ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1490-1509
DOI 10.5755/j01.itc.55.3.44650

DTRSF-Net: Dual-Temporal RGB-Skeleton Fusion for Fitness Action Recognition on National Fitness Platforms

Received 2026/03/27 | Accepted after revision 2026/06/26

HOW TO CITE: Tian, J., Zhou, S. (2026) DTRSF-Net: Dual-Temporal RGB-Skeleton Fusion for Fitness Action Recognition on National Fitness Platforms. *Information Technology and Control*, 55(3), 1489-1508. https://doi.org/10.5755/j01.itc.55.3.44650

DTRSF-Net: Dual-Temporal RGB-Skeleton Fusion for Fitness Action Recognition on National Fitness Platforms

Jie Tian

Department of Information and Media, Jiangsu City Vocational College, Suzhou, 215100, China; Department of Information and Media, Jiangsu United Vocational and Technical College, Suzhou, 215100, China;
e-mail: jerrylichunli@163.com

Songshan Zhou*

Chinese Martial Arts Department, Hebei Sport University, Shijiazhuang, 050041, China;
e-mail: Zhousongss@outlook.com

Corresponding author: Zhousongss@outlook.com

The national fitness platform is moving toward an intelligent coaching stage. Action recognition therefore needs to remain stable and reliable for users of different ages, physical abilities, and exercise types. Existing action recognition methods in fitness scenarios tend to be affected by local occlusion, lighting changes, and complex backgrounds. In addition, methods relying only on a single video modality encounter difficulties in balancing high accuracy with real-time inference efficiency in platform applications. To address these problems, this study proposes a Dual-Temporal RGB-Skeleton Fusion Network (DTRSF-Net) for fitness action recognition, which combines a video dual temporal sampling branch with a skeleton branch. The front end first slices the video and forms a sparse sampling sequence and a dense sampling sequence. It also performs keypoint estimation to obtain a normalized skeleton sequence. The back end uses dual branch three-dimensional convolution encoding to extract spatiotemporal features. It also introduces a skeleton action encoding network to obtain posture structure information. A fusion head then produces multimodal decision outputs. Experimental results show that after 200 training epochs, the Top-1 accuracy of DTRSF-Net reaches 98.04%.

The comparison and baseline models achieved 94.29% and 90.82%, respectively. The Top-5 accuracy reaches 99.70%. The comparison model and the baseline model achieve 98.40% and 96.79%. The results indicate that multimodal collaborative spatiotemporal feature learning achieves higher recognition accuracy in fitness action recognition tasks. The method provides intelligent action recognition capability for the national fitness platform across multiple scenarios. It also offers stable data support for functions such as online correction, action counting, and training records.

KEYWORDS: National fitness platform, Action recognition, SlowFast, Skeleton-based Learning, Multimodal fusion

1. Overview

National fitness has evolved from policy-promoted offline gym training to digital service models that focus on mobile Internet platforms. These models reduce the threshold for fitness training through course distribution, training planning, and interactive guidance. The concept of "anytime, anywhere training" has become a new norm [4]. With the popularity of home fitness training, lightweight equipment training, and artificial intelligence coaching, action recognition has become a new trend for mobile internet platforms. The results of action recognition are not only used for training records and action counting but also for important functional operations like real-time error correction, training rhythm control, and user safety reminders [10]. Current action recognition research is progressing beyond the limitations of single-stream feature modeling. It pays more attention to the joint use of appearance information and structural information, the effective modeling of key time segments, and the feasibility of deployment in complex application scenes. For example, a dual-branch Skeleton-RGB fusion framework incorporating a spatiotemporal attention mechanism was proposed by Lan et al. [8]. This shows that the explicit cooperation of appearance information and human structural information can improve action discrimination in complex scenes [8]. Wu et al. [18] further proposed a graph convolutional network with local and global self-attention and dynamic frame weighting. This indicates that representative frames and key motion segments ought to be emphasized in action recognition, rather than the uniform aggregation of the entire sequence [18]. Le et al. [9] combined a Long Short-Term Memory network with a depthwise separable convolutional neural network for skeleton-based action recognition. This shows that efficient convolution structures still have important value when recognition performance, computational cost, and deployment efficiency need to be balanced [9]. These studies show that recent action recognition methods pay more attention to structured fusion, discriminative temporal weighting, and lightweight model design. However, for national fitness platforms, the real challenge is not only to improve offline recognition accuracy, but also to keep recognition stable and efficient under repetitive fitness actions, viewpoint changes, partial occlusion, and heterogeneous terminal devices. Although existing studies have made some progress, there is still a clear gap in the practical application of national fitness platforms. When occlusion, illumination change, cluttered background, device heterogeneity, and repetitive action cycles appear at the same time, existing methods still tend to face problems such as insufficient feature alignment, unstable temporal decisions, or high deployment cost. Therefore, an action recognition framework for national fitness platforms needs not only joint modeling of appearance information and human structural information, but also consideration of temporal completeness, modality reliability, and inference efficiency in a unified architecture.

Nevertheless, there remains a critical deficiency in practical applications related to national fitness platforms: when "occlusion, lighting variations, complex backgrounds, device differences, and venue differences" exist together, there is a tendency towards insufficient feature alignment and unstable decision-making in single-modality or loosely coupled recognition systems, thereby reducing robustness in online real-time recognition. To address this deficiency, Three-Dimensional Convolutional Neu-

ral Network (3D CNN) modeling excels at capturing human appearance and local motion details, making it suitable for expressing fine-grained action differences in equipment training [11]. The dual temporal scale dual branch structure has a low computational cost while it can handle long-term and short-term dynamic information at the same time. Therefore, it is suitable for improving real-time inference efficiency in platform-oriented recognition scenarios [20]. The spatiotemporal graph modeling based on human skeletons can utilize the topology of joints to obtain structural patterns of action while suppressing background interference [17]. Therefore, this study proposes a Dual-Temporal RGB-Skeleton Fusion Network (DTRSF-Net) for fitness action recognition on national fitness platforms. The frontend of DTRSF-Net builds multi-temporal video inputs through video slicing and two-path sampling, and obtains normalized skeleton sequences through pose estimation. The backend extracts video spatiotemporal features through the dual-temporal 3D ResNet video encoder and obtains pose structural features through the ST-GCN-based skeleton action encoder. Multimodal feature alignment and fusion decision-making are achieved through a unified training framework. The contribution of DTRSF-Net is not the isolated use of SlowFast, 3D ResNet, ST-GCN, FPN, depthwise separable convolution, confidence weighting, or EMA, but their task-oriented coupling for platform fitness videos. Dual-temporal RGB clips and normalized skeleton sequences are generated from the same sliding window, aligned in a shared decision space, weighted according to modality reliability, and temporally stabilized for continuous recognition and action counting under occlusion, viewpoint changes, and cluttered backgrounds. To solve the low recognition accuracy of single-modality models in complex home environments, and to meet the deployment needs of national fitness platforms, this study chose SlowFast as the video encoder. In real platform applications, uploaded videos originate from mobile phones, tablets, and low-cost cameras. Therefore, DTRSF-Net is designed to reduce unnecessary computation while maintaining temporal modeling ability. Since the experiments were conducted on an RTX 4090 platform, the reported efficiency results are interpreted as computational-efficiency evidence rather than direct proof of edge-device deployment.

2. Methodology

2.1. SlowFast Sampling and Normalized Skeleton Sequence for Action Recognition

RGB video clips and normalized skeleton sequences from the same sliding window are utilized as inputs. The video branch incorporates slow and fast 3D CNN pathways to encode long-term and short-term motion patterns, and the skeleton branch uses Spatio-Temporal Graph Convolutional Network (ST-GCN) to model joint topology and temporal coordination. The three feature streams are fused by confidence-guided weighting, and an exponential moving average is applied to stabilize online predictions. Because user videos come from heterogeneous devices with different frame rates and resolutions, a unified slice generation mechanism is adopted, as shown in Equation (1).

$$X^{(k)}=\{I_t \mid t=ks_{\text{win}}, ks_{\text{win}}+1,\ldots,ks_{\text{win}}+T_{\text{clip}}-1\}. \quad (1)$$

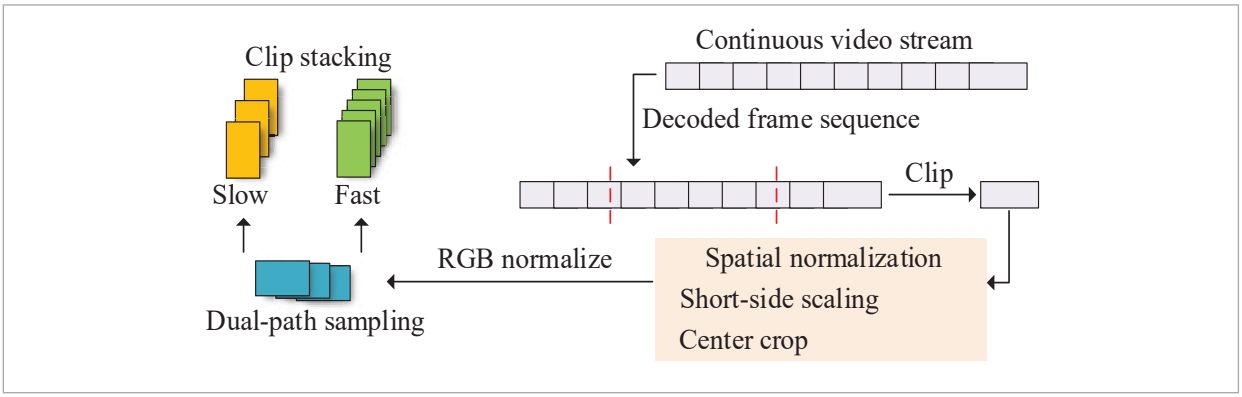
In Equation (1), I_t represents the t-th original frame decoded from the user side, and $X^{(k)}$ represents the k-th standard video segment, namely the clip. T_{clip} denotes the clip frame le ngth, and s_{win} denotes the sliding window stride used for continuous recognition window movement. After obtaining the basic video segments, geometric and pixel level normalization is required to eliminate the influence caused by differences in shooting distance and devices among users. The normalization process is shown in Figure 1.

