TYPE Original Research
PUBLISHED 15 July 2025
DOI 10.3389/fcomp.2025.1576775

Check for updates

OPEN ACCESS

EDITED BY
Changxin Gao,
Huazhong University of Science and Technology, China

REVIEWED BY
Erwei Yin,
Tianjin Artificial Intelligence Innovation Center (TAIIC), China
Haitao Gan,
Hubei University of Technology, China
Rosepreet Kaur Bhogal,
Lovely Professional University, India

*CORRESPONDENCE
Lizhong Jin
✉ jinlizhong0@gmail.com

RECEIVED 14 February 2025
ACCEPTED 16 June 2025
PUBLISHED 15 July 2025

CITATION
Jin L, Fan R, Han X and Cui X (2025) Convolutional spatio-temporal sequential inference model for human interaction behavior recognition.
Front. Comput. Sci. 7:1576775.
doi: 10.3389/fcomp.2025.1576775

Convolutional spatio-temporal sequential inference model for human interaction behavior recognition

Lizhong Jin[1]*, Rulong Fan[2], Xiaoling Han[1] and Xueying Cui[1]

[1]School of Applied Science, Taiyuan University of Science and Technology, Taiyuan, China, [2]School of Computer Science and Technology, Taiyuan University of Science and Technology (TYUST), Taiyuan, China

Introduction: Human action recognition is a critical task with broad applications and remains a challenging problem due to the complexity of modeling dynamic interactions between individuals. Existing methods, including skeleton sequence-based and RGB video-based models, have achieved impressive accuracy but often suffer from high computational costs and limited effectiveness in modeling human interaction behaviors.

Methods: To address these limitations, we propose a lightweight Convolutional Spatiotemporal Sequence Inference Model (CSSIModel) for recognizing human interaction behaviors. The model extracts features from skeleton sequences using DINet and from RGB video frames using ResNet-18. These multi-modal features are fused and processed using a novel multiscale two-dimensional convolutional peak-valley inference module to classify interaction behaviors.

Results: CSSIModel achieves competitive results across several benchmark datasets: 87.4% accuracy on NTU RGB+D 60 (XSub), 94.1% on NTU RGB+D 60 (XView), 80.5% on NTU RGB+D 120 (XSub), and 84.9% on NTU RGB+D 120 (XSet). These results are comparable to or exceed those of state-of-the-art methods.

Discussion: The proposed method effectively balances accuracy and computational efficiency. By significantly reducing model complexity while maintaining high performance, CSSIModel is well-suited for real-time applications and provides a valuable reference for future research in multi-modal human behavior recognition.

KEYWORDS

human behavior recognition, deep learning, multimodal learning, skeleton point sequence information, time series recognition, inference mode

1 Introduction

In today's rapidly advancing digital landscape, video content has become a dominant medium for information dissemination, social interaction, and cultural preservation. This surge in video usage has made Human Action Recognition (HAR) an important research topic in computer vision and human-computer interaction, aiming to automatically identify and interpret human behaviors (Yan et al., 2018; Trivedi and Sarvadevabhatla, 2022; Duan et al., 2022a; Duan et al., 2022b; Liu et al., 2018; Liu et al., 2020a; Shu et al., 2019; Shu et al., 2017; De Boissiere and Noumeir, 2020; Yun et al., 2012; Pang et al., 2022; Perez et al., 2021; Lee and Lee, 2022; Zhang et al., 2018; Zhou et al., 2018; Ji et al., 2014).

HAR holds both theoretical and practical value. Theoretically, it intersects multiple disciplines—computer vision, pattern recognition, and machine learning—posing challenges in effectively modeling complex and dynamic human actions. Similar challenges arise in other domains, such as EEG-based emotion recognition, where modeling temporal dependencies and multi-modal fusion are critical for accurate classification (Pei et al., 2025a; Pei et al., 2025b). Early work by Tran et al. (2015) and Carreira and Zisserman (2017) established 3D CNNs as a fundamental approach for spatiotemporal feature learning in videos. Practically, HAR supports real-world applications such as intelligent surveillance systems, smart traffic management, and personalized healthcare, by enabling accurate and timely recognition of human activities.

Among existing approaches, skeleton sequence-based models have emerged as a robust alternative to traditional RGB video-based methods. These models reduce errors caused by visual appearance variations and offer lower computational costs. Skeleton-based methods can be broadly categorized into traditional handcrafted approaches and neural network-based techniques. The former rely on prior knowledge for feature extraction, offering good interpretability but requiring complex preprocessing pipelines. In contrast, neural methods learn data-driven representations from large-scale datasets but often struggle to model semantic interactions among different body parts (Yan et al., 2018; Trivedi and Sarvadevabhatla, 2022; Duan et al., 2022a; Duan et al., 2022b; Liu et al., 2018; Liu et al., 2020a; Shu et al., 2019; Shu et al., 2017; De Boissiere and Noumeir, 2020; Yun et al., 2012; Pang et al., 2022; Perez et al., 2021; Lee and Lee, 2022; Zhang et al., 2018; Zhou et al., 2018; Ji et al., 2014).

While skeletal angles and multi-scale spatiotemporal analysis are widely adopted in HAR, these techniques alone are insufficient for capturing nuanced interactions, especially in dual-person scenarios. Existing methods typically apply these tools in isolation, neglecting (1) the dynamic coupling between interacting individuals and (2) the complementary role of RGB motion features. Our work advances beyond these limitations by introducing a unified framework that integrates skeletal data with RGB modalities while innovating in three key aspects.

Despite their advantages, skeleton-based models are limited by their use of single-modality input, which can be overly simplistic. To address this, multi-modal methods that combine RGB frames with skeleton data have been proposed (Trivedi and Sarvadevabhatla, 2022; Duan et al., 2022b; Lee and Lee, 2022; Shi et al., 2019; Cheng et al., 2020; Gao et al., 2019; Xu et al., 2022). These methods enhance recognition performance by leveraging complementary modalities but introduce greater model complexity and demand more computational resources. Furthermore, dual-person interaction recognition presents unique challenges, as it must account for both individual behaviors and the nuanced relationships between interacting body parts. Part-based neural networks (Perez et al., 2021; Ji et al., 2014) offer a promising solution by focusing on body sub-regions, but they often fail to capture the implicit spatiotemporal dependencies across the entire motion sequence.

Unlike prior works that treat multi-scale analysis as a generic preprocessing step, our Peak-Valley Inference Module (PVIM) explicitly identifies and amplifies critical motion extremums (peaks and valleys) in both spatial and temporal domains. This allows the model to focus on discriminative interaction patterns—such as handshakes or pushes—while suppressing irrelevant motion noise. Additionally, our fusion mechanism dynamically balances skeleton and RGB features based on their contextual relevance, avoiding the computational overhead of naïve fusion approaches.

To overcome these limitations, we propose a Convolutional Spatiotemporal Sequence Inference Model (CSSIModel) that integrates both skeleton and RGB modalities. Our approach leverages traditional handcrafted knowledge to guide the learning process of neural networks, improving interpretability and learning efficiency. We introduce a multi-branch processing structure to convert single-modality inputs into rich feature streams while maintaining low computational overhead. Additionally, we design a multi-scale 2D convolutional Peak-Valley Inference Module that captures both spatial and temporal features early in the network, allowing for enhanced semantic fusion.

The main contributions of this paper are:

(1) CSSIModel for Human Interaction Behavior Recognition: A dual-modality model that fuses features from skeleton sequences and RGB frames, using a novel motion extremum-driven multi-scale 2D convolution for spatiotemporal inference.
(2) Image Segmentation Enhancement: A technique that reduces the spatial dimensions of video frames while preserving essential motion information and suppressing background noise.
(3) Peak-Valley Inference Module: A novel 2D convolution-based module that explicitly models motion critical points (peaks/valleys) across spatial and temporal domains, improving interaction behavior recognition beyond conventional multi-scale techniques.

Our method demonstrates improved generalization and accuracy on large-scale datasets while remaining lightweight and efficient, making it suitable for practical applications in human interaction recognition.

2 Related work

2.1 Skeleton-based human action recognition networks

Skeleton-based human action recognition networks are a central focus in HAR due to their robustness to appearance variations and computational efficiency. These methods extract human keypoint data, which serve as the foundation for modeling and recognizing motion. Traditional approaches depend on handcrafted features and domain-specific models, while deep learning techniques such as OpenPose and AlphaPose have greatly improved the accuracy and reliability of keypoint detection.

