Ice and Snow Sports Behavior Recognition Technology Based on EM Algorithm and RepVGG Neural Network

Shu Yu
School of Winter Olympics, Harbin Sport University, China
15146618866@163.com

Jianxin Kang
School of Winter Olympics, Harbin Sport University, China
kangjianxin1102@163.com

Dawei Wu
School of Winter Olympics, Harbin Sport University, China
wudawei202506@163.com

Wang Wang
Graduate School, Harbin Sport University, China
17824940399@163.com

Abstract: *With the rapid development of intelligent sports analysis technology, automatic action recognition has become a key technology for performance evaluation and safety monitoring in winter sports. This study aims to improve the automatic recognition of complex behaviors in ice and snow sports and meet practical needs such as intelligent training and assisted judgment. The study proposes a behavior recognition method based on the Expectation-Maximization (EM) algorithm and the Re-parameterized Visual Geometry Group network (RepVGG). First, a multi-modal input feature channel is constructed by fusing residual maps and pose heatmaps extracted from video frames. This enhances the model's structural perception of the athlete and reduces background interference. Second, an Expectation-Maximization-based Refinement (EM-R) module is designed to perform soft clustering on frame-level features. By learning latent class distributions, this module introduces semantic priors to improve the classification capability of the subsequent backbone network. Finally, multi-scale convolution units and an EM-attention mechanism are integrated into the backbone to enhance the model's response to local dynamics and key action areas. The RepVGG-based architecture is further re-parameterized to enable efficient inference and lightweight deployment. Experiments are conducted on a skiing-related subset built from the public University of Central Florida (UCF101) dataset, including representative ice and snow behaviors such as "Skiing", "Ski Jumping", and "Snowboarding". Results show that the proposed method achieves 92.7% classification accuracy and a macro-average F1-score of 90.4% on the test set, outperforming the original RepVGG by 3.4% and 3.3%, respectively. The inference speed reaches 110 Frames Per Second (FPS), and real-time performance of 45 FPS is achieved on Jetson NX devices, demonstrating strong deployment capability and platform adaptability. Ablation studies confirm the critical roles of the EM-R module, multi-scale structure, and pose channel in improving overall performance.*

Keywords: *Ice and snow sports, Behavior recognition, RepVGG, EM algorithm, multimodal feature.*

Received July 09, 2025; accepted January 15, 2026
https://doi.org/10.34028/iajit/23/5/2

1. Introduction

With the promotion of international events such as the Winter Olympics, ice and snow sports have seen a continuous rise in global participation and viewership, leading to a growing demand for professional and intelligent sports technologies [17]. High-intensity disciplines such as skiing, skating, and ski jumping exhibit significant characteristics in terms of competitiveness and technical complexity, placing greater demands on the accurate capture and intelligent analysis of motion data [24]. Against this backdrop, behavior recognition technology has become a key component of intelligent sports systems and is widely used in scenarios such as action evaluation, training feedback, and safety warning systems [15]. Compared to conventional sports behavior recognition, ice and snow sports involve more complex spatial action combinations and more intense dynamic processes. Therefore, building a high-performance and real-time behavior recognition system for ice and snow sports not only holds theoretical research value but also provides practical benefits for athlete training support, referee decision assistance, and enhanced event broadcasting [9].

However, behavior recognition in ice and snow sports faces multiple technical challenges. First, these activities are often performed under conditions involving high-speed movement, complex lighting, and multi-angle camera capture, which can easily lead to image blur, partial occlusion, and pose distortion, thereby increasing the difficulty of distinguishing behavior categories [22]. Second, individual differences in movement rhythm, trajectory, and body structure introduce significant variability in the expression of similar actions, severely limiting the generalization capability of recognition models [28]. In addition, the highly uniform backgrounds of ice and snow

environments can interfere visually with the moving subjects, increasing the risk of background-related misclassification [13]. Existing behavior recognition methods mainly focus on 2D Convolutional Neural Network (CNN), 3D temporal modeling, and attention mechanisms. Although some progress has been made in specific tasks, these methods still face limitations under the extreme conditions of ice and snow sports, such as insufficient accuracy or low inference efficiency, making it difficult to meet the dual requirements of real-time deployment and high precision. Therefore, it is essential to explore recognition approaches that offer both strong structural representation and computational efficiency to accommodate the complexity of ice and snow sport behaviors.

To address the above challenges, this study proposes a novel framework for ice and snow sports behavior recognition that integrates the Expectation-Maximization (EM) algorithm with a Re-parameterized Visual Geometry Group network (RepVGG). The method incorporates multimodal feature inputs, where residual maps are used to enhance motion boundaries, and pose estimation is employed to generate human body structure heatmaps. This design enhances the model's spatial awareness of complex movements. For feature modeling, an Expectation-Maximization-based Refinement module (EM-R) is introduced to perform soft clustering on frame-level features. This module explicitly models the posterior distribution of latent behavior categories and embeds category-aware guidance into the backbone network. Regarding the backbone architecture, a multi-scale re-parameterized convolutional module is built upon the RepVGG network. A channel attention mechanism driven by behavioral priors is further introduced to enhance the network's response to key motion areas. The proposed method is trained in an end-to-end manner, jointly optimizing clustering representation and classification. It effectively balances recognition accuracy and computational cost, making it suitable for behavior analysis in high-dynamic and high-interference scenarios typical of ice and snow sports. Existing research on action recognition in winter sports usually focuses only on a single technical direction, such as convolutional networks based on Red-Green-Blue (RGB) features, shallow models based on skeletons, or deep structures based on temporal modeling. It lacks systematic modeling of the synergistic relationships among "background interference-posture structure-action semantics" in complex scenarios. Different from existing methods, this study does not simply superimpose multiple technical modules. Instead, centering on the characteristics of winter sports such as high speed, strong interference and ambiguous action phases, it constructs a unified recognition framework jointly driven by background removal, structure perception, semantic clustering and efficient network structure.

The structure of the remaining parts of this study is arranged as follows. Section 2 systematically reviews action recognition in winter sports and related deep learning methods, and analyzes the shortcomings of existing research. Section 3 details the proposed winter sports action recognition method based on the EM algorithm and RepVGG network, including multi-modal feature construction, EM-R clustering mechanism, and improved network structure design. Section 4 verifies the method performance through comparative experiments and ablation experiments, and analyzes the real-time performance and deploy ability of the system. Section 5 conducts an in-depth discussion on the experimental results. Section 6 summarizes the full text and prospects future research directions.

2. Related Works

In recent years, with the widespread application of computer vision in sports analytics, behavior recognition in ice and snow sports has emerged as a research hotspot [26, 29]. Chen *et al*. [4] proposed the Long and Short-term temporal difference Vision Transformer (LS-ViT) framework, which integrates short-term and long-term motion information into a Vision Transformer (ViT) architecture for action recognition. The method effectively enhances temporal feature representation and improves recognition performance for complex motion sequences. However, it mainly focuses on temporal dependency modeling and does not explicitly address the challenges of highly repetitive backgrounds, pose ambiguity, and semantic clustering in ice and snow sports scenarios. Ding *et al*. [7] constructed local motion energy maps from ski-jumping video sequences and applied temporal filters to enhance behavioral stability. Nonetheless, the model's performance degraded significantly under high-speed transformations and motion blur conditions. Hu *et al*. [11] introduced a shallow network model based on skeleton tracking to classify gliding and falling behaviors on ice. While this approach improved responsiveness to abnormal actions, its overall accuracy was constrained by skeleton extraction errors. Overall, most existing studies focus on small-scale, single-category behavior recognition and lack end-to-end modeling approaches for complex scenarios. In particular, the generalizability and deploy ability of current methods in real-world ice and snow environments remain notably limited.

