automation Constr.* 148, 104784. doi:10.1016/j.autcon.2023.104784

Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sens. Lett.* 14, 778–782. doi:10.1109/lgrs.2017.2681128

Lei, W., Yang, H., and Tang, M. (2021). Extraction of carbon emission feature in urban residential area based on remote sensing technology. *Int. J. Environ. Technol. Manag.* 24 (1/2), 120. doi:10.1504/IJETM.2021.10038735

Li, J., Zhuang, Y., Dong, S., Gao, P., Dong, H., Chen, H., et al. (2022). Hierarchical disentangling network for building extraction from very high resolution optical remote sensing imagery. *Remote Sens.* 14, 1767. doi:10.3390/rs14071767

Li, K., Wan, G., Cheng, G., Meng, L., and Han, J. (2020). Object detection in optical remote sensing images: a survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* 159, 296–307. doi:10.1016/j.isprsjprs.2019.11.023

Li, Y., Sun, X., Wang, H., Sun, H., and Li, X. (2012). Automatic target detection in high-resolution remote sensing images using a contour-based spatial model. *IEEE Geosci. Remote Sens. Lett.* 9 (5), 886–890. doi:10.1109/lgrs.2012.2183337

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollar, P. (2020). Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (2), 318–327. doi:10.1109/TPAMI.2018.2858826

Linchant, J., Lisein, J., Semeki, J., Lejeune, P., and Vermeulen, C. (2015). Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal. Rev.* 45, 239–252. doi:10.1111/mam.12046

Long, Y., Gong, Y., Xiao, Z., and Liu, Q. (2017). Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geoscience Remote Sens.* 55 (5), 2486–2498. doi:10.1109/TGRS.2016.2645610

Lv, Y., Zhang, X., Xiong, W., Cui, Y., and Cai, M. (2019). An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification. *Remote Sens.* 11, 3006. doi:10.3390/RS11243006

Ma, A., and Filippi, A. M. (2021). Computationally efficient sequential feature extraction for single hyperspectral remote sensing image classification. *ICA* 3, 1–2. doi:10.5194/ica-abs-3-189-2021

Mohsan, S. A. H., Othman, N. Q. H., Li, Y., Alsharif, M. H., and Khan, M. A. (2023). Unmanned aerial vehicles (UAVs): practical aspects, applications, open challenges, security issues, and future trends. *Intell. Serv. Robot.* 16, 109–137. doi:10.1007/s11370-022-00452-4

Panthi, K., and Iungo, G. V. (2023). Quantification of wind turbine energy loss due to leading-edge erosion through infrared-camera imaging, numerical simulations, and assessment against SCADA and meteorological data. *Wind energy* 26, 266–282. doi:10.1002/we.2798

Qingyun, F., and Zhaokui, W. (2022). Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognit.* 130, 108786. doi:10.1016/j.patcog.2022.108786

Reis, D., Kupec, J., Hong, J., and Daoudi, A. (2023). *Real-time flying object detection with YOLOv8*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: towards real-time object detection with re-gion proposal networks," in NeurIPS.

Shao, W., Yang, W., Liu, G., and Liu, J. (2012). "Car detection from high resolution aerial imagery using multiple features," in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Munich, Germany, 22-27 July 2012 (IEEE), 4379–4382.

Song, C., Zhang, F., Li, J., Xie, J., Yang, C., Zhou, H., et al. (2022). Detection of maize tassels for UAV remote sensing image with an improved YOLOX model. *J. Integr. Agric.* 22, 1671–1683. doi:10.1016/j.jia.2022.09.021

Sun, H., Sun, X., Wang, H., Li, Y., and Li, X. (2012). Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geosci. Remote Sens. Lett.* 9 (1), 109–113. doi:10.1109/lgrs.2011.2161569

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., et al. (2021). "Sparse R-CNN: end-to-end object detection with learnable proposals," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14454–14463. doi:10.1109/cvpr46437.2021.01422

Tellaeche, A., Pajares, G., Burgos-Artizzu, X. P., and Ribeiro, A. (2011). A computer vision approach for weeds identification through Support Vector Machines. *Appl. Soft Comput.* 11, 908–915. doi:10.1016/j.asoc.2010.01.011

Tian, M., Zhang, J., Yang, Z., Li, M., Li, J., and Zhao, L. (2024). Detection of early bruises on apples using near-infrared camera imaging technology combined with adaptive threshold segmentation algorithm. *J. food process Eng.* 47. doi:10.1111/jfpe.14500

Tianditu (2016). *Tianditu*. Available at: http://map.tianditu.com/map/index.html (Accessed February 01, 2016).

Wu, Z., Chen, Y., Zhao, B., Kang, X., and Ding, Y. (2021). Review of weed detection methods based on computer vision. *Sensors* 21, 3647. doi:10.3390/s21113647

Xiao, Z., Liu, Q., Tang, G., and Zhai, X. (2014). Elliptic Fourier transformation based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* 36 (2), 618–644. doi:10.1080/01431161.2014.999881

Xiao, Z., Liu, Q., Tang, G., and Zhai, X. (2015). Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* 36 (2), 618–644. doi:10.1080/01431161.2014.999881

Yao, X., Han, J., Cheng, G., Guo, L., and Qian, X. (2016). Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Trans. Geosci. Remote Sens.* 54 (6), 3660–3671. doi:10.1109/tgrs.2016.2523563

Zhang, F., Du, B., and Zhang, L. (2015). Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.* 53 (4), 2175–2184. doi:10.1109/tgrs.2014.2357078

Zhang, W., Sun, X., Fu, K., Wang, C., and Wang, H. (2014). Object detec tion in high-resolution remote sensing images using rotation invariant parts based model. *IEEE Geosci. Remote Sens. Lett.* 11 (1), 74–78. doi:10.1109/lgrs.2013.2246538