After spatial normalization, the image data require further normalization to accelerate network convergence. The unified scaling and normalization are shown in Equation (2).

$$I_t'=T_{\text{prep}}(I_t)=\frac{\text{Resize}_{H_{\text{in}}\times W_{\text{in}}}(\text{Crop}(I_t))-\mu_{\text{rgb}}}{\sigma_{\text{rgb}}}. \quad (2)$$

In Equation (2), T_{prep} represents the preprocessed image after short side scaling or center cropping. H_{in} and W_{in} represent the network input size. μ_{rgb} and σ_{rgb} are RGB normalization parameters used to reduce the influence of environmental lighting changes on feature extraction. To capture both fast motion de-

Figure 1
Preprocessing flow of dual-path sampling splitting.



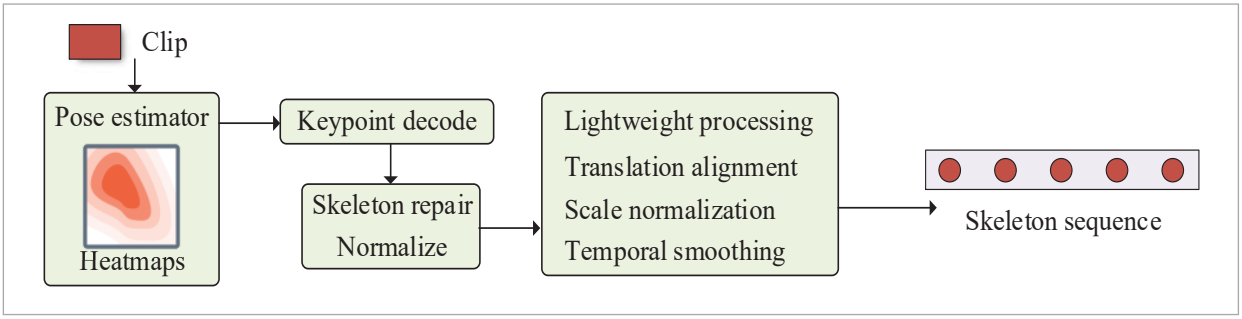
tails and slow posture changes in fitness actions, this study adopts the SlowFast dual pathway sampling strategy. The normalized clip is further partitioned into two streams. The diversion formula is shown in Equation (3).

$$\begin{cases} X_{\text{slow}} = (I'_{t_0+n\,\tau_{\text{slow}}})_{n=0}^{T_{\text{slow}}-1} \\ X_{\text{fast}} = (I'_{t_0+n\,\tau_{\text{fast}}})_{n=0}^{T_{\text{fast}}-1}, \quad \tau_{\text{slow}} = \alpha\tau_{\text{fast}} \end{cases}. \tag{3}$$

In Equation (3), $t_0 = ks_{\text{win}}$ represents the starting frame of the clip. τ_{slow} and τ_{fast} represent the sampling intervals of the two pathways. α represents the temporal rate of SlowFast. T_{slow} and T_{fast} represent the number of frames selected in the two pathways. The RGB branch inputs are X_{slow} and X_{fast}. The slow pathway uses sparse sampling to preserve long-range posture continuity, whereas the fast pathway uses dense sampling to retain short-term motion changes that are important for distinguishing actions with similar overall poses but different rhythms. In parallel, a keypoint detection branch converts human joints into heatmaps for skeleton extraction, as shown in Figure 2.

As shown in Figure 2, heatmap regression was employed to locate human keypoints. The skeleton sequence was extracted by a top-down pose estimator based on the HRNet-W32 backbone, and the pretrained weights were loaded from the COCO keypoint dataset [16]. The detector outputs 17 human joints per frame, and only the person with the highest confidence in the foreground fitness area was kept for later action modeling. Missing joints were estimated via temporal interpolation. Low-confidence joints were refined using a confidence-aware filtering strategy prior to coordinate normalization. Based on Gaussian heatmap encoding, the model learns the spatial distribution probability of human joints more accurately. The spatial distribution probability is shown in Equation (4).

Figure 2
Pose estimation front end and skeleton preprocessing.



$$H_{j,t}(x,y)=\exp\left(-\frac{(x-u_{j,t})^2+(y-v_{j,t})^2}{2\sigma_{\text{hm}}^2}\right). \quad (4)$$

In Equation (4), $(u_{j,t},v_{j,t})$ is the pixel coordinate of the j-th keypoint in the t-th frame, such as the left knee or right shoulder. $H_{j,t}(x,y)$ represents the response of the corresponding heatmap at (x,y), and σ_{hm} controls the diffusion range of the heatmap. This type of Gaussian encoding explains the supervision or decoding process from human regions to keypoints in the construction of skeleton branch input. After obtaining the heatmap response, this study locates the response peak and uses the least squares method for coordinate decoding to obtain accurate joint positions [19]. The least squares decoding from heatmap to coordinates is shown in Equation (5).

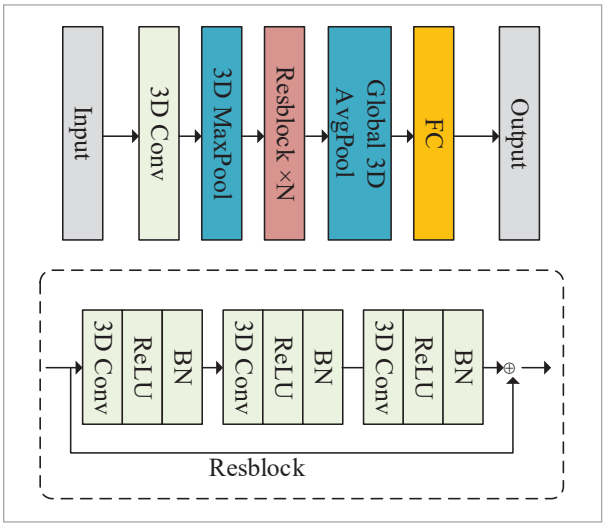
$$AY_{j,t}=B,\quad Y_{j,t}=(A^{\top}A)^{-1}A^{\top}B. \quad (5)$$

In Equation (5), $Y_{j,t}$ represents the keypoint coordinate to be solved. A and B represent the linear equation system constructed by (x,y,H) beams sampled within the neighborhood of the heatmap peak. Considering that users of National Fitness vary from children to adults to elderly people, their heights and shapes are greatly different. Using original coordinates for action recognition results in poor generalization. In this regard, a lightweight preprocessing approach is proposed for the skeleton sequence in spatiotemporal dimensions. The skeleton lightweight preprocessing is shown in Equation (6).

$$\hat{p}_{j,t}=\frac{p_{j,t}-p_{\text{root},t}}{d_{\text{scale},t}},\quad p_{j,t}^{\text{sm}}=\beta\hat{p}_{j,t}+(1-\beta)p_{j,t-1}^{\text{sm}}. \quad (6)$$

In Equation (6), $p_{j,t}$ represents the original keypoint coordinate. $p_{\text{root},t}$ represents the reference joint, such as the pelvis or torso center, for translation alignment. $d_{\text{scale},t}$ represents the scale factor, such as shoulder width or torso length, for scale normalization. $p_{j,t}^{\text{sm}}$ represents the temporally smoothed result, and β represents the smoothing coefficient. Figure 3 then illustrates the 3D ResNet-based spatiotemporal encoding process used to capture both spatial texture and temporal evolution of fitness actions.

Figure 3
3D ResNet spatiotemporal encoding block.



In Figure 3, the module of 3D ResNet adds a temporal dimension and extends the traditional 2D spatial convolution operation into a 3D spatiotemporal convolution operation. The input features are passed through a BN layer, a ReLU activation function, and a 3D CNN layer. Residual connections are employed to mitigate the gradient vanishing problem in deep networks. The action features are extracted while keeping the original spatiotemporal information flow. Based on this architecture, the 3D CNN operation encodes both spatial texture and temporal evolution of fitness actions simultaneously, as shown in Equation (7).

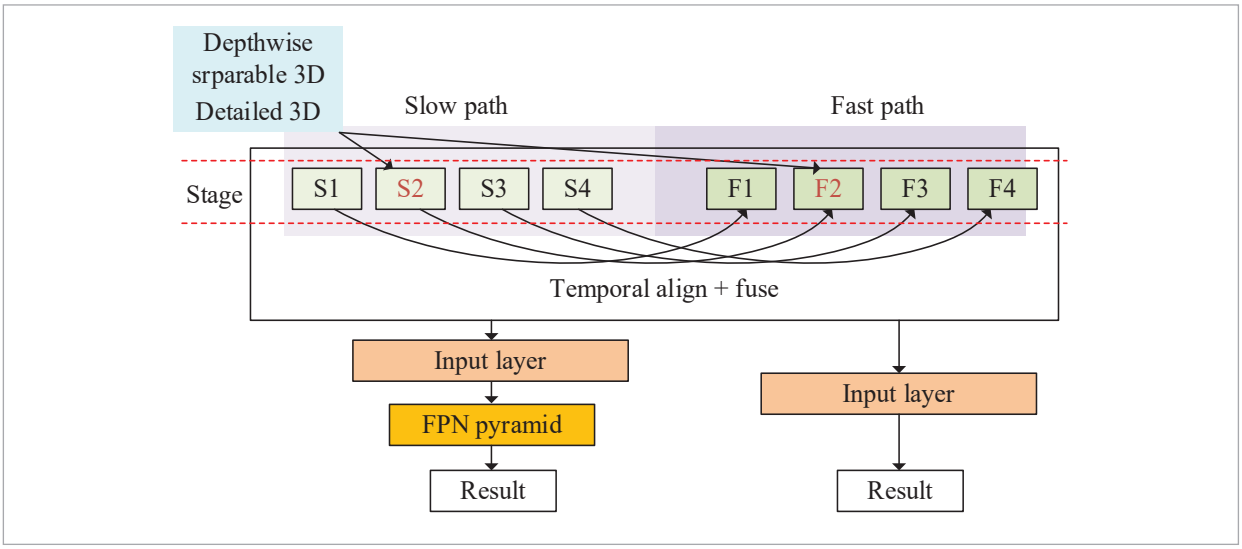
$$y_{i,j,k,l}=\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1}\sum_{r=0}^{R-1}w_{p,q,r,l}x_{i+p,j+p,k+r}. \quad (7)$$

In Equation (7), $x_{i,j,k}$ represents the sampling position of input features in temporal and spatial dimensions. $w_{p,q,r,l}$ represents the parameters of the convolution kernel. (P,Q,R) represents the kernel size in three dimensions. $y_{i,j,k,l}$ represents the response of the l-th output channel at (i,j,k). This response directly reflects the activation level of a specific motion pattern, such as arm raising or squatting.