Deep learning techniques offer powerful modeling capabilities for video-based action recognition. Existing methods are mainly categorized into three types: CNN, Recurrent Neural Network (RNN), and Transformer-based architectures. Kareem *et al*. [12] were among the first to propose the 3D-CNN for spatiotemporal modeling of video data, effectively capturing inter-frame temporal variations and laying a foundation for subsequent research. Yang *et al*. [31] introduced the

Inflated 3D ConvNet (I3D), which extended 2D convolutional kernels to 3D. This model achieved remarkable performance on the Kinetics dataset, but suffered from high computational cost and low inference speed, making it unsuitable for real-time applications. Zhao [34] proposed the Temporal Segment Network (TSN), which performed lightweight temporal modeling by sampling key frame segments. However, it had limited capability in capturing action boundaries and temporal continuity. In the RNN domain, Alhakbani *et al*. [2] introduced the Long-term Recurrent Convolutional Network (LRCN), which integrated CNN with Long Short-Term Memory (LSTM) network to model long behavior sequences. Nonetheless, it faced challenges such as gradient vanishing and unstable training. Recently, transformer-based architectures have been increasingly applied to action recognition. Cheng *et al*. [5] proposed the TimeSformer, which demonstrated strong performance in modeling long sequences. However, its high training cost and hardware dependency limited its practical deployment. Thus, developing efficient and lightweight recognition frameworks remains a key research focus, especially for complex scenarios.

RepVGG is a CNN architecture optimized for deployment, characterized by a "multi-branch during training, single-branch during inference" structural re-parameterization design [3]. Its core idea is to enhance representational capacity during training by introducing parallel 3×3 convolutions, 1×1 convolutions, and identity mappings. During inference, these branches are merged into an equivalent single convolution layer, significantly improving inference efficiency. This structure has demonstrated performance comparable to or even better than ResNet in tasks such as image classification and face detection [25]. Building on RepVGG, Han and Li [10] introduced multi-scale convolutional kernel combinations to enhance perception of small targets and local structures. Liu *et al*. [18] explored the application of RepVGG in action recognition tasks and incorporated a channel attention mechanism to improve the model focus on action-relevant regions. Despite its advantages in architectural design and engineering deployment, RepVGG still faces limitations in semantic feature modeling for action recognition. In particular, it lacks mechanisms to incorporate action category distributions or semantic priors into its framework.

The EM algorithm is a classical statistical method for handling latent variable problems and has been widely applied in tasks such as image processing and clustering modeling. Lopez-Lozada *et al*. [19], introduced the EM mechanism into deep clustering networks to optimize the semantic distribution in the feature space, guiding the network to learn more compact class boundaries. In the field of action recognition, Muhammad *et al*. [21], attempted to use the EM process to model latent behavioral states across video frames, enhancing the discrimination of anomalous behaviors. However, their model was structurally complex and exhibited unstable training, making it difficult to deploy in real-world systems. In summary, while the EM mechanism demonstrates strong class modeling capabilities and interpretability, further research is needed to develop efficient integration strategies with neural networks in deep action recognition frameworks.

The literature summary in the literature review section is shown in Table 1:

Table 1. Comparison and summary of related behavior recognition research.

Ref. No.	Research object/scene	Main method	Input mode/feature	Main conclusions (overview)	Limitations and research blank sources (used to export gap)
[4]	Ice and snow/skiing action stage (take-off, landing)	Manual feature+SVM	Motion/Statistical characteristics	Specific stages can be distinguished.	Strong background dependence and insufficient generalization of real and complex snow fields; lack of end-to-end learning
[7]	Snow/Ski jumping video sequence	Motion energy diagram+Time series filtering	RGB derived motion features	Improve behavioral stability	Insufficient robustness under high-speed blur and angle change; limited feature expression
[11]	Slide and fall on the ice, etc.	Skeleton tracking+Shallow network	Skeleton	More sensitive to abnormal behavior response	Skeleton error propagation is significant; lack of visual background suppression and multimodal fusion mechanism
[31]	General behavior recognition (long sequence/complex action)	I3D (3D-CNN extension)	RGB/Space-time body	Outstanding precision performance	Large amount of calculation and slow reasoning; it is difficult to meet real-time deployment requirements.
[34]	Motion recognition in physical education video	TSN (fragment sampling)	RGB fragment	Reasoning is relatively lightweight.	Insufficient modeling of action boundary and continuity; the high-speed stage is easy to be confused.
[2]	Universal video behavior recognition	LRCN (CNN+LSTM)	RGB sequence	Modelable long sequence dependence	Training instability and gradient problems; high deployment cost
[4]	Multi-modal motion recognition (cross-modal compensation)	Cross-modal compensation CNN	RGB-D /Multimodal	Multimodal can improve robustness	Strong dependence on acquisition conditions; there is no explicit treatment for snow and ice background interference.
[3]	Deployment optimization CNN (non-behavior recognition)	RepVGG variant	RGB	Re-parameterization of structure to improve reasoning efficiency	Behavior semantics/Category distribution modeling is not introduced; lack of action recognition scene verification
[25]	RepVGG is used to improve the classification task	RepVGG+Attention convolution	RGB	Improve feature focusing ability	Attention lacks "semantic transcendental" drive; not fused with clustering/hidden variable mechanism
[21]	General behavior recognition (abnormal/complex action)	Dilated CNN +Attention LSTM	RGB sequence	Strengthen the extraction of complex time series features	The network structure is complex and the reasoning efficiency is insufficient; lack of deployable design

Current research on ice and snow sports action recognition faces four major challenges:

1. Existing datasets are mostly limited to control experimental settings, leading to poor model generalization in real-world ice and snow environments.
2. Mainstream deep learning models tend to prioritize either accuracy or speed, lacking a unified architecture that balances both.
3. Although RepVGG offers deployment advantages, it lacks sufficient capacity for behavior-level semantic modeling.
4. While the EM algorithm performs well in image clustering, it has not yet been effectively integrated with re-parameterized network architectures.

Therefore, it is essential to develop a unified recognition framework that combines multimodal feature enhancement, semantic clustering modeling, and an efficient neural architecture. This framework aims to improve the modeling accuracy and inference efficiency of complex movement behaviors in real-world scenarios, which is the core issue addressed in this study.

3. Research Method of Ice and Snow Behavior Recognition Technology

This section focuses on the overall technical process of action recognition in winter sports, with an emphasis on the construction of multi-modal features, the modeling of action semantics, and the design of an efficient network structure. First, background suppression and posture structure enhancement are performed on the original video frames to construct input feature representations with stronger discriminative ability. Then, a hidden state clustering mechanism based on the EM algorithm is introduced to conduct semantic modeling of frame-level features, providing category prior guidance for the subsequent network. Finally, on this basis, an improved RepVGG backbone network is designed to balance recognition accuracy and inference efficiency. The above modules jointly constitute an end-to-end action recognition method for winter sports.