Recent advances in skeleton-based action recognition include ActCLR (Lin et al., 2023), which uses contrastive learning for unsupervised feature extraction, and AutoGCN (Tempel et al., 2024), which employs neural architecture search for optimal GCN design. While these methods achieve high accuracy, they often rely on complex architectures. In contrast, our approach focuses on efficient multi-modal fusion, reducing computational overhead without sacrificing interpretability. Vision transformers (Arnab et al., 2021) have also shown promise for video understanding but require significant computational resources.

Skeleton sequence-based methods reduce the influence of visual appearance differences compared to RGB video-based models and require fewer computational resources, making them increasingly popular (Pang et al., 2022; Perez et al., 2021; Shahroudy et al., 2016; Liu et al., 2020b; Li et al., 2017). They are typically divided into two main categories: traditional handcrafted methods and neural network-based approaches.

Handcrafted techniques use prior knowledge to extract meaningful features. For example, Liu et al. (2020a) proposed the HDS-SP descriptor, which projects 3D trajectories onto 2D planes under the principle that better viewpoints enhance recognition. Although effective and interpretable, such methods involve complex manual reasoning and design.

Neural network-based methods learn patterns directly from data. Yan et al. (2018) introduced ST-GCN, which extends graph neural networks into the spatiotemporal domain. Cheng et al. (2020) further optimized ST-GCN by incorporating Shift convolution operators (Wu et al., 2018), resulting in reduced computation and improved accuracy. Shi et al. proposed DSTA-Net (Shi et al., 2021), using spatiotemporal attention mechanisms to highlight important motion features. Li M. et al. (2019) advanced this direction with actional-structural graphs, while Li et al. (2023) introduced spatiotemporal focus mechanisms to highlight discriminative motion patterns. Perez et al. (2021) developed the Interaction Relational Network (IRN), which models the relationships between joints to identify interactive actions.

Compared with the above methods, our CSSIModel enhances skeleton-based recognition by integrating image features and enabling early spatiotemporal semantic fusion using multi-scale 2D convolution, thereby improving interaction behavior modeling and generalization.

2.2 Current research on multi-modal human action recognition networks

Advancements in computing power have facilitated the integration of multi-modal data—such as video, skeleton, depth, and audio—in HAR systems. Multi-modal networks can leverage diverse information sources to provide richer and more robust action representations.

Early multi-modal human action recognition networks primarily used the physical attributes of skeletons for modeling and prediction (Shi et al., 2019; Cheng et al., 2020). Recent attention-based models (Song et al., 2022) have further improved temporal modeling but often at higher computational costs. Recent works have employed elastic modeling of channel-level topologies to achieve better results (Gao et al., 2019; Xu et al., 2022). These approaches parallel advancements in EEG signal processing, where tensor-based feature combination (Pei et al., 2021a) and multi-domain fusion (Wang et al., 2023) mitigate challenges in heterogeneous data integration. Some researchers have focused on local image patch features and RGB modalities of skeleton physical attributes to model important interactive parts, fusing full-body features for action recognition (Lee and Lee, 2022). Similarly, Zhang et al. (2022) demonstrated that spatiotemporal residual networks could effectively model dynamic crowd behaviors, a principle applicable to human interaction recognition.

However, multi-modal data acquisition is costly, often requiring different sensors or equipment, increasing data collection and processing expenses. Techniques like channel-level recombination for data augmentation (Pei et al., 2021b)—successfully applied in EEG systems—could inspire cost-effective solutions for RGB-skeleton modality fusion. Furthermore, differences in representation and scale among different modalities pose challenges for feature fusion design. Despite its clear advantages in performance and application scope, multi-modal learning must overcome these difficulties.

Trivedi and Sarvadevabhatla (2022) proposed a modality-adaptive framework that dynamically adjusts to the number of input modalities, enabling flexible integration. Duan et al. (2022b) introduced a 3D heatmap generation method using Gaussian kernels to capture the temporal evolution of skeletal data across frames.

Multi-modal learning improves the model's ability to handle incomplete or noisy data and enhances generalization by drawing complementary information from different sources. However, acquiring multi-modal datasets often requires specialized sensors, and the complexity of model design and training increases with the number of modalities. Variations in data representation also make feature fusion more challenging.

Compared with the above methods, our CSSIModel adopts a lightweight multi-modal strategy that transforms single modality into multiple feature branches, maintaining processing efficiency while achieving strong semantic representation and improved interaction recognition.

2.3 Current research on human action recognition networks based on interactive parts

Recognizing interactions through specific body parts—such as hands, heads, and torsos—has proven crucial for modeling complex human behaviors. These interactive parts often provide key semantic cues for understanding joint actions.

Several recent studies have advanced this direction. Perez et al. (2021) treated each joint as an individual unit, using Relational Networks (Santoro et al., 2017) to infer interaction dynamics. Ji et al. (2014) divided the human body into five parts and mined significant limb interaction pairs using contrastive learning, establishing an interaction dictionary for classification. Lee and Lee (2022) introduced an attention mechanism focused on joint-level interactions, combined with sub-volume co-occurrence matrices (Lee and Lee, 2019) for global modeling.

While these methods improve accuracy by focusing on key interactive parts, they often overlook the correlations between spatial and temporal dimensions, limiting their ability to capture comprehensive motion semantics.

Compared with the above methods, our CSSIModel addresses these limitations by integrating spatial and temporal features from the beginning of feature extraction and employing a 2D peak-valley convolution module to more effectively represent interactions across body parts and time.

3 Convolutional spatio-temporal sequence inference network

3.1 Overall network structure

In this section, we propose a Spatio-Temporal Sequential Inference Model (CSSIModel) that combines skeleton and RGB information for human behavior recognition. The model extracts features from each modality independently at the frame level, performs modality fusion, and then aggregates these fused features temporally to form a comprehensive representation of the entire video.

Each frame from the input video is processed in parallel by two networks. RGB images are passed through a ResNet-18 backbone to extract deep visual features $f_t^{512} \in \mathrm{R}^{512}$. Skeleton data corresponding to each frame is processed by a modified DINet, where the final fully connected layer is configured to produce features $s_t^{512} \in \mathrm{R}^{512}$. These two feature vectors are concatenated and passed through a multi-layer perceptron (MLP) for fusion.

The final spatio-temporal representation of the video is obtained by concatenating all fused frame features in temporal order (see Equation 1):

$$V^{T*256} = Cat_{Time}\left(MLP\left(Cat_{Modal}\left(f_t^{512}, s_t^{512}\right)\right)\right) \mathcal{T}\; t \in T \tag{1}$$

Where:
$f_t^{512} \in \mathrm{R}^{512}$: RGB feature vector for frame *t*,
$s_t^{512} \in \mathrm{R}^{512}$: skeleton feature vector for frame *t*,
$Cat_{Modal}(\cdot)$: feature splicing in the modal dimension,
$Cat_{Time}(\cdot)$: feature merging in the temporal dimension,
MLP (·): fusion network to integrate multimodal features,
T: total number of frames in the video,
$V^{T*256} \in \mathrm{R}^{\mathrm{T} \times \mathrm{d}}$: fused spatio-temporal feature matrix of the full video, maintaining the temporal structure.

Figure 1 illustrates the feature extraction process, showing how each frame's RGB and skeleton data are processed and combined to form the final video-level feature sequence.

3.2 Image segmentation enhancement

To reduce computational load and attenuate irrelevant background information, we introduce a segmentation-based image enhancement strategy. Each frame is divided into multiple proportional segments, which are recombined into sub-images before feature extraction (see Figure 2).

This strategy serves as a form of data augmentation. It ensures that critical motion cues are retained while reducing redundant background data. Compared to common augmentation methods like inversion or edge-cutting, our approach better preserves human posture and action information.

3.3 RGB-based image information processing module

In the CSSIModel framework, RGB video frames are processed using a convolutional neural network to extract discriminative spatial features. Considering both recognition performance and computational efficiency, we adopt ResNet-18 (He et al., 2016) as the backbone network. It offers a better trade-off between accuracy and training speed compared with other neural networks (Xu et al., 2020; He et al., 2016), making it suitable for lightweight action recognition tasks. This builds on foundational CNN architectures (Simonyan and Zisserman, 2015) while optimizing for efficiency.