2.2. DTRSF-Net Architecture and Multimodal Collaborative Action Recognition Model

However, the original 3D ResNet still has limited efficiency and insufficient sensitivity to subtle actions

Figure 4
Dual-temporal 3D ResNet video encoder in DTRSF-Net.



under occlusion and complex home backgrounds. Therefore, this study introduces lightweight spatiotemporal factorization, depthwise separable convolution, and multi-scale feature enhancement to construct the dual-temporal 3D ResNet video encoder in DTRSF-Net, as shown in Figure 4.

In Figure 4, the slow pathway emphasizes spatial semantics, and the fast pathway emphasizes motion information. The video encoder in DTRSF-Net is improved by adding depthwise separable convolution to reduce computational cost and by integrating a feature pyramid network module into the lateral connections to improve feature fusion. The decomposition of 3D CNN is shown in Equation (8).

$$\begin{aligned} &\mathrm{Conv}_{7\times7\times7} \approx \mathrm{Conv}_{7\times1\times1} \circ \mathrm{Conv}_{1\times7\times7}, \\ &\mathrm{Conv}_{3\times3\times3} \approx \mathrm{Conv}_{3\times1\times1} \circ \mathrm{Conv}_{1\times3\times3} \end{aligned}. \tag{8}$$

In Equation (8), ∘ represents serial execution. Furthermore, to achieve a more lightweight design, this study adopts the depthwise separable 3D CNN strategy. This method decouples channel correlations from spatial correlations. The depthwise separable 3D CNN is shown in Equation (9).

$$F = \mathrm{PW}(\mathrm{DW}(X; W_{\mathrm{dw}}); W_{\mathrm{pw}}). \tag{9}$$

In Equation (9), DW represents channel wise 3D CNN that extracts independent spatiotemporal features for each channel. PW represents the 1×1×1 pointwise convolution used for channel mixing. W_{dw} and W_{pw} are the parameters of the two parts. Fitness actions show large motion amplitude variation, such as marching in place and burpees. Conventional convolution cannot balance receptive fields at different scales [21]. This study introduces dilated 3D CNN to expand the effective receptive field and capture larger range limb coordination information. The effective receptive field is shown in Equation (10).

$$K_{\mathrm{eff}} = k + (k+1)(a-1). \tag{10}$$

In Equation (10), k represents the original convolution kernel size. a represents the dilation rate. K_{eff} represents the effective convolution kernel size. During dual pathway feature extraction, the fast pathway and the slow pathway need effective interaction to compensate for the limited temporal resolution of the slow pathway [22]. This study performs spatiotemporal feature alignment and fusion through lateral connections. The expression is shown in Equation (11).

$$\begin{aligned} &F_{\mathrm{fast}\to\mathrm{slow}} = \mathrm{Conv}^{(\alpha)}_{5\times1\times1}(F_{\mathrm{fast}}), \\ &f_{\mathrm{rgb}} = \phi([F_{\mathrm{slow}}, F_{\mathrm{fast}\to\mathrm{slow}}]) \end{aligned}. \tag{11}$$

In Equation (11), F_{slow} and F_{fast} represent the feature tensors of the fast pathway and the slow pathway. $\text{Conv}_{5\times1\times1}^{(\alpha)}$ represents a convolution with kernel size 5×1×1 and temporal stride α. This operation downsamples the fast pathway along the temporal dimension to align it with the slow pathway. $\phi(\cdot)$ uses a 1×1×1 convolution to achieve channel fusion. To enhance the adaptability of the model to different shooting distances, this study integrates FPN into the backbone network to achieve multi-level feature aggregation. The recursive formulation is shown in Equation (12) [15].

$$P_l = \psi(C_l) + U(P_{l+1}). \tag{12}$$

In Equation (12), C_l represents the output feature of the backbone at the l-th stage. $\psi(\cdot)$ uses a 1×1×1 convolution for channel alignment. $U(\cdot)$ represents the feature after upsampling to the same scale as C_l. The addition produces the pyramid feature P_l. Beyond visual features, the skeleton branch processes human topology information through ST-GCN, as shown in Figure 5.

In Figure 5, the skeleton action encoder represents human joints as nodes in the graph, and connections between limbs and trajectories over time are represented as edges. Graph convolution operations are applied to extract coordinated motion features between human joints while preserving the topology of the human skeleton. This is due to the nature of the human skeleton, which is a natural graph. This enables GCN to efficiently extract coordinated motion relationships between human joints. The skeleton graph convolution is shown in Equation (13).

$$f_{\text{ske}} = \sum_{k=1}^{K} \Lambda_k^{-\frac{1}{2}} A_k \Lambda_k^{-\frac{1}{2}} \hat{S} W_k. \tag{13}$$

In Equation (13), $\hat{S}$ represents the normalized skeleton sequence, which stacks J keypoint coordinates along the temporal dimension. A_k represents the adjacency matrix of the k-th subgraph. Λ_k represents the corresponding degree matrix used for normalization. W_k represents learnable weights. f_{ske} represents the input normalized skeleton sequence. To address the failure of a single modality in specific scenarios, this study designs a modality fusion and training strategy during the training stage. The training framework is shown in Figure 6.

As shown in Figure 6, in the inference stage, a late fusion strategy is employed by the proposed study. The features are dynamically weighted according to the confidence of each modality to produce a decision result. When observation quality changes quickly, adaptive weighting is needed to maintain decision stability across adjacent clips. Specifically, the features of each modality are first projected into a shared embedding space. Then they are compared with the running class prototype or reference feature. A higher alignment score means that the current modality is more consistent with its historical discriminative pattern. Thus, it is more suitable as a decision basis in the current scene. This design is based on the fact that modality quality in fitness videos strongly depends on the scene. RGB features can degrade under occlusion or cluttered backgrounds, while skeleton features may become unstable due to missing keypoints or motion blur. Therefore, dy-

Figure 5

ST-GCN-based skeleton action encoder in DTRSF-Net.

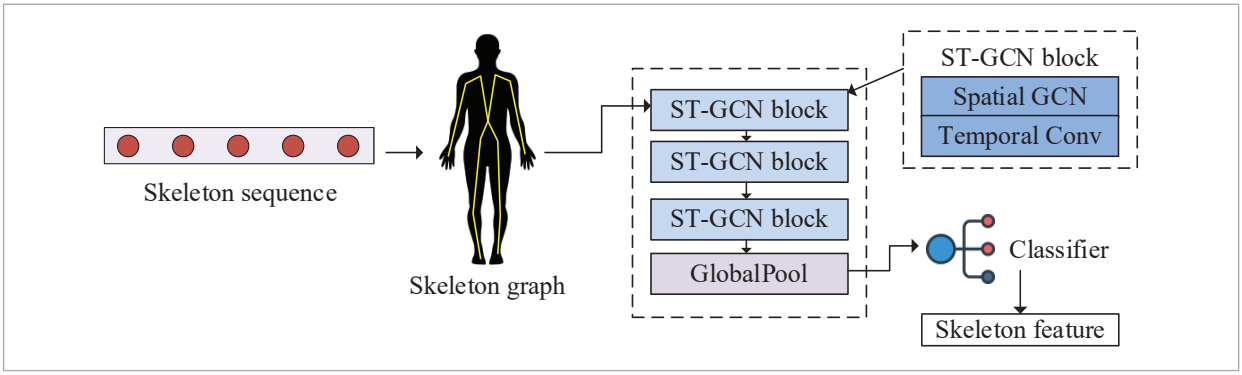
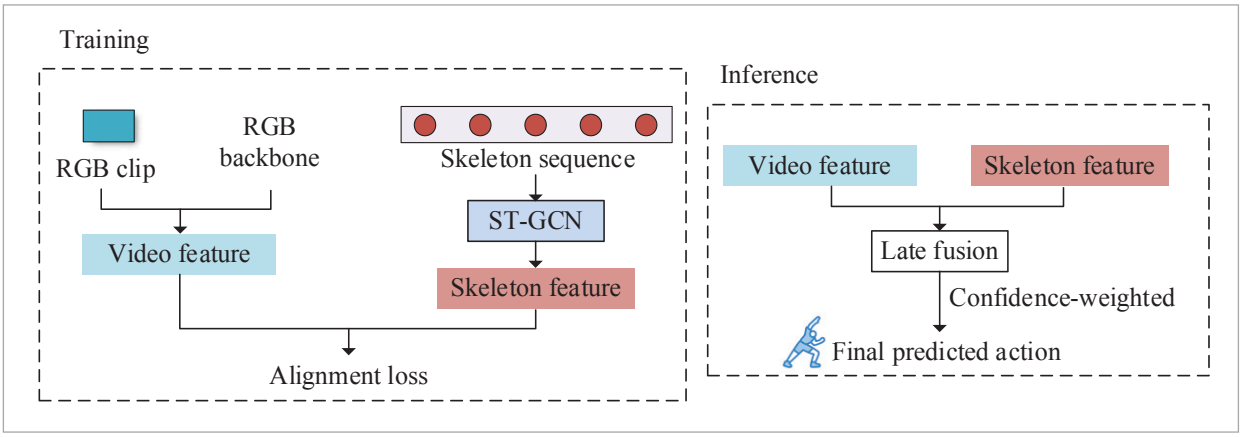


Figure 6
Skeleton mixed augmentation and joint training framework.