3.1. Multi-Modal Data Preprocessing and Feature Extraction

The original video sequence is expressed as a frame set $\mathcal{V} = F_t{}_{t=1}^{T}$, where $F_t \in \mathbb{R}^{H\times W\times 3}$ represents the image of the t frame. To reduce redundant information and unify the model input distribution, firstly, all frames are normalized in size and pixels, as shown in Equation (1):

$$\hat{F}_t(x,y,c) = \frac{F_t(x,y,c) - \mu_c}{\sigma_c}, c \in \{R,G,B\} \quad (1)$$

μ_c and σ_c are the global mean and standard deviation of channel c, respectively, to ensure the consistency of each frame in space and color domain. Equation (1) standardizes the three RGB channels to eliminate differences in lighting conditions and camera parameters among different video samples. This enables the network to focus more on the motion structure itself rather than changes in color distribution during training, thereby improving the stability of the model under cross-scenario conditions.

Snow and ice scenes are usually highly repetitive, such as snow, guardrails, audience areas, etc., which easily interfere with sports behavior characteristics [1, 32]. Therefore, this study introduces spatio-temporal background modeling and dynamic target separation mechanism, and adopts adaptive background removal algorithm based on multi-frame average model. The background model B_t is constructed as shown in Equation (2):

$$B_t = \frac{1}{k}\sum_{i=t-k}^{t-1} \hat{F}_i \quad (2)$$

The background estimation is formed by the average value of the previous k frames, and the dynamic region of the current frame is represented in the form of residual graph Δ_t, as shown in Equation (3):

$$\Delta_t = |\hat{F}_t - B_t| \quad (3)$$

The foreground mask M_t is obtained by threshold function I, as shown in Equation (4):

$$M_t(x,y) = \mathbb{I}(\Delta_t(x,y) > \theta) \quad (4)$$

θ is an empirical threshold, which is used to distinguish background from dynamic target. Parameter θ is used to control the separation intensity between the foreground and background, and its value directly affects the sparsity of the residual region. In practical applications, θ can be set according to the statistical characteristics of the residual distribution on the training set. It is usually selected within a threshold range that enables the foreground region to cover the main motion contours without excessively introducing noise. A smaller θ retains more dynamic details but may introduce background noise. A larger θ enhances the background suppression effect but may lead to the loss of fine-grained motion information. In the experiment, simple parameter tuning is performed on the validation set to select a threshold that balances recognition accuracy and stability, thereby ensuring the overall system performance.

The background modeling process described by Equations (2), (3), and (4) estimates static regions using inter-frame consistency within a short time window, thereby separating moving subjects from highly repetitive winter sports backgrounds. The generated residual map Δ_t can highlight motion boundaries and dynamically changing regions. It provides more discriminative input for subsequent feature extraction, and effectively reduces the impact of background interference on action recognition.

In addition, to enhance the model's understanding of spatial structure, a skeleton auxiliary channel based on attitude estimation is further introduced. A lightweight key point detection network is used to extract the human skeleton point set of each frame image, as shown in Equation (5):

$$\mathcal{P}_t = \{(x_j, y_j)\}_{j=1}^{J} \tag{5}$$

J is the number of key points. The key points are coded into the structural image H_t in the form of thermal map, and spliced with the residual image Δ_t as the input features of the enhanced structure, as shown in Equation (6):

$$\mathcal{X}_t = Concat(\Delta_t, \mathcal{H}_t) \tag{6}$$

3.2. The Behavior Hidden State Clustering Mechanism of EM Optimization

The spatiotemporal characteristics of ice and snow sports behaviors exhibit instability and phase ambiguity, making it difficult for single-frame features to accurately capture the complete behavioral state [14, 32]. To enhance the discrimination of complex actions, this study introduces a hidden state modeling mechanism based on the EM algorithm, enabling soft clustering and semantic prior enhancement in the feature space. Specifically, an EM-R module is designed to embed latent semantic distributions into the feature representation process, thereby assisting the RepVGG backbone in discriminative learning. By encoding human key points into posture heatmaps H_t and concatenating them with residual maps, the model can perceive the position changes of motion regions, and explicitly model human structural relationships. Such structure-aided information is particularly important for complex movements such as high-speed gliding, take-off and aerial twisting. It helps alleviate the recognition difficulties caused by posture ambiguity and view changes.

Given the preprocessed video frame feature sequence $Xtt=1^T$, the preliminary representation $z_t \in \mathbb{R}^d$ is extracted through the shallow convolution network as the input of EM clustering. Suppose there are K potential behavior categories, each category corresponds to the Gaussian distribution center μ_k and covariance matrix Σ_k. The behavior category is the hidden variables $Z_t \in 1, ..., K$, with the objective of maximizing the logarithmic likelihood of the observed data, as shown in Equation (7):

$$\mathcal{L} = \sum_{t=1}^{T} log\left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(z_t \mid \mu_k, \Sigma_k)\right) \tag{7}$$

π_k is the prior probability of class k, which satisfies $\sum_{k=1}^{K} \pi_k = 1$. The EM algorithm is used to iteratively optimize the model, and the steps are as follows:

E-step (Expectation): estimate the posterior probability that each sample belongs to category k based on the current parameters, as shown in Equation (8):

$$\gamma_{tk} = \frac{\pi_k \cdot \mathcal{N}(z_t \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(z_t \mid \mu_j, \Sigma_j)} \tag{8}$$

M-step (Maximization): update Gaussian parameters and prior distribution to maximize the likelihood under the current expectation;

$$\pi_k^{(new)} = \frac{1}{T}\sum_{t=1}^{T} \gamma_{tk} \tag{9}$$

$$\mu_k^{(new)} = \frac{\sum_{t=1}^{T} \gamma_{tk} z_t}{\sum_{t=1}^{T} \gamma_{tk}} \tag{10}$$

$$\sum\nolimits_k^{(new)} = \frac{\sum_{t=1}^{T} \gamma_{tk}(z_t - \mu_k)(z_t - \mu_k)^{\top}}{\sum_{t=1}^{T} \gamma_{tk}} \tag{11}$$

In Equations (9), (10), and (11), m inner loop iterations are carried out in each training batch. After stable convergence, the posterior probability matrix $\Gamma=\gamma_{tk}$ is introduced into the backbone network as a category awareness feature. Build an EM-R module, remap Γ to the channel dimension through convolution, and generate a feature enhancement map, as shown in Equations (12) and (13):

$$F_{EM} = \phi(\Gamma; \theta) \tag{12}$$

$$X_{refined} = \mathcal{X}_t \oplus F_{EM} \tag{13}$$

$\phi(\cdot)$ represents the EM channel mapping function and $\oplus$ represents the channel-level splicing operation. In essence, this module guides RepVGG backbone to pay attention to the differences between potential categories through semantic clustering, effectively suppresses background disturbance and category overlap, and improves the discrimination clarity of behavior classification boundaries.