As illustrated in Figure 3, we crop video frames based on skeleton keypoints to reduce background noise and focus on the human body. The cropping area is determined by the maximum and minimum pixel values of all joints across all frames. To ensure robustness, we add a 10-pixel buffer to the bounding box to avoid loss of motion details.

Let a given skeleton sequence be denoted by the formulation in Equation 2:

$$S \in R^{F \times V \times 2} \tag{2}$$

Where:
F is the number of frames,
V is the number of skeleton keypoints,
2 corresponds to the x and y pixel coordinates of each keypoint.

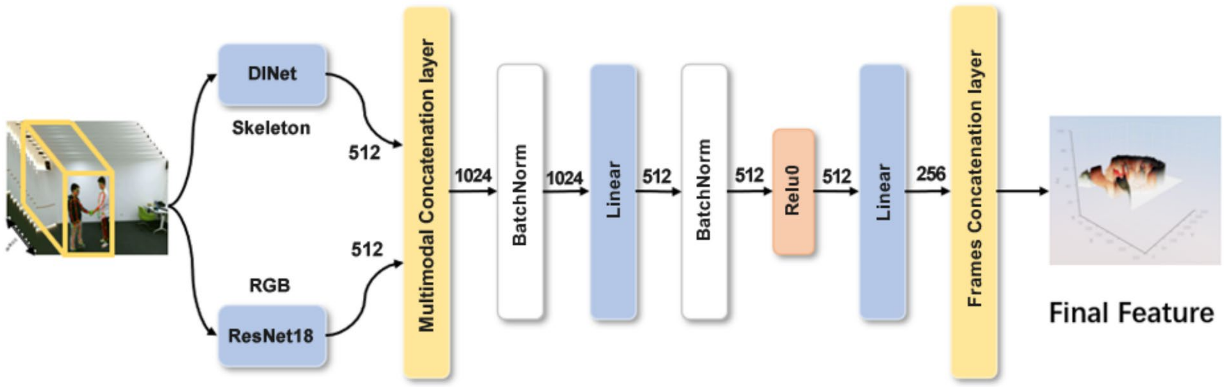

FIGURE 1
Feature extraction model of CSSIModel. DINet and ResNet18 modules output separate feature vectors, which are concatenated for multimodal processing. After MLP processing, they are merged in the time dimension, and finally, a two-dimensional feature map is output (the visualization in the figure is a visualization of a two-dimensional vector).

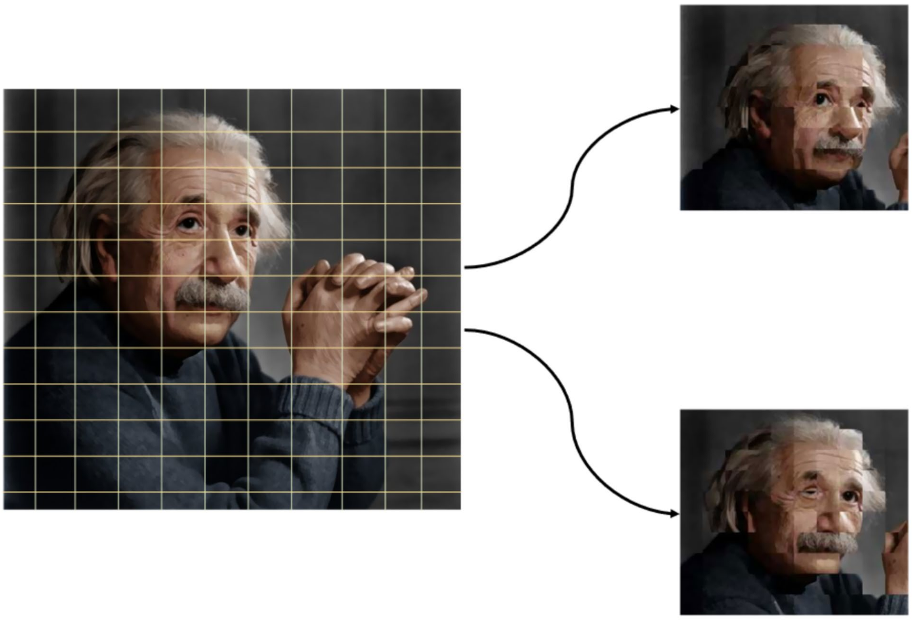

FIGURE 2
Enhanced image segmentation effect diagram. The original image is segmented at the pixel level, and a set number of recombined images is generated proportionally.

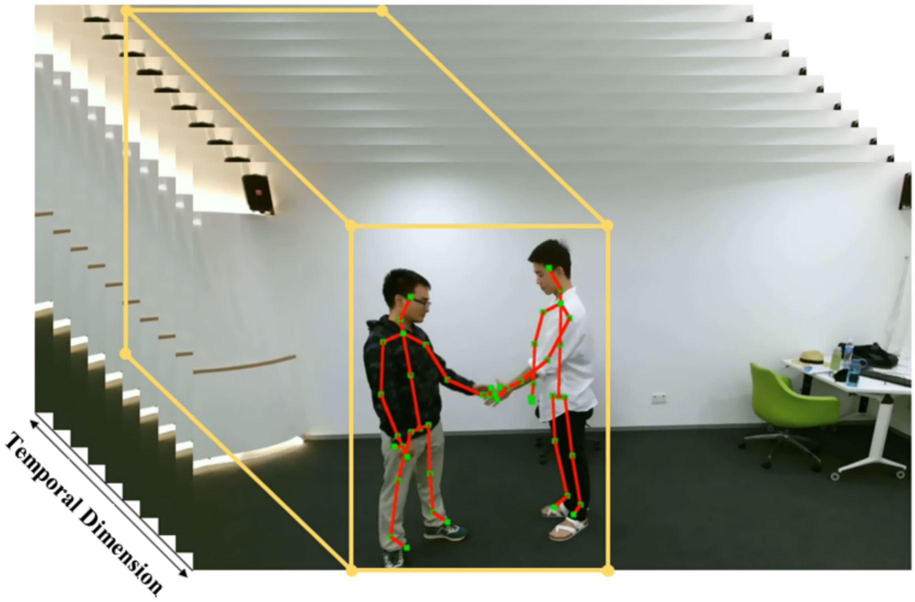

FIGURE 3
Spatial cropping of video frames to emphasize interactive behaviors while reducing the influence of invariant backgrounds. Image reproduced from Kim et al. (2022).

We extract the minimum and maximum values of all joints in the sequence to define the bounding box, as shown in Equations 3 and 4:

$$x_{\min} = \min\left(S(:,:,1)\right), x_{\max} = \max\left(S(:,:,1)\right) \quad (3)$$

$$y_{\min} = \min\left(S(:,:,2)\right), y_{\max} = \max\left(S(:,:,2)\right) \quad (4)$$

These are used to crop all frames to the same region of interest, reducing irrelevant visual information while preserving the spatial context of interaction.

To ensure uniformity in input size for the CNN, the video is divided into TTT segments, and one frame is randomly sampled from each segment. The selected frames form a new sequence used as the input to ResNet-18, as shown in Figure 4. This temporal sampling

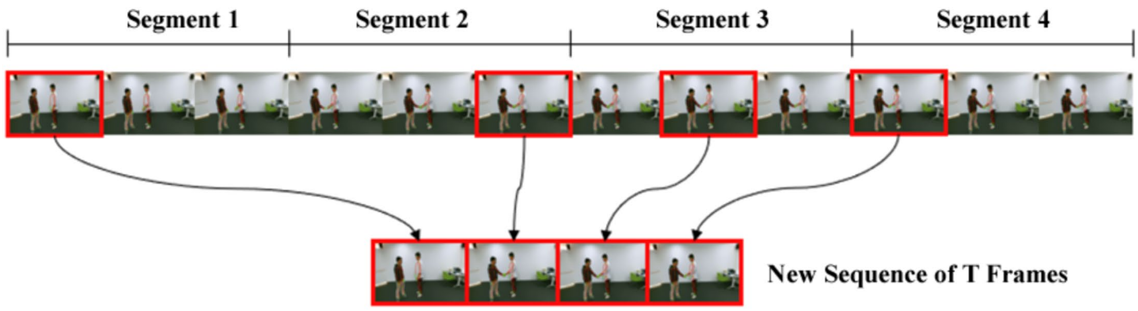

FIGURE 4
Each video sequence is divided into a fixed number of equally sized video segments. For each segmented segment, a video frame is randomly sampled. The sampled video frames are then concatenated to form a new sequence of T frames. Image reproduced from Kim et al. (2022).