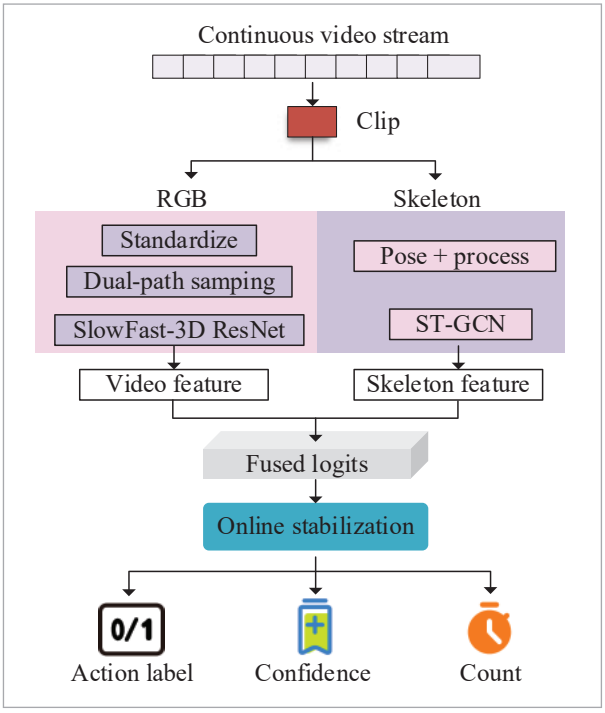
namic weighting can actively reduce the influence of unreliable modalities and keep stable modalities. Hence, it is more suitable for online fitness action recognition than fixed fusion coefficients. The weight calculation of late fusion is shown in Equation (14).

$$\alpha_m = \frac{\exp(\mathrm{sim}(F_m, F_{\mathrm{ref}}))}{\sum_{j=1}^{N} \exp(\mathrm{sim}(F_j, F_{\mathrm{ref}}))}, \quad z_{\mathrm{final}} = \sum_{m=1}^{N} \alpha_m F_m . \tag{14}$$

In Equation (14), F_m represents the m-th feature vector. This study sets $F_1 = z_{\mathrm{rgb}}$ and $F_2 = z_{\mathrm{ske}}$. F_{ref} represents the reference feature. α_m represents the late fusion weight. This study used the inner product as a compact confidence proxy because it can reflect both feature magnitude consistency and alignment with the class reference direction. After score normalization, the obtained weight represents the relative reliability of the i-th modality in the current clip, rather than an absolute probability. This form was used because it adds almost no extra computational cost, and it is also suitable for real-time multimodal fusion on national fitness platforms with limited computing resources. And the fused outputs are further stabilized by an exponential moving average in the final prediction stage, as shown in Figure 7.

As illustrated in Figure 7, DTRSF-Net includes dual-pathway video sampling, skeleton extraction, dual-stream feature encoding, and late fusion. It is noteworthy that the proposed architecture emphasizes the online stabilization module in the last stage. The online stabilization module refines the output of the continuous time window and provides the action category, confidence, and the timestamp for the count. With the Exponential Moving Average (EMA) algorithm, the model outputs smooth and stable prediction results and satisfies the accuracy requirement of platform counting. The EMA stabilization for continuous recognition and the joint training objective are shown in Equation (15).

Figure 7
DTRSF-Net model architecture.



$$\begin{cases} z_t^{\text{ema}} = \lambda_{\text{ema}} z_{t-1}^{\text{ema}} + (1-\lambda) z_{\text{final},t} \\ \tilde{S} = P(S^{(a)}, S^{(b)}) = \alpha_{\text{mix}} S^{(a)} + (1-\alpha_{\text{mix}}) S^{(b)}. \\ L = L_{\text{ce}}(z_{\text{final}}, y) + \lambda_{align} \left\| z_{\text{rgb}} - z_{\text{ske}} \right\|_2^2 \end{cases} \quad (15)$$

In Equation (15), $z_{\text{final},t}$ represents the fused logits or embeddings of the t-th sliding window. z_t^{ema} represents the stabilized output after EMA. λ_{ema} controls the strength of fluctuation suppression. During training, skeleton mixed augmentation is represented by $P(\cdot)$. Two sample skeletons $S^{(a)}$ and $S^{(b)}$ are linearly mixed with weight α_{mix} to obtain $\tilde{S}$. The classification cross entropy L_{ce} provides the main supervision. $\|z_{\text{rgb}} - z_{\text{ske}}\|_2^2$ provides modality alignment regularization.

3. Results and Analysis

3.1. Experimental Environment Setup

DTRSF-Net was evaluated using Kinetics-400, Fitness-101, and M3GYM with different experimental roles rather than through a single merged split. Kinetics-400 was used to initialize the RGB video encoder by generic action pre-training. Fitness-101 was used for task-specific training, validation, and testing, and its videos were divided at the video level into training, validation, and test subsets at a ratio of 7:1:2. M3GYM was reserved for external fitness-scene evaluation of online inference, environmental interference, and venue transfer, and it was not used for model selection. Sparse and dense sampling were used for the slow and fast pathways, respectively, and normalized keypoint sequences were used as inputs to the skeleton branch. The model was implemented in Python 3.10 with PyTorch 2.1 and CUDA 12.1 on an NVIDIA RTX 4090 GPU and an Intel Core i9 platform with 64 GB RAM. AdamW was used with an initial learning rate of 1×10^{-4}, cosine annealing with 5-epoch warm-up, weight decay of 1×10^{-4}, batch size 16, and 200 epochs. The comparison models were 3D ResNet-50 and SlowFast-YOLOv8. Ablation experiments were conducted on single-branch settings, equal-weight fusion, alignment removal, and key lightweight modules under the same data split, input resolution, and training budget.

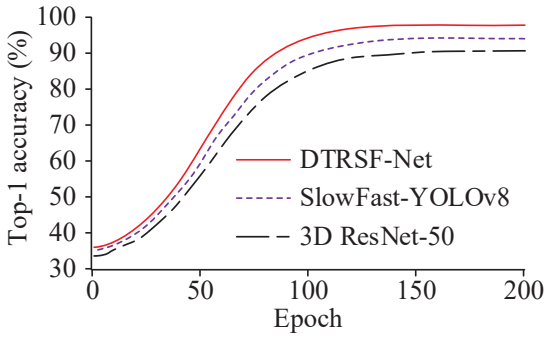
3.2. Recognition Performance Analysis

To compare the convergence speed and final performance differences of each model during training, this study validated the changes in Top-1 accuracy and Top-5 accuracy of each model during 200 training epochs, as shown in Figure 8.

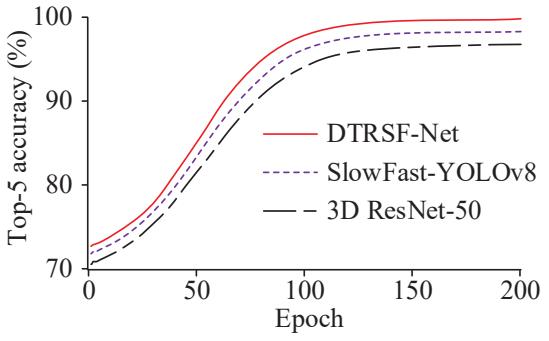
In Figure 8(a), the Top-1 accuracy of DTRSF-Net reached 98.04% at the 200th epoch of the training process. It outperformed SlowFast-YOLOv8 and 3D ResNet-50, which achieved Top-1 accuracies of 94.29% and 90.82%, respectively. This is because DTRSF-Net enhanced the joint modeling ability for action rhythm and pose structure by the collaborative representation and fusion decision-making of the multi-temporal scale video branches and the

Figure 8

Comparison of Top-1 accuracy and Top-5 accuracy convergence during training.

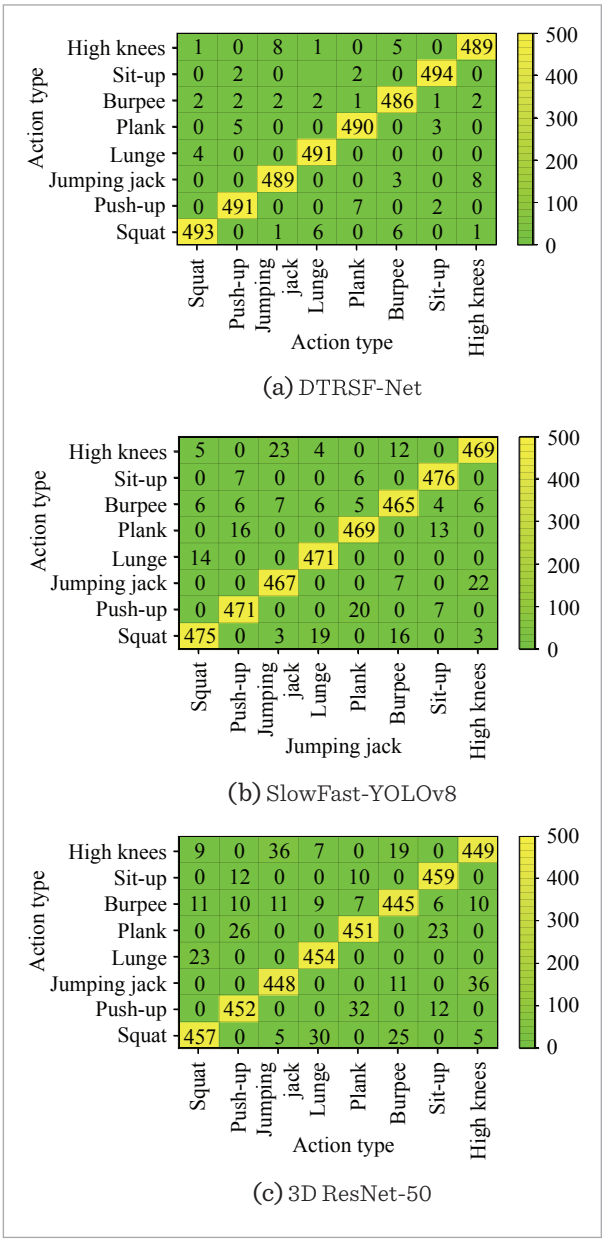


(a) Comparison of Top-1 accuracy across models



(b) Comparison of Top-5 accuracy across models

Figure 9
Classification confusion matrices for different fitness actions.