3.3. Improved RepVGG Network Structure Design

To enhance both feature representation and inference efficiency in ice and snow sports behavior recognition, this study reconstructs the original RepVGG network by introducing a multi-scale convolutional fusion mechanism. Additionally, a prior-guided attention module is incorporated to strengthen the network's response to key action regions. The improved architecture maintains a simple convolutional form during inference while significantly enhancing multi-branch representational capacity during training. This achieves a synergistic modeling of structural re-parameterization and behavior-aware feature guidance [23].

The original RepVGG network adopts the following re-parameterization strategy: In the training stage, parallel 3×3 convolution, 1×1 convolution and identity mapping branches are used for feature learning. The output of each layer of the network can be expressed as Equation (14):

$$y = ReLU((w_{3\times3} * x) + (w_{1\times1} * x) + x) \tag{14}$$

During inference, the above multi-branch structure is

merged into an equivalent single-branch convolutional layer, balancing accuracy and efficiency. Based on this framework, this study proposes multi-scale rep units, which expand multi-scale parallel convolutional paths during training. Specifically, these include standard 3×3 convolutions to capture medium-scale action structures, dilated 5×5 convolutions to enhance perception of long-range poses, and deformable convolutions to adapt to abnormal pose variations and non-rigid motions.

Multi-scale outputs participate in structural reparameterization after weighted fusion channel by channel [33]. Let $F3\times3$, $F5\times5$ and F_{def} represent three kinds of convolution path outputs respectively, then the fusion feature is Equation (15):

$$y_{multi} = \alpha \cdot \mathcal{F}_{3\times3} + \beta \cdot \mathcal{F}_{5\times5} + \gamma \cdot \mathcal{F}_{def} \quad (15)$$

$\alpha+\beta+\gamma=1$. The weights are trained and learned, and finally mapped into equivalent single convolution form to participate in inference.

EM-attention mechanism is designed, and the posterior probability matrix Γ is introduced into the attention module of the backbone network channel [20, 27]. The specific form is as follows:

Let the convolution feature map be $X\in\mathbb{R}^{C\times H\times W}$ and the category posteriori be $\Gamma\in\mathbb{R}^{K}$, then the channel attention weight is generated as shown in Equation (16):

$$A_c = \sigma(W_2 \cdot ReLU(W_1 \cdot \Gamma)), X'_c = A_c \cdot X_c \quad (16)$$

$W_1\in\mathbb{R}^{d\times K}$ and $W_2\in\mathbb{R}^{C\times d}$ is the attention parameter matrix. $\sigma(\cdot)$ is the sigmoid function. Finally, the enhanced feature map X' remains unchanged in the spatial dimension, and only the channel response is prior modulated by EM clustering, thus realizing the mapping process of "behavioral semantics→channel activation".

3.4. Overall Framework of Ice and Snow Sports Behavior Recognition

To achieve accurate and rapid recognition of ice and snow sports behaviors, an end-to-end recognition system is constructed based on the EM algorithm and an enhanced RepVGG architecture. The overall system follows a modular design principle and consists of four main components: the input layer, a multimodal feature aggregation layer (integrating the EM-R module), a behavior discrimination backbone network, and a behavior classification output layer. These modules are seamlessly connected via standard tensor interfaces, facilitating model deployment and training optimization, as illustrated in Figure 1.

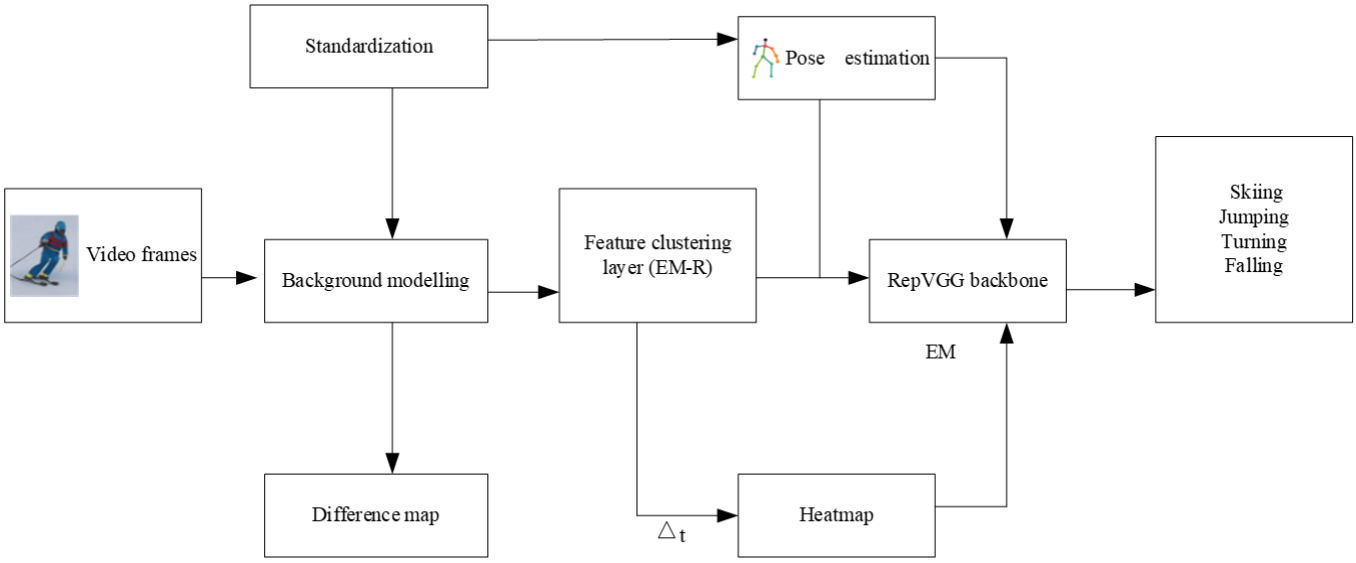

Figure 1. Identification system architecture.

The input of the system is the standardized video frame sequence $\hat{F}_t$. The multi-modal input feature Δ_t is constructed by combining the residual image H_t and the attitude heat map X_t, and then enters the feature clustering layer for soft clustering operation guided by behavior semantics. This module clusters and optimizes the frame-level features in each training period, and generates a posterior probability matrix Γ. It is used to guide the subsequent networks to pay attention to the differences between potential behavior categories.

The cluster-enhanced feature vectors are fed into the backbone network, which consists of multiple re-parameterized convolutional modules. During training, the network adopts a multi-branch, multi-scale structure to learn diverse semantic responses. In the inference stage, it is simplified into an efficient single-branch convolutional network. Additionally, the EM-attention mechanism is integrated to dynamically adjust the convolutional channel weights based on the posterior category vectors provided by the EM-R module, thereby enhancing the model's responsiveness to key regions and anomalous actions.

The global feature vector output from the backbone is processed by average pooling and then passed to the behavior classifier, which is implemented as a linear discriminative layer that outputs the behavior category label. The entire system forms a unified pipeline from video input to action label output, requiring no external

handcrafted features or post-processing, thus demonstrating strong deploy ability.

In the training stage, end-to-end supervised learning strategy is adopted. Define the multi-task loss function $\mathcal{L}_{total}$, which includes two parts: main classification loss and EM clustering auxiliary loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{em} \tag{17}$$

In Equation (17), $\mathcal{L}_{cls}$ is the cross-entropy loss, which measures the difference between the final classification output and the real label. $\mathcal{L}_{em}$ is the expected maximum reconstruction error of EM module, which is used to constrain the clustering quality. λ is the weight coefficient.