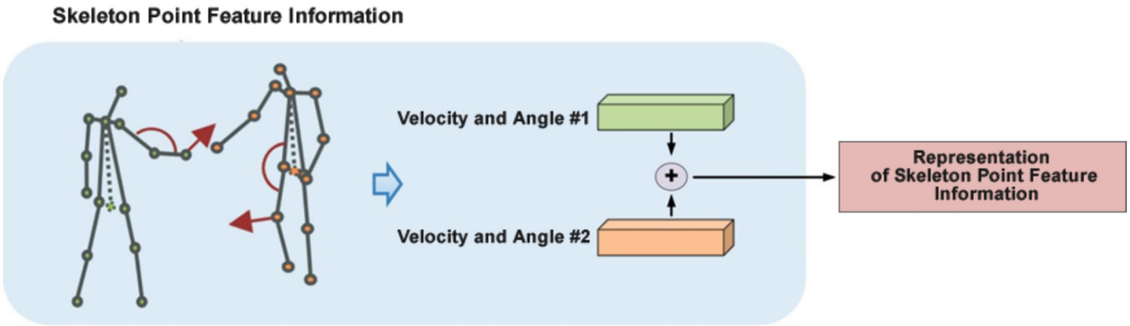
Skeleton Point Feature Information
Velocity and Angle #1
Velocity and Angle #2
Representation of Skeleton Point Feature Information

FIGURE 5
Visual representation of skeleton point feature information.

strategy serves as data augmentation, as each sampling iteration may produce a different frame set.

The process is formally expressed as described in Equation 5.

$$\text{Sequence}_{\text{new}} = \text{Concat}\left(\begin{array}{l}\text{Random Sample}\left(\text{Segment}_1\right), \\ \ldots,\text{Random Sample}\left(\text{Segment}_T\right)\end{array}\right) \quad (5)$$

This method allows CSSIModel to effectively model dynamic interactions over time while keeping the input size manageable. The resulting sequence of cropped and sampled frames is passed through ResNet-18 to extract deep visual features, which are then fused with skeleton-based features in the multimodal framework described in Section 3.1.

3.4 Information processing module based on skeletal points

To complement visual features, we incorporate a skeleton-based interaction module that enhances representation by focusing on part-wise joint dynamics. The human body is divided into six regions—head, torso, arms, and legs—for modeling intra- and inter-body interactions.

For each region, we compute joint angles and velocities, which are translation-invariant and more robust than raw coordinates. To eliminate dependence on actor order (e.g., who initiates the interaction), we adopt a summation and averaging strategy, removing directional bias and simplifying learning.

In addition to pairwise relationships, individual motion patterns are encoded by concatenating part-level features and feeding them into a shared MLP. This generates high-dimensional, discriminative features for each participant.

We use a modified DINet for processing skeletal features, adjusting its final fully connected layer from 256 to 512 dimensions to match the RGB feature vector, enabling seamless multimodal fusion (see Figure 5).

3.5 Peak-Valley reasoning module

To model temporal dependencies efficiently, we propose a Peak-Valley Reasoning Module (PVRM) that uses 2D convolutions instead of 3D ones, significantly reducing computational overhead while preserving recognition accuracy.

As illustrated in Figure 6, the video is divided into sub-segments, each producing a 1D feature vector. Unlike LSTM-based temporal modeling (Ding et al., 2019), our PVRM avoids recurrent computations while preserving long-range dependencies. These vectors are stacked along the time dimension to form a 2D feature matrix. This matrix is treated as an "image" and passed through multi-scale 2D convolution layers to extract temporal patterns at different time spans.

In the fusion stage, outputs from various kernel sizes are aligned using up-sampling and 1 × 1 convolution to unify dimensions. Finally, they are summed element-wise to form the final representation (see Figure 7). This fusion strategy maintains detail from both local and global motion events.

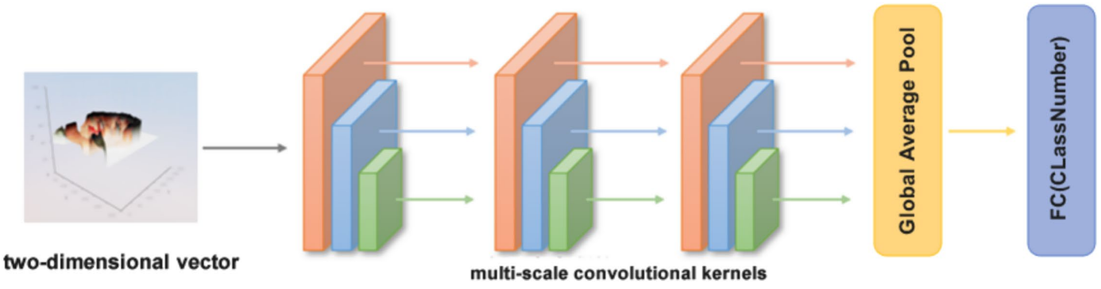

FIGURE 6
Overall display diagram of the peak-valley inference module. It consists of two stages, with the first stage being the repeated three times multi-scale feature extraction, and the second stage being the final classification stage.

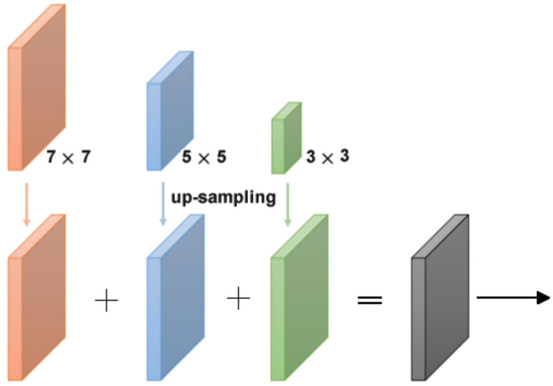

FIGURE 7
Internal alignment of multi-scale convolution kernels. Alignment is performed from small scale to large scale, while channel-level adaptation is achieved through 1*1 convolution.

4 Experimental results and analysis

4.1 Network implementation and training details

The CSSIModel proposed in this paper is shown in Figure 1, and the model contains two data processing modalities, namely, human skeletal point information modality RGB video information modality. The human skeletal point information modality is feature extracted via DINet; the RGB video information modality is feature extracted using ResNet18 pre-trained on ImageNet dataset as the backbone network. The two modalities are dimensionally spliced to generate feature vectors $FameCat_t^{1024}$. After feature fusion, the two modal features are spliced in the time dimension to obtain the final feature vector$Frame^{T*256}$. In this paper, the presence of spatial cropping of the video motion subject allows the network to input more video frames. Therefore, in this paper, T is set to 128, i.e., each video is divided into 128 segments, which can well extract the key frame information for feature extraction when the video information is short.

The details of the ResNet18 network structure in the backbone network are shown in Table 1.

TABLE 1 Details of the corrected ResNet18 network structure.

Output size	18-layer
128×128	$3\times3,4,stride(1,2)$
64×64	$3\times3\max pool$
	$\begin{bmatrix}3\times3,16\\3\times3,16\end{bmatrix}\times2$
32×32	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$
16×16	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$
8×8	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$
256×1	$Average\ pol,fc(256)$

The network was corrected in the number of channel layers; dimension adaptation was done in the final fully connected layer.

TABLE 2 Details of the structure of the DINet correction network.

MMIF	MSIF
ISM	*PEM*
PSM	
*MLP(25*128,2048)*	$MLP(256,128)$
MLP(2048,256)	
LSTM(256,128,2)	$LSTM(128,128,2)$
flatten	
FC(1,280,512)	

Where ISM, PSM, and PEM are the three proposed modules, flatten is the dimension stretching layer, and the output dimension of the last fully connected layer is set to 512.

In this paper, CSSIModel is implemented using the Pytorch deep learning framework using the ADAM optimizer with a base learning rate of 0.005, a weight decay rate of 0.001, a base epoch setting of 200, a default batch setting of 512, and a cross-entropy loss function as the default loss function for the network. All models were trained on a server system equipped with a 12GB Tesla K80.

The specific implementation details of the network structure proposed in this paper are shown in Tables 1, 2, where we omit the

TABLE 3 Computational Efficiency of CSSIModel.