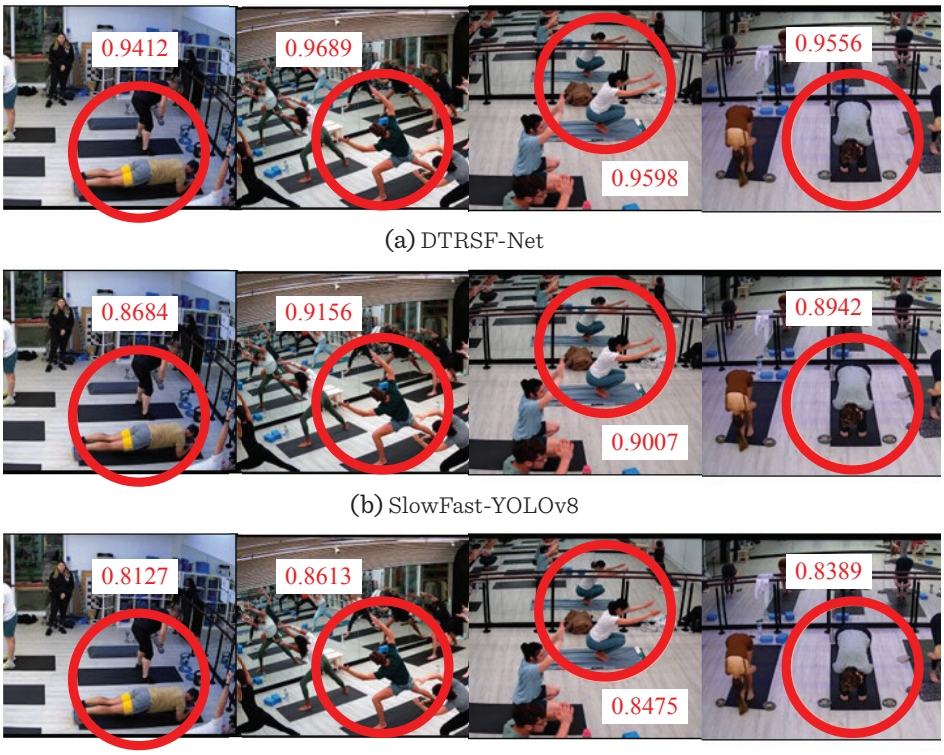
skeleton branches. This reduced the noise of the features caused by occlusion, illumination changes, and background clutter. In Figure 8(b), the Top-5 accuracy of DTRSF-Net reached 99.70% at the 200th epoch. The advantage of DTRSF-Net lay in the fact that multimodal spatiotemporal feature alignment and fusion enhanced inter-class separability and fine-grained differentiation of similar actions, allowing Top-5 to reach a high level earlier and maintain the lead. Subsequently, this study examined the discriminative boundaries and confusion situations of the model on specific action categories. With 500 groups of each of 8 action types, the recognition performance of DTRSF-Net and comparison models on different fitness action recognition tasks is shown in Figure 9.

In Figure 9(a), the correct recognition numbers for the 8 action classes of DTRSF-Net were essentially between 486 and 494 (Sit-up at 494, Burpee at 486, High knees at 489), with the most errors appearing in "High knees → Jumping jack" at 8 cases. The results indicated that DTRSF-Net had the clearest discriminative boundaries for various action types and maintained low misclassification even between similar actions. These results indicate that 3D spatiotemporal modeling, integrated with temporal motion cues and skeleton structural cues, enables the model to capture both action rhythm and key pose structure simultaneously. In Figure 9(b), the actual recognition of the High knees action of the SlowFast-YOLOv8 was 469, while the number of incorrect recognitions of the action as Jumping jack was 23. The SlowFast-YOLOv8 was more likely to be confused between the pair of actions based on their rhythms and the upper limb support actions. The separability of the actions was less compared to the DTRSF-Net. As per Figure 9(c), the 3D ResNet-50 had the most incorrect recognitions. Out of the total of 449 correct recognitions of the High knees action of the 3D ResNet-50, as many as 36 were incorrectly recognized as Jumping jack action. The above results indicated that the baseline model had insufficient capability to characterize fine-grained action differences, with the lowest inter-class separation. This was likely because single 3D convolutional features under complex backgrounds and similar rhythm conditions more easily expanded intra-class variance and compressed inter-class intervals, and the discriminative boundaries for similar actions were more unstable without complementary constraints.

3.3. Online Inference and Computational Efficiency Evaluation

Subsequently, regarding the stability issue of action recognition in video streams, Figure 10 compared the recognition accuracy fluctuations of the three models when processing consecutive frames of different action videos to evaluate their online inference capability.

Figure 10
Comparison of recognition accuracy for different actions.



Note: Source from M3GYM. Faces have been masked, meeting ethical standards.

In Figure 10(a), the most stable and highest recognition accuracy of DTRSF-Net in consecutive frame conditions with different actions and different interference backgrounds was retained. At the same time, the recognition accuracy of the four action frames reached 94.12%, 96.89%, 95.98%, and 95.56%, respectively. Among them, the lowest single frame recognition accuracy remained above 94%, reflecting the more comprehensive aggregation of spatiotemporal dynamics and key pose cues of DTRSF-Net and the ability of the network to extract relatively robust discriminative features under occlusion, viewpoint differences, and multi-person backgrounds. In Figure 10(b), the recognition accuracies for the four action frames of SlowFast-YOLOv8 were 86.84%, 91.56%, 90.70%, and 89.42%, overall showing fluctuations where accuracy was higher for more typical actions and lower with more complex occlusion and backgrounds. This indicated that SlowFast-YOLOv8 had certain consecutive frame recognition capability, but its stability and upper limit accuracy were significantly weaker than DTRSF-Net under complex scenarios. In Figure 10(c), the recognition accuracies for the four action frames of 3D ResNet-50 were all below 87%, consistent with the confusion matrix examination results, with the highest misclassification risk. This meant that the 3D ResNet-50 had poor coverage of the discriminative features and was more dependent on the coarse features of the appearance. This was due to the multi-person motion interference, incomplete pose detection, and transition frames. This caused the performance of the network to deteriorate. To evaluate the computational efficiency and online counting stability of DTRSF-Net in platform-oriented applications, this study recorded the counting error (Mean Absolute Error, MAE) and operational efficiency (Frames Per Second, FPS) performance of different model variants under different time window settings. The specific results are shown in Figure 11.

Table 1
Ablation and comparison results of counting error metrics for different models.

Model	MAE (ms/clip)	RMSE (ms/clip)	MAPE (%)	Max error (ms/clip)	Std of error
HCTT	0.61	0.84	3.27	2.56	0.59
DBSA	0.57	0.80	3.05	2.43	0.55
S (slow only)	0.92	1.24	4.96	3.70	0.87
F (fast only)	0.88	1.19	4.71	3.58	0.83
K (skeleton only)	0.81	1.10	4.25	3.31	0.78
SF (slow+fast only)	0.69	0.95	3.63	2.94	0.66
SF+K (equal fusion)	0.63	0.87	3.31	2.71	0.61
w/o alignment	0.59	0.82	3.14	2.58	0.57
w/o EMA	0.58	0.86	3.11	3.02	0.64
DTRSF-Net	0.50	0.73	2.68	2.21	0.49

Note: S, F, and K denote the slow RGB pathway, fast RGB pathway, and skeleton branch in DTRSF-Net, respectively; SF and SF+K denote their ablated combinations.

Table 2
Ablation and comparison results of computational efficiency metrics for different models.

Model	Params (M)	FLOPs (G)	Inference latency (ms/clip)	FPS	GPU memory (GB)
HCTT	58.7	88.5	28.7	34.8	7.9
DBSA	50.9	79.4	24.9	40.2	7.1
S (slow only)	24.6	36.2	14.8	67.6	3.9
F (fast only)	26.1	39.8	15.6	64.1	4.1
K (skeleton only)	8.7	7.9	9.7	103.1	2.2
SF (slow+fast only)	34.8	53.7	18.9	52.9	5.1
SF+K (equal fusion)	41.2	62.1	20.8	48.1	5.9
w/o alignment	42.6	66.5	20.4	49.0	6.0
w/o EMA	43.2	68.7	20.2	49.4	6.1
DTRSF-Net	43.5	69.2	20.6	48.5	6.2

As shown in Table 1, DTRSF-Net achieved the lowest MAE, RMSE, MAPE, and maximum error among all compared models. This indicates that it had the best counting stability. Although HCTT and DBSA also showed competitive performance, their counting errors were still slightly higher than those of the proposed method. The ablation results further show that removing branch complementarity, alignment learning, or EMA stabilization leads to a clear drop in performance. This is especially reflected in the maximum error and error variance. This means that online counting benefits not only from stronger recognition accuracy, but also from stable outputs across adjacent clips. Then, this study further evaluated the parameter size, Floating Point Operations (FLOPs), inference latency, Frames Per Second (FPS), and GPU memory consumption of each model. The efficiency comparison is shown in Table 2.

In Table 2, DTRSF-Net did not achieve the absolute lowest computational cost, as the comprehensive multimodal design inevitably introduces additional parameters and memory consumption. However, DTRSF-Net kept a good balance between efficiency and accuracy. Its performance was better than that of the heavier baseline models, and it was also more practical than Transformer-based competing models in latency and FPS. Although the single-branch variants were faster, their counting errors related to accuracy were clearly worse. This confirms that DTRSF-Net provides a more suitable trade-off for real national fitness platform applications.

3.4. Environmental Interference Resistance and Generalization Performance Evaluation

To verify the model's anti-interference capability under complex environments, this study set up three different interference conditions: partial occlusion, lighting fluctuations, and cluttered backgrounds. The trend of Top-1 accuracy changes for each model under these conditions is shown in Figure 11.

In Figure 11(a), at 10 min, the Top-1 accuracies of DTRSF-Net were 95.04% (partial occlusion), 96.15% (lighting changes), and 95.49% (cluttered background), with the lighting change condition consistently the highest and partial occlusion the lowest. This demonstrates that DTRSF-Net achieves a more comprehensive aggregation of temporal context and key discriminative cues, enabling it to use longer time windows to smooth instantaneous feature shifts caused by local occlusion or background noise. As shown in Figure 11(b), at 10 min, the Top-1 accuracies of 3DResNet-50 under partial occlusion, lighting changes, and cluttered background were 89.16%, 90.72%, and 89.57%, respectively. It is observed that the SlowFast-YOLOv8 model exhibits a certain degree of interference resistance. However, the upper bound of the accuracy of the model under strong interference was greatly limited. As shown in Figure 11(c), under the three conditions of partial occlusion, lighting changes, and cluttered background, the Top-1 accuracy of 3D ResNet-50 starts at the lowest level across all three conditions and improves at a slower rate, reaching 84.39% (partial occlusion), 85.81% (illumination variation), and 84.85% (cluttered background) by 10 minutes. Considering the impact of diverse user filming habits on recognition accuracy, this study further validated the recognition accuracy of each model from different camera angles and shooting distances, with results shown in Figure 12.

Figure 11

Comparison of Top-1 accuracy under occlusion and background interference.