4. Analysis on the Results of Ice and Snow Behavior Recognition Technology

To validate the effectiveness of the proposed method for ice and snow sports action recognition, experiments are conducted using relevant subsets of the publicly available University of Central Florida (UCF101) action recognition dataset (https://www.crcv.ucf.edu/data/UCF101.php). Specifically, three representative action categories "Skiing", "Ski Jumping", and "Snowboarding"-are selected. These categories cover key motion patterns such as gliding, takeoff, and aerial rotation, forming a representative subset for evaluation. Each video segment is labeled with a complete action category, meeting the fundamental requirements of classification tasks.

Since the original UCF101 dataset provides video-level labels, this study adopts an equal-interval frame sampling strategy to convert the videos into frame sequences. Each video is uniformly sampled at 25 Frames Per Second (FPS) and segmented into fixed-length clips of 5 seconds. To enhance the model's ability to capture structural information, the OpenPose key point detection model is used to automatically generate a pose heatmap for each frame. These heatmaps are combined with residual images to form multimodal input features. The final dataset consists of 1,200 video clips with balanced class distribution, and is split into training, validation, and test sets in a ratio of 8:1:1.

The experiments are conducted on a system equipped with an Nvidia RTX 3090 Graphics Processing Unit (GPU), an Intel i9-13900K Central Processing Unit (CPU), and 32GB of Random Access Memory (RAM). The training environment was built on PyTorch 2.0 with Compute Unified Device Architecture (CUDA) 11.8. Performance evaluation metrics include classification accuracy, Macro-averaged F1-score (Macro-F1), number of model parameters (Params), and inference time (ms/frame). These metrics are used to comprehensively assess the model's recognition accuracy, discriminative capability, and real-time performance. All comparative experiments are conducted using the same data splits and hardware configuration to ensure the consistency and fairness of the evaluation results.

Among these metrics, classification accuracy is used to measure the proportion of correctly predicted samples, which is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) denote the numbers of TP, TN, FP, and FN samples, respectively.

The definitions of *Precision* and *Recall* are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

F1 is defined as follows:

$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{21}$$

Since all action categories in the experimental dataset have equal importance, this study adopts the macro-average *F1-score* to average the F1-scores of all categories to avoid the impact of category distribution on the evaluation results.

4.1. Baseline Method and Comparative Experiment

To evaluate the performance advantages of the proposed EM-RepVGG method in ice and snow sports action recognition, four mainstream video action recognition models are selected as comparative baselines. These include: the traditional 2D convolutional network ResNet50 [8], the temporal modeling network TSN [6], the 3D convolutional network I3D [30], and the original RepVGG-A0 model without structural enhancements or clustering guidance [16].

All models are trained using the same input data, number of training epochs, and loss function settings. The inputs consist of frame-extracted image sequences. Except for I3D, which processes 3D video inputs, the remaining models use 2D-RGB images as input. For the proposed method, the input features are constructed from a multimodal combination of residual maps (Δ_t) and pose heatmaps (H_t), and the model integrates the EM-R module along with the multi-scale RepVGG architecture. The experiments evaluate each model on the test set in terms of action classification accuracy, Macro-F1, per-frame inference time, number of model parameters, and model size. A comprehensive comparison of the results is presented in Figure 2.

In Figure 2, the proposed method outperforms all baseline models in both accuracy and F1-score. Compared to the original RepVGG, it achieves a 3.4% improvement in accuracy and a 3.3% gain in F1-score, indicating that the structural re-parameterization and multi-scale enhancement significantly improve the

model's ability to distinguish behavior features. Benefiting from the lightweight design of the RepVGG architecture, the inference speed is maintained at 9.1ms per frame, which is notably faster than that of TSN and I3D, both of which involve more complex temporal modeling structures.

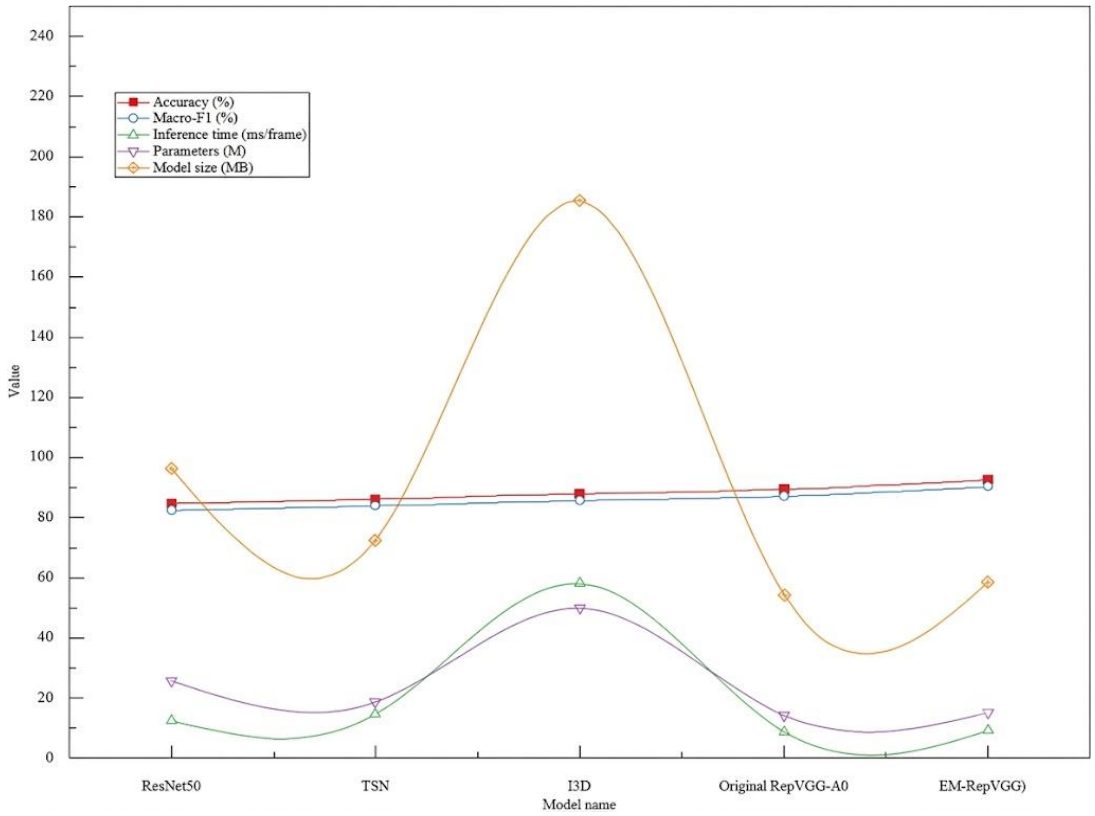

Figure 2. Performance comparison of different models in snow and ice movement behavior recognition task.

4.2. Ablation Experiment

To further verify the individual contributions of each component to the overall performance improvement, systematic ablation experiments are conducted. Key structures or feature inputs are removed one at a time, and the resulting performance is compared against that of the full model. All experiments are conducted under identical training epochs and data splits to ensure consistency and comparability of the results.