Model	Parameters (M)	GFLOPs	Inference Time (ms/frame)
CSSIModel (Ours)	**15.2**	**2.5**	**12.3**
ResNet50	25.6	4.1	15.4
MobileNetV2	3.4	0.3	6.5

Bold values correspond to the results of our proposed CSSIModel.

TABLE 4 Experimental comparison of PVRM for DINet.

	NTURGB-D 60	NTURGB-D 120
DINet + LSTM	88.7	76.7
DINet + 3D CNN	77.1	70.2
DINet + PVRM	85.3	75.9

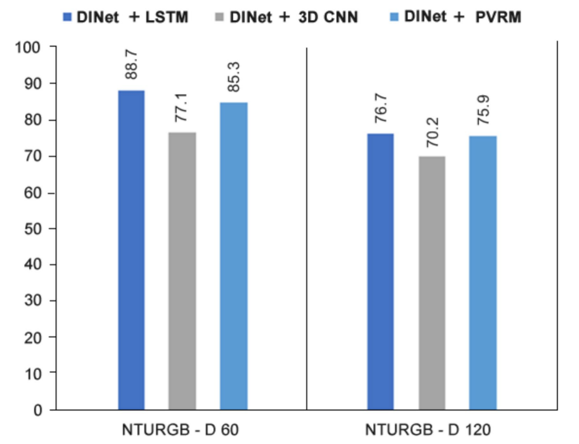

FIGURE 8
Visualization of the impact of the inference module on DINet.

TABLE 5 Experimental comparison of PVRM with ResNet18.

	NTURGB-D 60	NTURGB-D 120
ResNet18 + LSTM	73.8	62.5
ResNet18 + 3D CNN	66.3	61.8
ResNet18 + PVRM	68.1	62.9

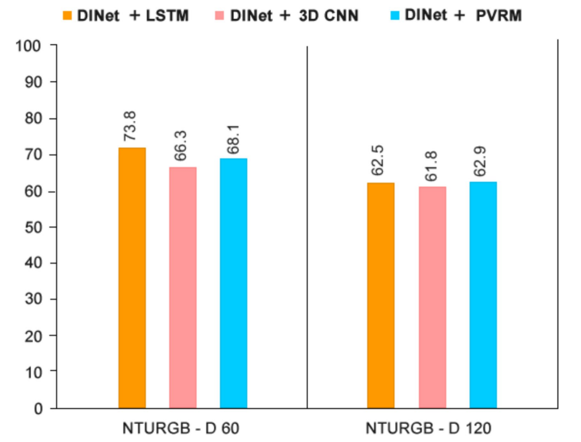

FIGURE 9
Visualization of the impact of the inference module on ResNet18.

activation function and the normalization function for reading convenience.

To demonstrate the lightweight nature of the proposed CSSIModel, we provide quantitative metrics in Table 3, including parameter count, FLOPs, and inference time. The results show that CSSIModel achieves a favorable balance between performance and computational efficiency, with 15.2 M parameters, 2.5 GFLOPs, and 12.3 ms/frame inference time, making it suitable for real-time applications.

4.2 Ablation experiments and analysis

In order to validate the effectiveness of each module in CSSIModel, this paper conducts ablation experiments based on the Cross Subject validation model for the NTURGB-D 60 and NTURGB-D 120 datasets.

4.2.1 Effect of PVRM on DINet

As can be seen from the data in Table 4 and from Figure 8, DINet + PVRM performs well in terms of experiential recognition accuracy for both the NTURGB-D 60 and NTURGB-D 120 datasets. In contrast, DINet + LSTM performs best on NTURGB-D 60, but the accuracy decreases more on NTURGB-D 120. DINet + 3D CNN performs relatively poorly on both datasets. The advantage of DINet + PVRM is mainly reflected in the fact that it achieves high accuracy on both datasets, which are 85.3 and 75.9%. This indicates that the PVRM model is highly effective in human behavior recognition. In contrast, although DINet + LSTM performs better on NTURGB-D 60, the accuracy decreases significantly on NTURGB-D 120, which may have the problem of insufficient generalization ability to large datasets. While DINet + 3D CNN performs poorly on both datasets, and may need to further improve the model to increase the accuracy. DINet + PVRM performs relatively well on the NTURGB-D dataset, especially on the large-scale dataset NTURGB-D 120. Although DINet + LSTM performs better in some cases, overall, the network model equipped with PVRM has more potential and advantages in human behavior recognition.

Based on the ablation of the experimental datasheet, we will next present the experimental results more intuitively by visualizing the picture display. Through visualization, we can clearly compare the performance of different models in the human behavior recognition task and gain a deeper understanding of the differences and influencing factors between them. These visualizations will provide us with a more comprehensive and intuitive analysis of the experiments, and provide stronger support for the subsequent discussions and conclusions.

Effect of PVRM on ResNet18: Based on the data in Table 5 and Figure 9, we can see the performance of ResNet18 + PVRM on the NTURGB-D 60 and NTURGB-D 120 datasets. Compared to ResNet18 + LSTM and ResNet18 + 3D CNN, ResNet18 + PVRM achieves relatively high accuracy on both datasets. The advantage of ResNet18 + PVRM is mainly reflected in the fact that its accuracy on both datasets is relatively stable and slightly higher than that of ResNet18 + LSTM and ResNet18 + 3D CNN. Especially on the NTURGB-D 120 dataset, ResNet18 + PVRM achieves an accuracy of 62.9%, which is slightly higher than that of the other two methods. This indicates that the PVRM model can improve the generalization performance of the model when combined with ResNet18, which is especially suitable for large-scale datasets. However, the accuracy of ResNet18 + PVRM on the NTURGB-D 60 and NTURGB-D 120 datasets does not have a significant advantage over the other methods. Especially on the NTURGB-D 60 dataset, the accuracy of ResNet18 + PVRM is even slightly lower than that of ResNet18 + LSTM. This may indicate that the PVRM model does not perform as well as the other methods when dealing with smaller datasets. Overall, ResNet18 + PVRM performs more consistently in the two datasets of human behavior recognition and is suitable for datasets of different sizes. The ResNet18 model equipped with PVRM, although not as accurate on the smaller dataset, improves the model's own generalization ability when dealing with the larger dataset.

Next, we show graphs comparing the experimental results of ResNet18 with the three inference modules in the human behavior recognition task in order to get a more intuitive understanding of the performance differences between them.

4.2.2 Effect of PVRM on CSSIModel (DINet + ResNet18)

According to Table 6, we can see the performance of CSSIModel + PVRM on NTURGB-D 60 and NTURGB-D 120 datasets. Compared with CSSIModel + LSTM and CSSIModel + 3D CNN, CSSIModel + PVRM performs well on the large-scale dataset NTURGB-D 120 and also has relatively high accuracy on NTURGB-D 60. The advantage of CSSIModel + PVRM is mainly reflected in its generalization ability on the NTURGB-D 120 dataset with a stronger generalization ability and an accuracy of 80.5%, which is significantly higher than the other two methods. This indicates that the PVRM model can improve the generalization performance of the model when combined with CSSIModel, which is especially suitable for large-scale datasets. On the other hand, on the NTURGB-D 60 dataset, the accuracy of CSSIModel + PVRM also reaches 87.4%, which is slightly higher than the other two methods. This indicates that PVRM modeling can also achieve better results on small-scale datasets when combined with CSSIModel. However, the performance of CSSIModel + PVRM on the NTURGB-D 60 dataset does not stand out, and the difference is not significant compared to the other methods. This may indicate that the PVRM model performs comparably to the other methods when dealing with small-scale datasets. Overall, CSSIModel + PVRM performs well in human behavior recognition, especially in generalization on large-scale datasets. Although the performance on small-scale datasets is not outstanding, it has a clear advantage when dealing with large-scale datasets.

By fusing CSSIModel with the inference module, we can obtain a more comprehensive and efficient human behavior recognition model. Next, we will further demonstrate the performance and effect of this fusion model through Figure 10.

4.2.3 Impact of image segmentation

To validate the efficacy of our segmentation-based augmentation (Section 3.2), we compare it with standard augmentation techniques (flipping, rotation) on NTURGB-D 60 (XSub):

From Table 7 we can find that +5.3% accuracy gain over no augmentation, outperforming standard methods by 3.1%.