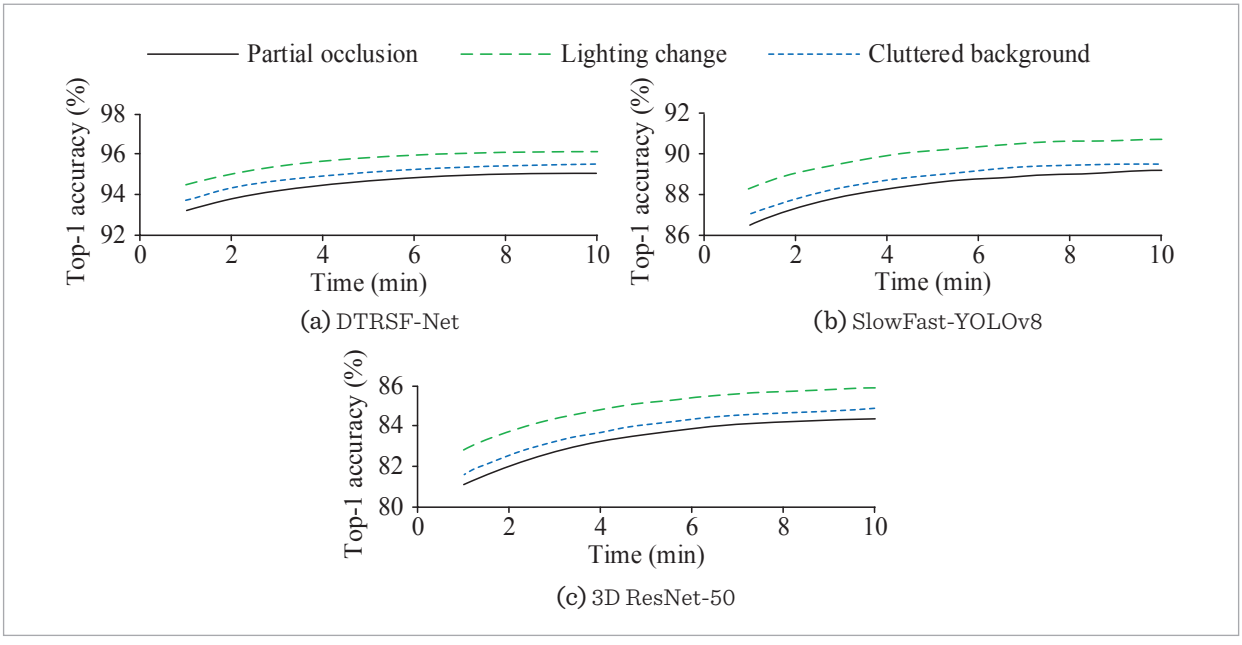
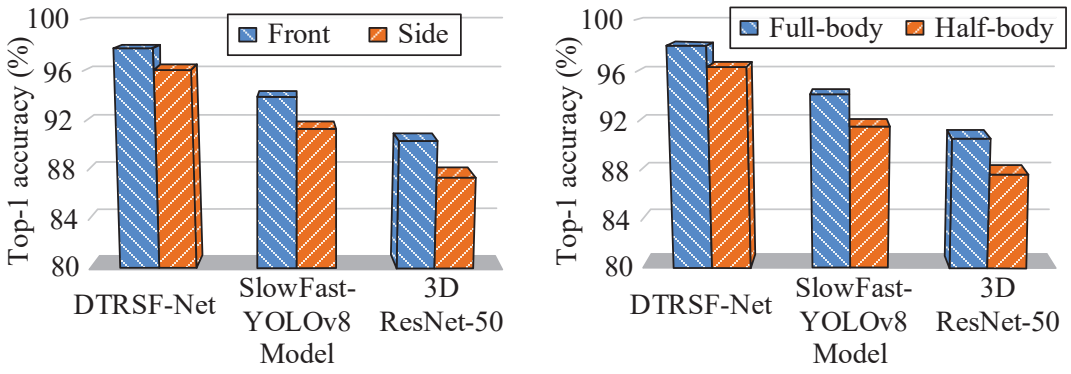


Figure 12
Comparison of Top-1 accuracy under different camera angles and distances.



(a) Comparison of Top-1 accuracy across different camera angles

(b) Comparison of Top-1 accuracy across different shooting distances

Table 3
Comparison of Top-1 accuracy in different venues (%).

Venue	Time (min)	DTRSF-Net	SlowFast-YOLOv8	3D ResNet-50
Gym	1	96.02	91.45	87.34
	2	96.53	92.12	88.15
	3	96.88	92.61	88.75
	4	97.12	92.97	89.19
	5	97.30	93.23	89.53
	6	97.42	93.43	89.77
	7	97.50	93.57	89.96
	8	97.56	93.67	90.10
	9	97.60	93.74	90.20
	10	97.63	93.80	90.28
Home	1	93.87	87.90	82.99
	2	94.48	88.68	83.91
	3	94.91	89.25	84.60
	4	95.21	89.66	85.11
	5	95.41	89.97	85.49
	6	95.56	90.19	85.77
	7	95.66	90.35	85.98
	8	95.73	90.47	86.14
	9	95.78	90.56	86.26
	10	95.82	90.63	86.34

In Figure 12(a), DTRSF-Net achieved Top-1 accuracies of 97.79% and 96.13% for frontal and side views, respectively. In comparison, SlowFast-YOLOv8 decreased from 93.96% to 91.40%, and 3D ResNet-50 decreased from 90.45% to 87.47%. DTRSF-Net led SlowFast-YOLOv8 and 3D ResNet-50 in both angles because its action discrimination relied more on relatively stable cues such as spatiotemporal dynamics and pose structure rather than strongly depending on appearance texture under specific viewpoints, resulting in smaller performance degradation under joint occlusion and morphological projection changes caused by side views. In Figure 12(b), the trends of the three models were consistent, with Top-1 accuracy higher for full-body than half-body shots. DTRSF-Net had achieved 97.99% and 96.33% Top-1 accuracy on full-body shots and half-body shots, respectively. For SlowFast-YOLOv8, these were 94.16% and 91.65%, while for 3D ResNet-50, these were 90.65% and 87.72%. This indicated that DTRSF-Net could more effectively aggregate key action segments and relative motion information of key joints within the locally visible region without relying on full body context or remote information from the limbs. To verify the generalization performance of the model under various application conditions, Table 3 lists the change in Top-1 accuracy for each model under two different venues: gym and home.

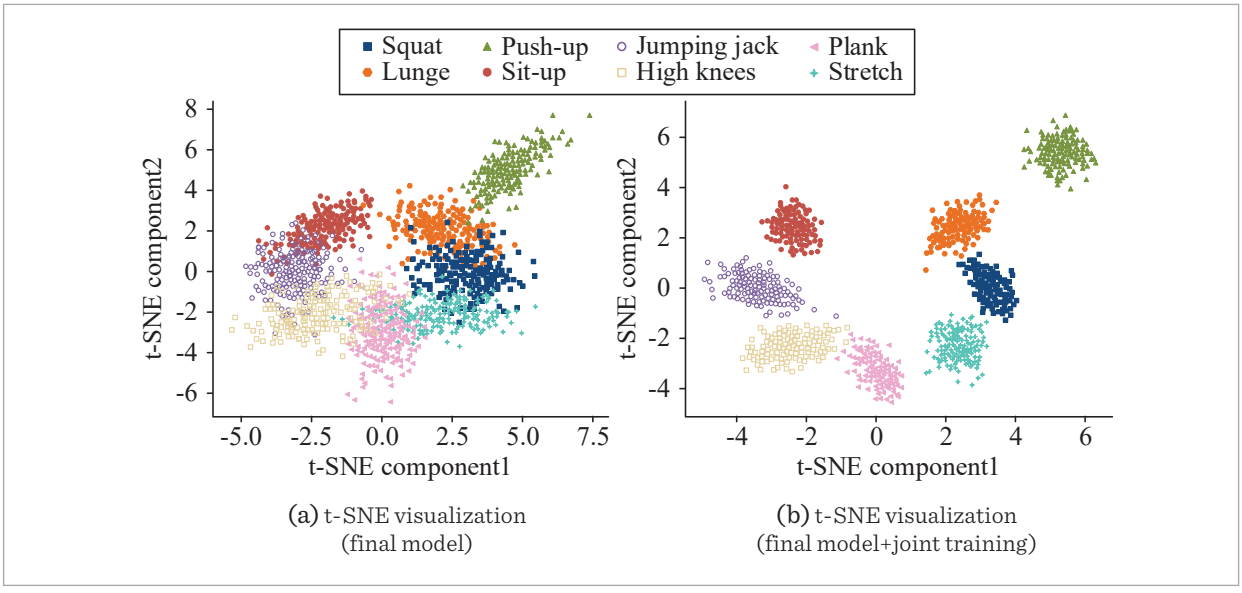
In Table 3, the Top-1 accuracy of the three models increases steadily and stabilizes with the extension of the observation duration from 1 min to 10 min. DTRSF-Net increased from 96.02% to 97.63% in the gym and from 93.87% to 95.82% at home. The comparative model SlowFast-YOLOv8 increased from 91.45% to 93.80% in the gym and from 87.90% to 90.63% at home. 3D ResNet-50 increased from 87.34% to 90.28% in the gym and from 82.99% to 86.34% at home. DTRSF-Net suppressed interference caused by lighting, background, and viewpoint differences through more comprehensive spatiotemporal feature modeling, skeleton structural information, and reliability-guided fusion. Therefore, under both venue types and all settings, DTRSF-Net maintained the highest accuracy, demonstrating more stable cross-scenario recognition capability.

3.5. Feature Space Separability Analysis

Finally, to verify the enhancement effect of the joint training strategy on feature space separability, this study visualized and compared the feature distribution and clustering effect of DTRSF-Net before and after incorporating the joint training framework using t-SNE technology. The results are shown in Figure 13.