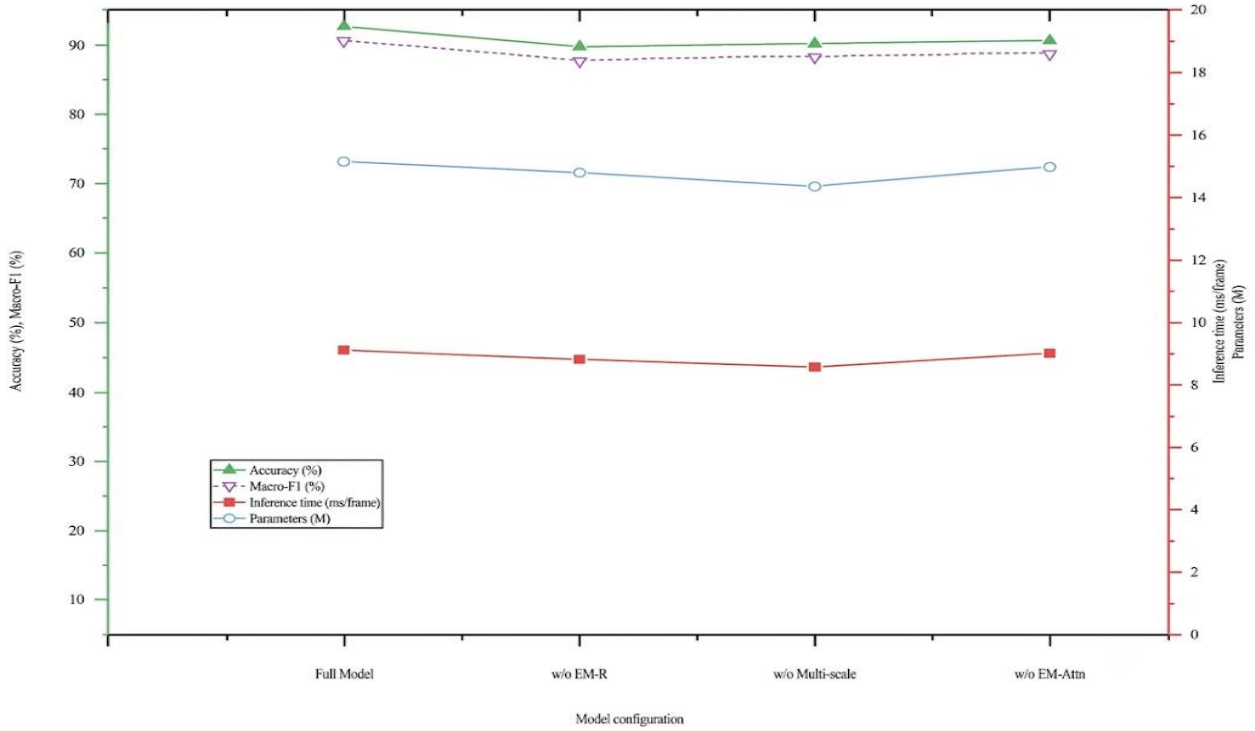

Figure 3. Ablation comparison of each sub-module of model structure.

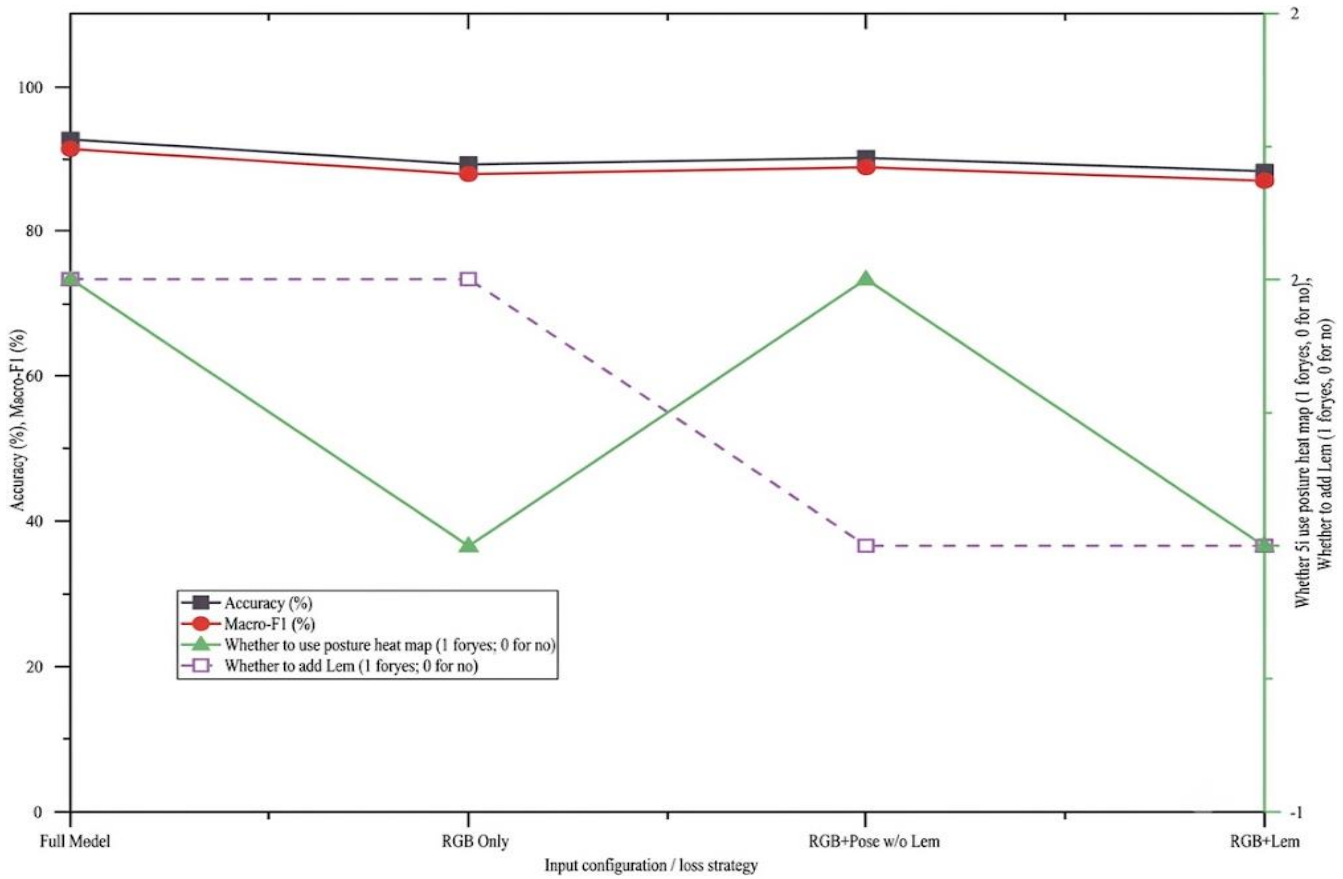

Figure 4. Effects of different input characteristics and loss strategies on model performance.