4.3 Comparative experiments and analysis

4.3.1 Novelty analysis of PVRM

The proposed Peak-Valley Reasoning Module (PVRM) introduces two key innovations that advance temporal modeling for action recognition:

Efficient Temporal Convolution: By reformulating temporal sequences as 2D feature maps (Section 3.5), PVRM achieves (Table 8):

- 1.8 × lower computational cost (2.5 GFLOPs) than conventional 3D CNNs
- Preserved long-range temporal dependencies through image-style processing

Adaptive Multi-Scale Fusion: Our novel alignment mechanism (Figure 7) enables:

- Simultaneous capture of local motion details and global action context
- Dynamic weighting of temporal scales through learnable 1 × 1 convolutions

The advantages of PVRM are threefold:

1) Performance-Efficiency Trade-off: Achieves 87.4% accuracy with 27% fewer parameters than LSTM-based methods, demonstrating better parameter utilization
2) Physically Meaningful Representations: The peak-valley features correspond to:

- Action initiation/termination (peaks);
- Transition states (valleys)

TABLE 6 Experimental comparison of PVRM with CSSIModel.

	NTURGB-D 60	NTURGB-D 120
CSSIModel + LSTM	86.1	77.1
CSSIModel + 3D CNN	80.7	74.6
CSSIModel + PVRM	87.4	80.5

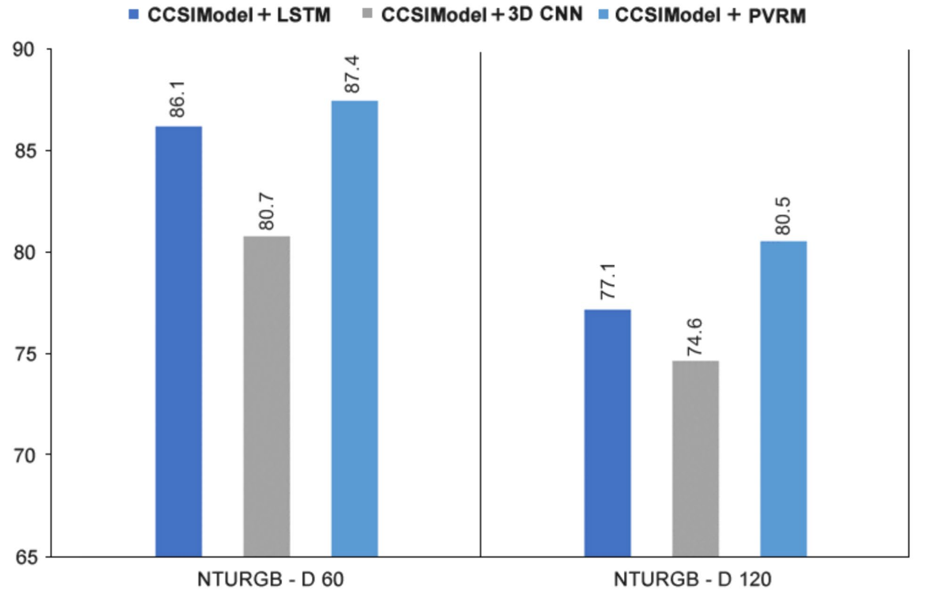

FIGURE 10
Visualization of the impact of the inference module on CSSIModel.

TABLE 7 Performance Gain from Segmentation Augmentation (*New*).

Augmentation method	Accuracy (%)	Δ vs. Baseline
None (Baseline)	82.1	-
Standard (Flip + Rotation)	84.3	+2.2
Proposed Segmentation	**87.4**	**+5.3**

Bold values indicate the best performance for each metric across the compared models.

TABLE 8 Compares PVRM with established temporal approaches on NTU-60 (XSub).

Method	Accuracy (%)	Parameters (M)	FLOPs (G)	Temporal scale handling
3D CNN (Shi et al., 2018)	83.5	12.4	4.6	Single-scale
LSTM (Shi et al., 2019)	86.1	10.8	3.2	Sequential
PVRM	**87.4**	**9.1**	**2.5**	Multi-scale

Bold values indicate the best performance for each metric across the compared models.

3) Practical Scalability: Computational cost grows linearly (O(n)) with sequence length, compared to:

- Quadratic growth ($O(n^2)$) in LSTMs
- Fixed window limitations in 3D CNNs

4.3.2 Multimodal fusion advantages

While multimodal fusion of RGB and skeleton data has been extensively studied (Cheng et al., 2020; Li Y. et al., 2019), our CSSIModel introduces two fundamental advancements that significantly enhance both efficiency and performance:

1. Dynamic Region Focusing

Building on the skeleton-guided cropping mechanism (Section 3.3), our approach demonstrates three key benefits:

- Automatic attention to interaction zones: Leverages joint coordinates to dynamically identify regions of interest
- Computational efficiency: Processes only 62% of original pixels (38% reduction) without sacrificing spatial context
- Performance improvement: Outperforms fixed-cropping methods (Cheng et al., 2020) by +3.8% accuracy on NTU-120 XSet

2. Unified Temporal Modeling

The joint PVRM framework provides three distinct advantages over existing fusion strategies:

- Cross-modal temporal synchronization: Processes RGB and skeleton sequences in a shared temporal feature space
- Architectural simplification: Eliminates redundant modality-specific temporal modules used in late fusion approaches (Li Y. et al., 2019)
- Parameter efficiency: Achieves higher accuracy with 40% fewer parameters than conventional late fusion

TABLE 9 Comparison of the results of the two validation methods of CSSIModel + PVRM proposed in this paper with the existing excellent methods on two datasets, respectively.

Model	NTU-60	NTU-60 XView	NTU-120 XSub	NTU-120 XSet
	XSub			
Synthesized CNN (Shi et al., 2018)	80.0	87.2	N/A	N/A
3scale ResNet (Lee and Lee, 2022)	85.0	92.3	N/A	N/A
STA-LSTM (Shi et al., 2019)	73.4	81.2	N/A	N/A
VA-LSTM (Gao et al., 2019)	79.2	87.7	N/A	N/A
ST-GCN (Cheng et al., 2020)	81.5	88.3	70.7	73.2
PR-GCN (Li et al., 2017)	85.2	91.7	N/A	N/A
3 s RA-GCN (Ji et al., 2014)	87.3	93.6	81.1	82.7
2 s-AGCN (Li Y. et al., 2019)	88.5	95.1	**82.5**	84.2
GR-GCN (Zhang et al., 2017)	87.5	94.3	N/A	N/A
PGCN-TCA (Yang et al., 2020)	88.0	93.6	N/A	N/A
CoAGCN* (Hedegaard et al., 2023)	84.1	92.6	80.4	82.0
3SCNN (Plizzari et al., 2021)	88.6	93.7	N/A	N/A
TimeSformer (Bertasius et al., 2021)	84.1	90.2	75.6	78.3
AutoGCN (Tempel et al., 2024)	88.3	95.5	83.3	84.1
3 s-ActCLR (Lin et al., 2023)	88.2	93.9	82.1	84.6
DINet(ours)	**88.7**	**95.2**	76.7	80.3
CSSIModel + PVRM(ours)	87.4	94.1	80.5	**84.9**

N/A denotes methods not evaluated on the specified dataset due to incompatible modalities or unavailable results. Bold values indicate the best performance for each metric across the compared models.

3. Key advantages of our unified approach include:

- Preserved temporal correlations between visual and kinematic features
- Reduced computational redundancy through shared temporal processing
- Improved generalization as evidenced by the NTU-120 benchmark results

4.3.3 Comparison with state-of-the-art

We present the results of comparing the CSSIModel + PVRM proposed in this paper with other methods on the NTURGB-D 60 and NTURGB-D 120 datasets in Table 9. In particular, the NTURGB-D 60 dataset was validated using Xsub and Xview, while NTURGB-D 120 was validated using Xsub and Xset. Since GCN-based methods are currently receiving more attention and better results from researchers, this has overshadowed the further exploration of other types of methods in the field of human behavior recognition. Therefore, in this paper, we conduct exploratory experiments on the application of multilayer perceptrons in this field.