As can be seen in Figure 13(a), there were obvious adjacency and overlapping phenomena among multiple class samples, and the intra-class distributions were

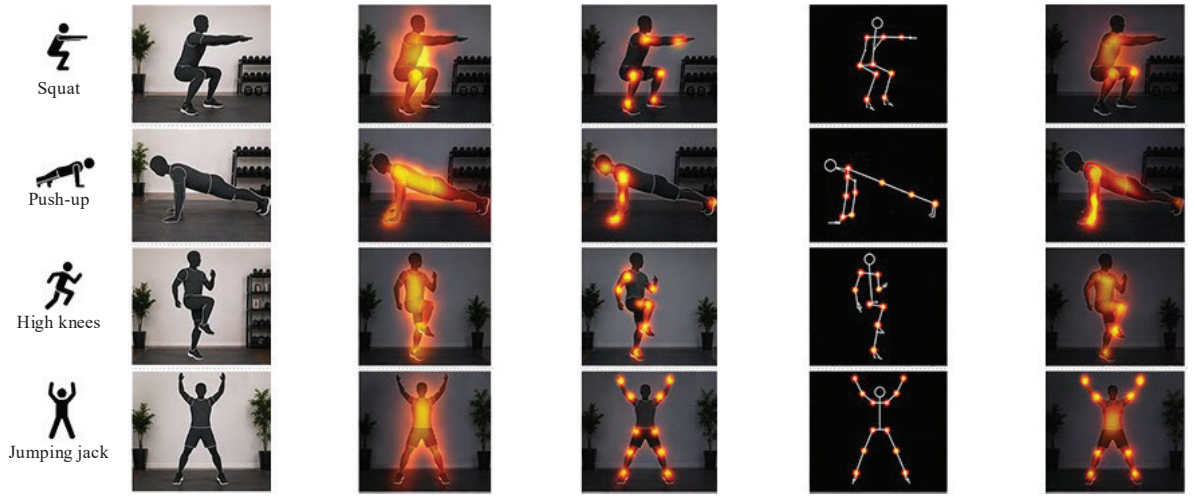
Figure 13
t-SNE visualization comparison of DTRSF-Net.



(a) t-SNE visualization (final model)

(b) t-SNE visualization (final model+joint training)

Figure 14
Interpretability heatmaps.



relatively loose, especially with blurred boundaries among similar action categories. It indicated that the model of DTRSF-Net had learned some class structures, yet the discriminative intervals in the embedding space were still not sufficient and might cause confusion among similar actions. This was because feature learning was more likely to be dominated by appearance texture and local rhythm under a single training paradigm, resulting in larger intra-class variance and smaller inter-class intervals, hence resulting in clusters of point clouds that were not compact enough and were mutually invasive. In Figure 13(b), there were more compact intra-class clusters and clearer inter-class intervals, with fewer outliers and better inter-cluster separation, and similar categories of actions were easier to be separated. This showed that joint training played a regularization and structuring role in representation learning by introducing additional constraint signals, improving the separability and robustness of the embedding space, enabling DTRSF-Net to form a more stable discriminative structure at the feature level. To more clearly show the attention regions and decision basis of the proposed model in fitness action recognition, this study further generated gradient-based activation heatmaps for the slow pathway, the fast pathway, and the fused output, as shown in Figure 14.

As shown in Figure 14, the slow pathway mainly focuses on the trunk, hips, and the overall posture contour. This demonstrates its effectiveness in maintaining coarse posture continuity across adjacent frames. In contrast, the fast pathway produces stronger activations in rapidly changing limb regions, including the wrists, elbows, knees, and lower legs. This shows that it is more sensitive to short-term motion bursts and phase reversals in repetitive fitness actions. After the skeleton branch is added, the model attention becomes more concentrated on functionally relevant joints, and the influence of background objects is reduced. The fused response map further confirms that the proposed framework benefits from complementary temporal-scale cues and cross-modal structural guidance. This explains its stronger robustness under occlusion, viewpoint changes, and cluttered home scenes.

4. Discussion

4.1. Analysis of Recognition Accuracy and Multimodal Collaboration Advantages

From the experimental results, the Top-1 accuracy of the proposed DTRSF-Net was 98.04%, and the Top-5 accuracy was 99.70%, which was higher than that

of SlowFast-YOLOv8 (Top-1 accuracy of 94.29%, Top-5 accuracy of 98.40%) and that of the 3D ResNet-50 (Top-1 accuracy of 90.82%, Top-5 accuracy of 96.79%). This showed that the proposed method was more comprehensive in the spatiotemporal representation and discriminative boundary learning. Compared to the method proposed by Lan et al. that employed the Skeleton-RGB dual-branch structure and strengthened the modal coupling effect through the spatiotemporal attention mechanism, the method proposed in this study emphasized the attention-driven cross-modal interaction more, while the accuracy was improved more directly using the lightweight fusion and training strategy [8]. Under different viewpoints and framing ranges, DTRSF-Net achieved Top-1 accuracies of 97.79% and 96.13% for frontal and side shots, respectively, and 97.99% and 96.33% for full-body and half-body shots, respectively. The accuracy degradation amplitude under viewpoint or cropping changes was smaller than comparison models, indicating better adaptability to missing visible joints and projection changes. The method proposed by Gunasekara et al. [5] considered "selectivity of temporal aggregation" as the central point of gain, which was different from the complementary constraints of dual temporal scale video and skeleton structural information. However, from the point of view of temporal modeling, the method of "organizing multi-scale spatiotemporal information at action granularity" in the proposed model was consistent with the research of Chen et al. [2].

4.2. Discussion on Fine-Grained Action Discrimination and Skeleton Graph Modeling

The confusion matrix results showed that DTRSF-Net had the highest action recognition accuracy, with relatively high misrecognition only manifested as High knees → Jumping jack (8 cases), while SlowFast-YOLOv8 had 23 cases for High knees → Jumping jack and 22 cases for Jumping jack → High knees. 3D ResNet-50 reached 36 cases for the same pair and had more action recognition errors such as Push-up → Plank with 32 cases, indicating that the proposed model had more stable differentiation for similar rhythms and similar support postures. Compared with the STANet approach of Li et al. [12], which explicitly characterized motion trends using bidirectional temporal differences to balance static semantics and dynamic changes, their method treated "difference-driven key motion extraction" as the core, while this study indirectly achieved emphasis on key dynamics through multi-temporal-scale video branches and skeleton structural priors. From the perspective of the skeleton graph model, the current branch of the skeleton graph may have the ability to further reduce the confusion of similar action pairs. Compared with the DeGCN proposed by Myung et al., the method proposed by the authors of the present study adopted a more fixed skeleton topology and cross-modal fusion to enhance separability, whereas the DeGCN proposed by Myung et al. [14] adopted a more flexible skeleton structure and spatiotemporal sampling positions through deformable graph convolution. On the other hand, if the skeleton branch is improved using a more powerful structural GCN to fully exploit the high-dimensional structural information and alleviate the over-smoothing problem caused by the aggregation of multiple scales, the MSS-GCN and graph weight annealing proposed by Jang et al. provided a pathway of "structural correlation + graph scheduling," whereas the main contribution of the present study came from the cross-modal complementarity [7].

4.3. Analysis of complex environment robustness and generalization performance

Under the interference of occlusion, lighting, and cluttered background, the Top-1 accuracy of DTRSF-Net and other comparison models showed a trend of increase with time and a certain degree of saturation. However, the curve of DTRSF-Net remained the highest among all, at 95%-96%, and was higher than those of other models, namely, 89%-91% for SlowFast-YOLOv8 and 84%-86% for 3D ResNet-50 at 10 min. This further showed that the discriminative features of DTRSF-Net were more stable under complex background interference and occlusion. Compared with the SDP model presented by Dai et al. [3], where background interference was suppressed by "diversity prototype learning" for first-person view scenarios and adaptively handled for different action durations, the model in this study improved its anti-interference capability by the complementary relationship between skeleton structural cues and video spatiotemporal cues. Cross-venue results further reflected generalization differences. In two scenarios, gym and home, DTRSF-Net improved from 96.02% to 97.63% and from 93.87% to 95.82%, respectively, from 1 min

to 10 min, and consistently led SlowFast-YOLOv8 and 3D ResNet-50 at all durations. Compared with Skeleton-CutMix of Liu et al. [13], which performed supervised domain adaptation in a "mixed domain" through skeleton swapping to reduce domain shift, their method treated cross-domain augmentation as the core, while this study currently mainly relied on implicit robustness brought by conventional training and fusion. Considering that the domain shift in home scenarios was stronger, if generalization capability at home were to be improved without relying on large amounts of annotation, self-supervised hierarchical spatiotemporal mask contrastive learning provided a more suitable direction for pre-training reinforcement. Compared with Cao et al. [1], who learned transferable representations through semantic-guided masks and cross-layer distillation, this study still focused mainly on supervised learning and could use their paradigm to enhance cross-scenario transferability in the future. Meanwhile, as far as the problem of difficulty in distinguishing similar actions is concerned, the strategy adopted by the micro-action recognition field of enhancing category semantic separation by joint embedding loss is very inspiring for this study. In comparison to Guo et al. [6], who adopted a strategy of incorporating a form of semantic matching-style joint embedding loss into MANet to enhance the category boundaries between fine-grained classes, this study mainly relied on a strategy involving feature fusion and spatiotemporal modeling, which could be enhanced by incorporating category semantics in the future to suppress confusion between similar classes.

4.4. Limitations and Case Analysis

Although DTRSF-Net performs better than the compared models in overall recognition accuracy, online stability, and robustness in complex scenes, the experimental results reveal that the most prevalent limitation of DTRSF-Net is still the residual confusion in transition frames between similar actions. This is especially shown in the unstable discrimination between High knees and Jumping jack at the moment of action switching. Compared with complete action phases, these transition frames usually keep only part of the intermediate posture features. For example, the trunk may still stay upright, the lower limbs may already show an upward motion trend, but the upper limbs may not have fully opened or returned yet. As a result, different actions share similar local shapes in a short time. Under these circumstances, the slow pathway maintains the continuity of the overall posture contour, yet remains susceptible to misclassifying these intermediate states as similar low-frequency posture patterns. The fast pathway is more sensitive to sudden limb motion, but when the transition is very short and local blur or incomplete framing appears at the same time, the key discriminative cues may not be fully encoded in time. At the same time, the skeleton branch is also easily affected by rapid changes in joint angles and fluctuations in keypoint localization in such frames. This further compresses the class boundaries in the shared embedding space, and finally leads to residual misclassification between similar actions.

The current framework is better at recognizing typical action patterns in stable phases. But for short intermediate states during phase switching, it still lacks more fine-grained explicit constraints. From the platform application view, this kind of error may not last for a long time, but it can directly affect the boundary of action counting and the timing of online correction prompts. Thus, it is more important than ordinary static misclassification. In future research, auxiliary supervision for action phases or explicit modeling of transition states could be introduced. This can separate complete actions, transition actions, and recovery actions, and then further reduce the discriminative overlap between actions with similar rhythm.