First, the effectiveness of each sub-module within the model architecture is examined. Figure 3 presents the performance variations when the EM-R module, the multi-scale convolutional units, and the EM-attention mechanism are removed individually.

In Figure 3, each structural module contributes significantly to the overall performance. Among them, the removal of the EM-R module results in the most pronounced performance drop, indicating that semantic clustering-guided feature refinement plays a crucial role in this task. The multi-scale convolutional structure and the attention mechanism also led to performance gains of approximately 1.9% to 2.3%, respectively, confirming the effectiveness of spatial perception and semantic weighting in complex behavior classification.

Next, further analysis is conducted from two perspectives: input modality and loss function design. Figure 4 illustrates the model's performance under different input modalities (RGB vs. RGB+Pose) and the inclusion or exclusion of the EM auxiliary loss term (L_{EM}).

In Figure 4, incorporating pose heatmaps significantly enhances the model's structural perception ability, resulting in approximately a 3.5% increase in accuracy. Additionally, introducing the EM auxiliary loss further stabilizes the clustering-guided process, leading to clearer classification boundaries in ambiguous action segments. The final configuration achieves the best performance, confirming the synergistic effect of enhanced input features and multi-task loss design.

4.3. Real-Time and Deployable Analysis of the System

In practical applications of ice and snow sports analysis, behavior recognition systems are required not only to achieve high classification accuracy but also to meet comprehensive demands for real-time performance, resource efficiency, and deployment flexibility. Therefore, this section evaluates the proposed method's performance from the perspectives of operational efficiency and deployment friendliness, comparing it with mainstream models. All tests are conducted under the same experimental environment, with inference evaluations performed on an Nvidia RTX 3090 GPU and simulated inference on the Jetson NX edge device. Figure 5 presents a comparison of key system metrics under challenging scenarios.

In Figure 5, the proposed method maintains an inference speed close to the original RepVGG (110 FPS) without significantly increasing memory usage, outperforming traditional models such as ResNet50 and TSN. Additionally, it achieves an inference speed of 45 FPS on the Jetson NX platform, making it suitable for real-time deployment on edge devices. Compared to the I3D model, which has a large parameter size and high resource consumption, the proposed method demonstrates stronger device adaptability and better

energy efficiency. Moreover, the model's design supports multi-branch structures during training and structural re-parameterization compression during inference, enabling deployment without modifying the inference graph, thereby significantly reducing model conversion and platform adaptation complexity. In terms of training convergence speed, the proposed method stabilizes within 28 epochs, faster than all compared models, saving both training time and computational resources.

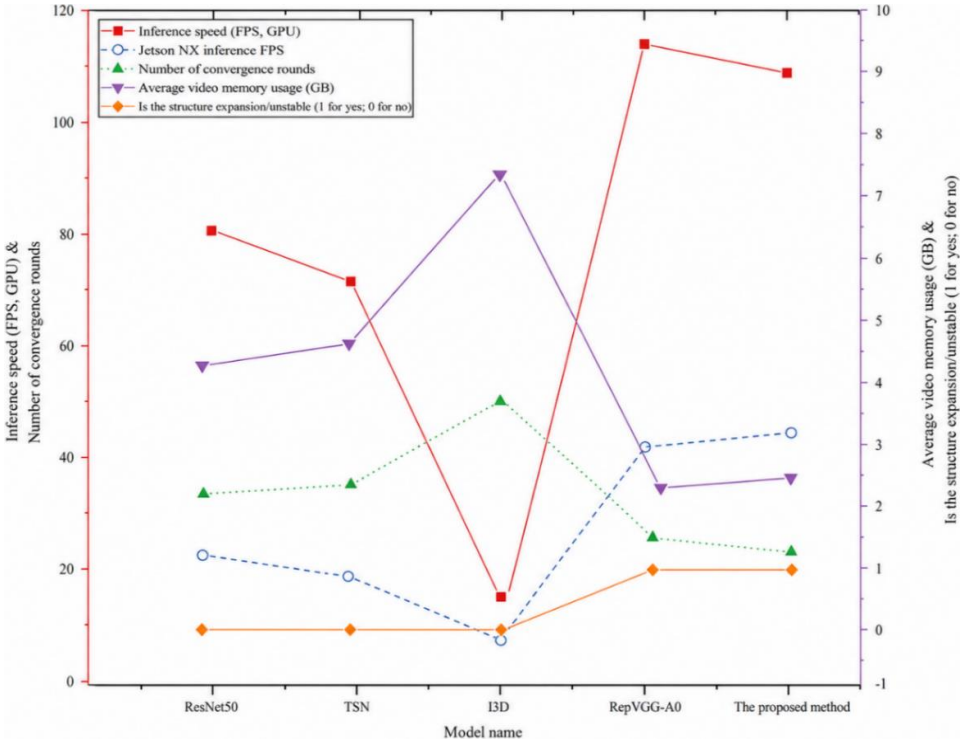

Figure 5. Comparison of operational efficiency and deployment feasibility of each model.

5. Discussions

First, the fast movements and high background interference in ice and snow sports make it difficult for models relying solely on RGB input to effectively separate the subject from the background. The proposed multimodal input strategy, which combines residual maps and pose heatmaps, significantly enhances structural sensitivity and regional discrimination. Pose estimation provides stable key point localization across frames, compensating for misclassification risks caused by color feature failures in complex actions such as jumps and turns.

Second, the EM-R module performs dynamic soft clustering in the feature space by modeling the probability distribution of latent categories. This mechanism enables the model to acquire prior knowledge of behavior semantic boundaries during training, resulting in stronger decision stability in classification tasks. Especially in ambiguous or transitional actions, the posterior guidance from EM-R effectively suppresses category drift, explaining its critical role in performance improvements observed in ablation studies. The improved RepVGG structure leverages multi-scale convolutional modules and the EM-attention mechanism to capture motion patterns at different levels.

Compared with traditional action recognition methods based on 2D or 3D convolutional networks [26, 31], the proposed method does not simply rely on spatiotemporal convolutional structures to model motion information. Instead, it focuses on the moving subjects themselves through explicit background suppression and posture structure modeling. This design is particularly important in winter sports, because high-speed movements and highly similar snow backgrounds tend to weaken the ability of convolutional features to perceive action differences.

Hu *et al*. [11] and Muhammad *et al*. [21] have attempted to model human actions using skeleton sequences or RNNs, achieving certain progress in structural modeling. However, such methods are usually highly sensitive to the accuracy of key point extraction and lack effective utilization of visual contextual information. By taking posture heatmaps as auxiliary channels instead of the sole input, this study retains the advantages of structural information while avoiding the excessive impact of skeleton extraction errors on the overall recognition performance.

In terms of efficient network structures, Balakrishnan and Sengar [3] and Shi *et al*. [25] have applied RepVGG to image classification and other visual tasks, verifying the advantages of structural re-parameterization in inference efficiency. However, these studies mainly focus on feature expression or attention enhancement, and do not introduce a modeling mechanism at the action semantic level. This study further embeds the EM

clustering guidance mechanism into the RepVGG backbone, and realizes the explicit regulation of semantic priors on feature learning through the EM-R module and EM-attention. Thus, it improves the discriminative ability of complex actions while maintaining efficient inference. The combination of "semantic clustering+structural re-parameterization" has not been systematically explored in existing research on action recognition in winter sports.