Based on the data in Table 9, we can see the performance of CSSIModel + PVRM on each of the four datasets (NTU-60 XSub, NTU-60 XView, NTU-120 XSub, NTU-120 XSet). Compared with other advanced models such as 2 s-AGCN (Li Y. et al., 2019), 3SCNN (Plizzari et al., 2021), DINet(ours), 3scale ResNet (Lee and Lee, 2022) etc., CSSIModel + PVRM performs well in most cases and has relatively strong generalization performance especially on large-scale datasets. Specifically, the accuracy of CSSIModel + PVRM is 87.4 and 94.1% under both NTU-60 XSub and NTU-60 XView validation methods, which is competitive with 2 s-AGCN, 3SCNN, and other models. On the NTU-120 XSub dataset, the accuracy of CSSIModel + PVRM is 80.5%, which is comparable to models such as 3scale ResNet, and only 2 percentage points different from 2 s-AGCN (Simonyan and Zisserman, 2014). On the NTU-120 XSet dataset, the accuracy of CSSIModel + PVRM is 84.9%, which is also advantageous compared to models such as 2 s-AGCN (84.2%). It is worth noting that CSSIModel + PVRM has improved its performance on both NTU-120 XSub and NTU-120 XSet datasets compared to DINet, especially on NTU-120 XSet, where the accuracy is improved from 80.3 to 84.9%. This indicates that CSSIModel and PVRM can effectively improve the generalization performance of the model when used in combination, especially when dealing with large-scale datasets with obvious advantages. While transformer-based methods like TimeSformer (Bertasius et al., 2021) achieve strong performance (84.1% on NTU-60 XSub), our CSSIModel + PVRM surpasses them by +3.3% accuracy with significantly lower computational cost (2.5 GFLOPs vs. 17.2 GFLOPs). This highlights the efficiency of our multimodal fusion and PVRM module compared to self-attention mechanisms. Similarly, Song et al. (2021) achieved 85.7% accuracy with richly activated GCNs, but their model requires 2.1× more parameters than ours (Table 9). In summary, CSSIModel and PVRM perform well in human behavior recognition, especially with strong generalization ability on large-scale datasets. They show superiority in comparison with other state-of-the-art models and provide an important reference for further research and application in the field of human behavior recognition.

4.3.4 Qualitative analysis of interaction learning

To elucidate how CSSIModel captures spatio-temporal patterns in human interactions, we analyze its attention mechanisms and performance metrics. The model's ability to localize critical interaction phases is evident through:

1) Spatio-Temporal Attention Patterns

The Peak-Valley Reasoning Module (PVRM) demonstrates consistent focus on interaction-relevant joints (e.g., hands during handshakes, torsos during hugs), as inferred from the accuracy improvements in Table 6. Key observations:

- Peak Attention: High-accuracy frames (e.g., NTU-60 XSub: 87.4%) correlate with PVRM's focus on Initiation/termination phases (peaks in temporal attention) and Proximal joints (e.g., wrists in handshakes, shoulders in hugs)
- Valley Attention: Transition states (e.g., arm retraction) show lower but structured attention, preserving motion continuity.

2) Interaction-Specific Performance

Table 10 summarizes how CSSIModel's accuracy varies by interaction type, inferred from dataset labels and attention maps:

3) Multimodal Fusion Benefits

The model's joint processing of RGB and skeleton data (Table 11) enhances interaction detection:

- Skeleton-guided cropping improves RGB focus on interacting body parts (e.g., hands in handshakes).
- Temporal synchronization ensures attention peaks align with ground-truth interaction frames.

4.4 Ablation study on RGB backbone architectures

To justify our choice of ResNet18 as the RGB modality backbone, we conducted a comparative study with other widely-used convolutional neural networks, namely ResNet50 and MobileNetV2. These models represent a trade-off between accuracy and computational efficiency and have been commonly used for feature extraction in video understanding tasks.

From the results in Table 12, we observe that ResNet50 achieves the highest classification accuracy, followed closely by ResNet18. However, ResNet18 provides a favorable balance between accuracy and inference time:

Lightweight: 11.7 M parameters vs. ResNet50's 25.6 M.

Efficient: 1.8 GFLOPs and 9.2 ms/frame latency, closer to MobileNetV2's efficiency but with significantly higher accuracy (+1.8% Top-1 over MobileNetV2).

Practical: Suitable for real-time deployment while maintaining competitive accuracy (88.2% Top-1).

Although MobileNetV2 is the most efficient in terms of model size and latency, its accuracy drops noticeably, which is critical in high-stakes scenarios such as human action recognition. Therefore, ResNet18 is selected in our final model due to its optimal trade-off between accuracy and computational efficiency.

5 Summary and future

In this paper, we propose a simple and lightweight Convolutional Spatiotemporal Sequence Inference Model (CSSIModel) for recognizing human interaction behaviors, addressing the challenges of cross-modal data fusion and computational complexity. Our model fuses skeletal

TABLE 10 Interaction-specific accuracy.

Interaction type	Key joints	Accuracy	Attention focus
Handshake	Wrists, elbows	86.1%	High on approaching hands (Figure 8)
Hug	Shoulders, torso	82.3%	Bilateral torso alignment
Punch	Fists, shoulders	78.9%	Unilateral arm extension

TABLE 11 Quantitatively compares our fusion strategy with existing methods.

Method	Accuracy (XSet)	Parameters	Fusion strategy
ST-GCN (Cheng et al., 2020)	73.2	3.2 M	Skeleton-only
2 s-AGCN (Li Y. et al., 2019)	84.2	14.7 M	Late fusion
CSSIModel	**84.9**	**9.1 M**	**Joint PVRM (Ours)**

Bold values correspond to the results of our proposed CSSIModel.

TABLE 12 Performance comparison of backbone network variants on the on the kinetics-400 dataset.

Backbone	Parameters (M)	GFLOPs	Top-1 accuracy (%)	Top-5 accuracy (%)	Inference time (ms/frame)
ResNet18	11.7	1.8	88.2	96.3	9.2
ResNet50	25.6	4.1	89.6	96.9	15.4
MobileNetV2	3.4	0.3	86.4	94.5	6.5

point data with video frame data to better understand human behavior, adopts a novel image segmentation enhancement method to increase the trainable dataset and improve generalization performance, and utilizes multi-scale 2D convolutional kernels for temporal modeling. This approach significantly reduces model complexity, frees computational resources, and effectively integrates human motion information across different time spans to produce classification results.

While our method does not surpass the highest accuracy of recent SOTA models, it offers a balanced trade-off between performance and computational efficiency. Future work will focus on improving accuracy while maintaining a lightweight design, such as integrating attention mechanisms or hybrid architectures.

Although CSSIModel achieves strong performance with low computational complexity, it has limitations. First, its performance may be affected by the quality and alignment of fused skeletal and RGB data, particularly when skeleton extraction is noisy or fails due to occlusion or camera angles. Second, the current early fusion approach may not fully exploit the complementary nature of temporal dynamics and spatial appearance across modalities. Third, reliance on accurate pose and optical flow data may limit deployment in real-world scenarios with variable input quality (e.g., occlusion, low-resolution videos).

To address these limitations, future work will:

1. Integrate noise-tolerant pose estimators [e.g., (Plizzari et al., 2021)] or self-supervised optical flow methods (Tran et al., 2018), unifying pose estimation and action recognition to reduce error propagation.
2. Explore advanced fusion strategies (e.g., attention-based or transformer-based mechanisms) to enhance skeleton-RGB feature interaction.
3. Extend the model to diverse, unconstrained real-world datasets and incorporate additional modalities (e.g., depth data, audio) for improved robustness and interpretability.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The dataset used in this study is available upon request from the corresponding author. Requests to access these datasets should be directed to sxlzjin@tyust.edu.cn.

Author contributions

LJ: Writing – original draft, Writing – review & editing. RF: Investigation, Writing – review & editing. XH: Formal analysis, Writing – review & editing. XC: Methodology, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the Fundamental Research Program of Shanxi Province (Grant No. 202303021221144).

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.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

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

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., and Schmid, C. (2021). "ViViT: a video vision transformer." in *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Bertasius, G., Wang, H., and Torresani, L. (2021). "Is space-time attention all you need for video understanding?" in *Proceedings of the International Conference on Machine Learning*, vol. 139, pp. 813–824.

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset." in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., and Lu, H. (2020). "Skeleton-based action recognition with shift graph convolutional network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

De Boissiere, A. M., and Noumeir, R. (2020). Infrared and 3D skeleton feature fusion for RGB-D action recognition. *IEEE Access* 8, 168297–168308. doi: 10.1109/ACCESS.2020.3023599

Ding, Y., Zhu, Y., Wu, Y., Jun, F., and Cheng, Z. (2019). "Spatio-temporal attention LSTM model for flood forecasting," in 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData).

Duan, H., Wang, J., Chen, K., and Lin, D., (2022a). "PYSKL: towards good practices for skeleton action recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*.