5. Summary and Outlook

The promotion of online training services on national fitness platforms has increased the requirements for action recognition accuracy. However, under real-scenario perturbations, the recognition stability and cross-domain usability of single-modality video models remain major bottlenecks. Therefore, this study constructed the DTRSF-Net, which coordinated RGB dual-temporal-scale branches and a skeleton branch. The proposed model performs post-fusion decision-making through a fusion head and incorporates a joint training strategy to improve cross-modal feature alignment, aiming to enhance the stable output capability required for online platform applications. The experimental results indicated that under the interference of complex environ-

ments such as partial occlusion, changes in lighting, and messy backgrounds, the model had a higher accuracy in recognizing images. At the same time, in the test of variations under different camera angles and shooting distances, the model also presented an extremely high level of system robustness. In conclusion, through multi-modal collaboration and joint training, the model achieves enhanced robustness when addressing complex environments and usability in real-time error correction and counting in the platform side.

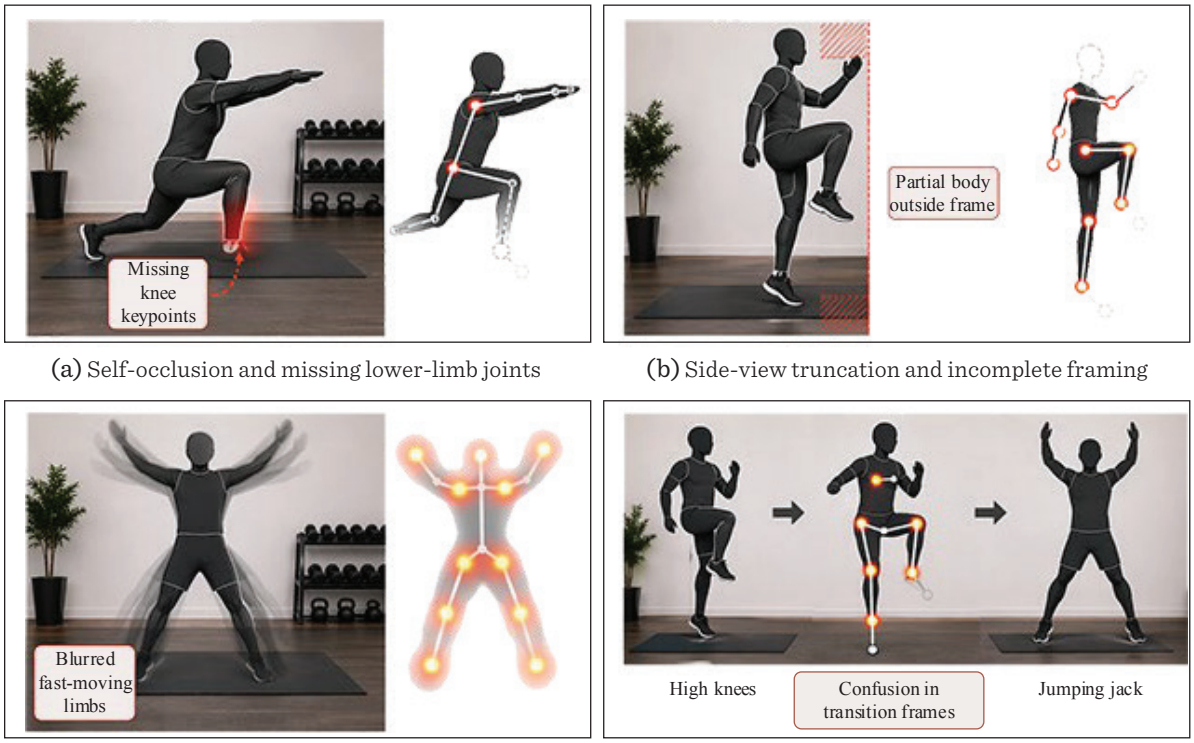
Funding

The research is supported by Hebei Province Federation of social science circles: The Approach to Reshaping the Cultural Ecology of Rural Martial Arts in the Context of Rural Cultural Revitalization (Grant No. 202402242).

Appendix A.

Supplementary Figure S1

Representative failure cases.



(a) Self-occlusion and missing lower-limb joints

(b) Side-view truncation and incomplete framing

(c) Motion blur under rapid limb movement

(d) Transition ambiguity between similar actions

References

1. Cao, W., Zhang, A., He, Z., Zhang, Y., Yin, X. Hierarchical Spatial-Temporal Masked Contrast for Skeleton Action Recognition. IEEE Transactions on Artificial Intelligence, 2024, 5(11), 5801-5814. https://doi.org/10.1109/TAI.2024.3430260
2. Chen, Y., Ge, H., Liu, Y., Cai, X., Sun, L. AGPN: Action Granularity Pyramid Network for Video Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8), 3912-3923. https://doi.org/10.1109/TCSVT.2023.3235522

3. Dai, G., Shu, X., Yan, R., et al. SlowFast Diversity-Aware Prototype Learning for Egocentric Action Recognition. Proceedings of the 31st ACM International Conference on Multimedia, 2023, 7549-7558. https://doi.org/10.1145/3581783.3612144
4. Erdemir, E., Kaplanoglu, E., Uyanik, C., Akgun, G. Motion Trajectory Estimation for Hand Grasping States Using a Deep Learning Approach. SWT, 2025, 1, A5, 1-8. https://doi.org/10.47852/bonviewSWT52025659
5. Gunasekara, S. R., Li, W., Yang, J., Ogunbona, P. O. Asynchronous Joint Based Temporal Pooling for Skeleton Based Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(1), 357-366. https://doi.org/10.1109/TCSVT.2024.3465845
6. Guo, D., Li, K., Hu, B., Zhang, Y., Wang, M. Benchmarking Micro-Action Recognition: Dataset, Methods, and Applications. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7), 6238-6252. https://doi.org/10.1109/TCSVT.2024.3358415
7. Jang, S., Lee, H., Kim, W. J., Lee, J., Woo, S., Lee, S. Multi-Scale Structural Graph Convolutional Network for Skeleton-Based Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8), 7244-7258. https://doi.org/10.1109/TCSVT.2024.3375512
8. Lan, X., Zhang, J., Zhou, J. Dual Branch Skeleton-RGB Fusion for Human Action Recognition with Spatio-Temporal Attention. Alexandria Engineering Journal, 2025, 129(4), 205-215. https://doi.org/10.1016/j.aej.2025.05.083
9. Le, H., Lu, C. K., Hsu, C. C., Huang, S. K. Skeleton-Based Human Action Recognition Using LSTM and Depthwise Separable Convolutional Neural Network. Applied Intelligence, 2025, 55(5), 298. https://doi.org/10.1007/s10489-024-06082-w
10. Li, C., Li, S., Gao, Y., Chen, P., Li, J., Li, W. Unsupervised Feature Enrichment and Fidelity Preservation Learning Framework for Skeleton-Based Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(7), 7116-7128. https://doi.org/10.1109/TCSVT.2025.3540292
11. Li, M., Chen, K., Bai, Y., Pei, J. Skeleton Action Recognition Via Graph Convolutional Network with Self-Attention Module. Electronic Research Archive, 2024, 32(4), 2848-2864. https://doi.org/10.3934/era.2024129
12. Li, Z., Li, J., Ma, Y., Wang, R., Shi, Z., Ding, Y., Liu, X. Spatio-Temporal Adaptive Network with Bidirectional Temporal Difference for Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9), 5174-5185. https://doi.org/10.1109/TCSVT.2023.3250646
13. Liu, H., Liu, Y., Mu, T. J., Huang, X., Hu, S. M. Skeleton-CutMix: Mixing Up Skeleton with Probabilistic Bone Exchange for Supervised Domain Adaptation. IEEE Transactions on Image Processing, 2023, 32, 4046-4058. https://doi.org/10.1109/TIP.2023.3293766
14. Myung, W., Su, N., Xue, J. H., Wang, G. DeGCN: Deformable Graph Convolutional Networks for Skeleton-Based Action Recognition. IEEE Transactions on Image Processing, 2024, 33, 2477-2490. https://doi.org/10.1109/TIP.2024.3378886
15. Qiu, H., Hou, B., Ren, B., Zhang, X. Spatio-Temporal Segments Attention for Skeleton-Based Action Recognition. Neurocomputing, 2023, 518, 30-38. https://doi.org/10.1016/j.neucom.2022.10.084
16. Sun, K., Xiao, B., Liu, D., Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 5693-5703. https://doi.org/10.1109/CVPR.2019.00584
17. Tong, J., Wang, F. Basketball Sports Posture Recognition Technology Based on Improved Graph Convolutional Neural Network. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2024, 28(3), 552-561. https://doi.org/10.20965/jaciii.2024.p0552
18. Wu, Z., Ding, Y., Wan, L., Li, T., Nian, F. Local and Global Self-Attention Enhanced Graph Convolutional Network for Skeleton-Based Action Recognition. Pattern Recognition, 2025, 159, 111106. https://doi.org/10.1016/j.patcog.2024.111106
19. Xiang, L., Li, J., Wang, Z. Least-Squares Estimation of Keypoint Coordinate for Human Pose Estimation. In: Pattern Recognition and Computer Vision, PRCV 2022. Lecture Notes in Computer Science, vol. 13536. Springer, Cham, 2022, 448-460. https://doi.org/10.1007/978-3-031-18913-5_35
20. Xu, F., Shi, P., Zhang, X. Skeleton-Based Human Action Recognition with Spatial and Temporal Attention-Enhanced Graph Convolution Networks. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2024, 28(6), 1367-1379. https://doi.org/10.20965/jaciii.2024.p1367
21. Zhang, H., Zhang, X., Wang, D., Zhou, F., Yan, J. MCNet: A Unified Multi-Center Graph Convolutional Network Based on Skeletal Behavior Recognition. Alexandria Engineering Journal, 2025, 120(2), 116-127. https://doi.org/10.1016/j.aej.2025.01.118
22. Zhou, T., Li, L., Chen, L., Wang, Y., Liu, Z., Liu, J., Sun, L. Topology Learning by Context Embedding and Channel Refinement for Skeletal Behavior Recognition. IEEE Access, 2024, 12, 181106-181115. https://doi.org/10.1109/ACCESS.2024.3492332

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).