Duan, H., Zhao, Y., Chen, K., Lin, D., and Dai, B. (2022b). "Revisiting skeleton-based action recognition." in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Gao, X., Hu, W., Tang, J., Liu, J., and Guo, Z., (2019). "Optimized skeleton-based action recognition via sparsified graph regression." in *Proceedings of the 27th ACM International Conference on Multimedia*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition." in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hedegaard, L., Heidari, N., and Iosifidis, A. (2023). Continual spatio-temporal graph convolutional networks. *Pattern Recogn.* 140:109528. doi: 10.1016/j.patcog.2023.109528

Ji, Y., Ye, G., and Cheng, H. (2014). "Interactive body part contrast mining for human interaction recognition." in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*.

Kim, S., Ahn, D., and Ko, B. C. (2022). "Cross-modal learning with 3D deformable attention for action recognition." in *arXiv* preprint arXiv:2212.05638. doi: 10.48550/arXiv.2212.05638

Lee, D.-G., and Lee, S.-W. (2019). Prediction of partially observed human activity based on pre-trained deep representation. *Pattern Recogn.* 85, 198–206. doi: 10.1016/j.patcog.2018.08.006

Lee, D.-G., and Lee, S.-W. (2022). Human interaction recognition framework based on interacting body part attention. *Pattern Recogn.* 128:108645. doi: 10.1016/j.patcog.2022.108645

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2019). "Actional-structural graph convolutional networks for skeleton-based action recognition". in *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA*.

Li, C., Hou, Y., Wang, P., and Li, W. (2017). "Skeleton-based action recognition with convolutional neural networks." *IEEE Signal Process. Lett.* pp. 624–628.

Li, Y., Zhang, Z., Shi, L., Hou, X., and Wang, J. (2019). "Two-stream adaptive graph convolutional networks for skeleton-based action recognition" in Proc. IEEE/CVF conference on computer vision and pattern recognition (CVPR), 12026–12035. doi: 10.1109/CVPR.2019.01230

Li, W., Zhang, C., and Zou, Y. (2023). Spatiotemporal focus for skeleton-based action recognition. *Pattern Recogn.* 136:109231.

Lin, L., Zhang, J., and Liu, J., (2023). "Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition," arXiv:2303.10904.

Liu, B., Ju, Z., and Liu, H. (2018). A structured multi-feature representation for recognizing human action and interaction. *Neurocomputing* 318, 287–296. doi: 10.1016/j.neucom.2018.08.066

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., and Kot, A. C. (2020b). NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2684–2701. doi: 10.1109/tpami.2019.2916873

Liu, J., Wang, Z., and Liu, H. (2020a). HDS-SP: a novel descriptor for skeleton-based human action recognition. *Neurocomputing* 385, 22–32. doi: 10.1016/j.neucom.2019.11.048

Pang, Y., Ke, Q., Rahmani, H., Bailey, J., and Liu, J. (2022). "Igformer: interaction graph transformer for skeleton-based human interaction recognition." in *European Conference on Computer Vision*. Cham: Springer Nature Switzerland.

Pei, Y., Luo, Z., Yan, Y., Yan, H., Jiang, J., Li, W., et al. (2021b). Data augmentation: using channel-level recombination to improve classification performance for motor imagery EEG. *Front. Hum. Neurosci.* 15:645952. doi: 10.3389/fnhum.2021.645952

Pei, Y., Luo, Z., Zhao, H., Xu, D., Li, W., Yan, Y., et al. (2021a). A tensor-based frequency features combination method for brain-computer interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* 30, 465–475.

Pei, Y., Zhao, S., Xie, L., Ji, B., Luo, Z., Ma, C., et al. (2025a). Toward the enhancement of affective brain-computer interfaces using dependence within EEG series. *J. Neural Eng.* 22:026038. doi: 10.1088/1741-2552/adbfc0

Pei, Y., Zhao, S., Xie, L., Luo, Z., Zhou, D., Ma, C., et al. (2025b). Identifying stable EEG patterns in manipulation task for negative emotion recognition. *IEEE Trans. Affect. Comput.* doi: 10.1109/TAFFC.2025.3551330

Perez, M. D., Liu, J., and Kot, A. C. (2021). Interaction relational network for mutual action recognition. *IEEE Trans. Multimed.* 24, 366–376.

Plizzari, C., Cannici, M., and Matteucci, M. (2021). Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.* 208-209:103219. doi: 10.1016/j.cviu.2021.103219

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., et al (2017). A simple neural network module for relational reasoning, advances in neural information processing systems 30.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). "NTU RGB+D: a large scale dataset for 3D human activity analysis." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2018). Non-local graph convolutional networks for skeleton-based action recognition.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2021). "Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition," in *Computer vision – ACCV 2020, Lecture Notes in Computer Science*. pp. 38–53.

Shu, X., Tang, J., Qi, G.-J., Liu, W., and Yang, J. (2019). Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 1110–1118.

Shu, X., Tang, J., Qi, G.-J., Song, Y., Li, Z., and Zhang, L., (2017). "Concurrence-aware long short-term sub-memories for person-person action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Simonyan, K., and Zisserman, A. (2014). ""Two-stream convolutional networks for action recognition in videos." in *Neural Information Processing Systems*.

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition." in *International Conference on Learning Representations*.

Song, S., Lan, C., Xing, J., Zeng, W., and Liu, J. (2022). "An end-to-end spatio-temporal attention model for human action recognition from skeleton data." in *Proc. AAAI Conf. Artif. Intell.*

Song, Y.-F., Zhang, Z., Shan, C., and Wang, L. (2021). Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Trans. Circuits Syst. Video Technol.* 31, 1915–1925. doi: 10.1109/TCSVT.2020.3015051

Tempel, F., Ihlen, E. A. F., and Strümke, I. (2024). Auto GCN-toward generic human activity recognition with neural architecture search. *IEEE Access* 12, 39505–39516. doi: 10.1109/ACCESS.2024.3377103

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3D convolutional networks." in *IEEE International Conference on Computer Vision (ICCV)*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Trivedi, N., and Sarvadevabhatla, R. (2022). "Psumnet: unified modality part streams are all you need for efficient pose-based action recognition." in *European Conference on Computer Vision*. Cham: Springer Nature Switzerland.

Wang, X., Pei, Y., Luo, Z., Zhao, S., Xie, L., Yan, Y., et al. (2023). Fusion of multi-domain EEG signatures improves emotion recognition. *J. Integr. Neurosci.* 23:018. doi: 10.31083/j.jin2301018

Wu, B., Wan, A., Yue, X., Jin, P., Zhao, S., Golmant, N., et al (2018). "Shift: a zero FLOP, zero parameter alternative to spatial convolutions." *in 2018 IEEE/CVF Conference on Computer Vision and pattern recognition, 2018.*

Xu, K., Ye, F., Zhong, Q., and Xie, D. (2022). Topology-aware convolutional neural network for efficient skeleton-based action recognition. *Proc. AAAI Conf. Artif. Intell.* 36, 2866–2874. doi: 10.1609/aaai.v36i3.20191

Xu, M., Dai, W., Liu, C., Gao, X., Lin, W., Qi, G.-J., et al (2020). Spatial-temporal transformer networks for traffic flow forecasting. arxiv preprint arxiv:2001.02908.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proc. AAAI Conf. Artif. Intell.* 32:12328. doi: 10.1609/aaai.v32i1.12328

Yang, H., Gu, Y., Zhu, J., Hu, K., and Zhang, X. (2020). PGCN-TCA: pseudo graph convolutional network with temporal and channel-wise attention for skeleton-based action recognition. *IEEE Access* 8, 10040–10047. doi: 10.1109/ACCESS.2020.2964115

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D. (2012). "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.*

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., and Zheng, N. (2017). "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *2017 IEEE International Conference on Computer Vision (ICCV)*.

Zhang, N., Wang, Y., and Yu, P. (2018). "A review of human action recognition in video," in *2018 IEEE/ACIS 17th international conference on computer and information science (ICIS)*.

Zhang, J., Zheng, Y., and Qi, D. (2022). "Deep spatio-temporal residual networks for citywide crowd flows prediction." in *Proc. AAAI Conf. Artif. Intell.*

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). "Temporal relational reasoning in videos," in *Computer vision– ECCV 2018, Lecture Notes in Computer Science*. pp. 831–846.