Without significantly increasing inference cost, it enhances the model's responsiveness to local pose variations and overall dynamics. This lightweight yet expressive architecture is well suited to the "fast speed+complex structure" characteristics of ice and snow sports. Finally, from a system perspective, the proposed method balances accuracy, efficiency, and deploy ability. It maintains stable frame rates on edge devices while delivering high classification performance, demonstrating strong practical potential. Notably, the structural re-parameterization mechanism separates optimization between training and inference phases, reducing structural complexity during deployment and improving engineering feasibility.

6. Conclusions

This study addresses the challenges of insufficient accuracy and real-time performance in complex behavior recognition for ice and snow sports by proposing a multimodal recognition method that integrates an EM clustering mechanism with an improved RepVGG architecture. The proposed method outperforms traditional 2D and 3D models as well as the original RepVGG backbone in terms of accuracy and F1-score. It achieves high precision while maintaining fast inference speed and a compact model size. Its deployment on Jetson edge devices further validates its practical engineering applicability. Furthermore, ablation studies demonstrate the independent effectiveness of each structural module and the synergistic gains from their combination. Despite the clear performance advantages, certain limitations remain. The model's generalization capability to abnormal behavior categories needs further improvement, and some cross-individual behavioral variations cause mismatches in pose estimation. Additionally, the dataset relies solely on a single public dataset, lacking validation across diverse real-world scenarios. Future work will focus on expanding multisource sensor fusion, incorporating temporal Transformer architectures to enhance long-term behavior modeling, and leveraging large-model pretraining techniques to improve recognition performance in zero-shot or few-shot scenarios.

References

[1] Abdelbaky A. and Aly S., "Human Action Recognition Using Three Orthogonal Planes with Unsupervised Deep Convolutional Neural Network," *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 20019-20043, 2021. https://link.springer.com/article/10.1007/s11042-021-10636-2

[2] Alhakbani N., Alghamdi M., and Al-Nafjan A., "Design and Development of An Imitation Detection System for Human Action Recognition Using Deep Learning," *Sensors*, vol. 23, no. 24, pp. 1-16, 2023. DOI: 10.3390/s23249889

[3] Balakrishnan T. and Sengar S., "RepVGG-GELAN: Enhanced GELAN with VGG-Style Convnets for Brain Tumour Detection," *in Proceedings of the 9th International Conference* on *Computer Vision and Image Processing*, Chennai, pp. 417-430, 2024. https://link.springer.com/chapter/10.1007/978-3-031-93688-3_30

[4] Chen D., Wu P., Chen M., Wu M., and et al., "LS-ViT: Vision Transformer for Action Recognition Based on Long and Short-Term Temporal Difference," *Frontiers in Neurorobotics*, vol. 18, pp. 1-12, 2024. https://doi.org/10.3389/fnbot.2024.1457843

[5] Cheng J., Ren Z., Zhang Q., Gao X., and Hao F., "Cross-Modality Compensation Convolutional Neural Networks for RGB-D Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1498-1509, 2022. https://ieeexplore.ieee.org/document/9417202

[6] Ding G., Sener F., and Yao A., "Temporal Action Segmentation: An Analysis of Modern Techniques," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 1011-1030, 2023. DOI: 10.1109/TPAMI.2023.3327284

[7] Ding I., Zheng N., and Hsieh M., "Hand Gesture Intention-Based Identity Recognition Using Various Recognition Strategies Incorporated with VGG Convolution Neural Network-Extracted Deep Learning Features," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 7775-7788, 2021. https://doi.org/10.3233/JIFS-189598

[8] Elpeltagy M. and Sallam H., "Automatic Prediction of COVID-19 from Chest Images Using Modified ResNet50," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26451-26463, 2021. DOI: 10.1007/s11042-021-10783-6

[9] Guo Y., Ju R., Li K., Lan Z., and et al., "A Smart Ski Pole for Skiing Pattern Recognition and Quantification Application," *Sensors*, vol. 24, no. 16, pp. 1-14, 2024. DOI: 10.3390/s24165291

[10] Han J. and Li J., "A Video Action Recognition Method via Dual-Stream Feature Fusion Neural Network with Attention," *International Journal of Uncertainty, Fuzziness and Knowledge-Based*

Systems, vol. 32, no. 4, pp. 673-694, 2024. https://doi.org/10.1142/S0218488524400130

[11] Hu K., Jin J., Zheng F., Weng L., and Ding Y., "Overview of Behavior Recognition Based on Deep Learning," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 1833-1865, 2023. https://doi.org/10.1007/s10462-022-10210-8

[12] Kareem I., Ali S., Bilal M., and Hanif M., "Exploiting the Features of Deep Residual Network with SVM Classifier for Human Posture Recognition," *PLoS ONE*, vol. 19, no. 12, pp. 1-26, 2024. DOI: 10.1371/journal.pone.0314959

[13] Korban M., Youngs P., and Acton S., "A Semantic and Motion-Aware Spatiotemporal Transformer Network for Action Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6055-6069, 2024. https://ieeexplore.ieee.org/document/10472872

[14] Li H. and Reddy M., "Remote Dance Action Correction Based on EM and Min-Min Algorithms," *Mobile Networks and Applications*, vol. 30, pp. 1037-1049, 2025. DOI:10.1007/s11036-024-02427-4

[15] Li T., Wang J., Wiltos K., and Wozniak M., "An Intelligent Proofreading for Remote Skiing Actions Based on Variable Shape Basis," *Mobile Networks and Applications*, vol, 30, pp. 1008-1021, 2026. DOI: 10.1007/s11036-024-02419-4

[16] Lin C., Zheng Y., Xiao X., and Lin J., "CXR-RefineDet: Single-Shot Refinement Neural Network for Chest X-Ray Radiograph Based on Multiple Lesions Detection," *Journal of Healthcare Engineering*, vol. 2022, no. 1, pp. 1-11, 2022. DOI: 10.1155/2022/4182191

[17] Liu J., Huang G., Hyyppä J., Li J., and et al., "A Survey on Location and Motion Tracking Technologies, Methodologies and Applications in Precision Sports," *Expert Systems with Applications*, vol. 229, pp. 120492, 2023. https://doi.org/10.1016/j.eswa.2023.120492

[18] Liu Y., Zhang H., Li Y., He K., and Xu D., "Skeleton-Based Human Action Recognition via Large-Kernel Attention Graph Convolutional Network," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2575-2585, 2023. DOI: 10.1109/TVCG.2023.3247075

[19] Lopez-Lozada E., Sossa H., Rubio-Espino E., and Montiel-Perez J., "Action Recognition in Videos Through a Transfer-Learning-Based Technique," *Mathematics*, vol. 12, no. 20, pp. 1-17, 2024. https://doi.org/10.3390/math12203245

[20] Ma L., Yang H., and Jin G., "A Spatio-Temporal Feature Representation of Multimodal Surveillance Images for Behavioral Recognition," *The International Arab Journal of Information Technology*, vol. 22, no. 4, pp. 832-843, 2025. https://doi.org/10.34028/iajit/22/4/15

[21] Muhammad K., Mustaqeem., Ullah A., Imran A., and et al., "Human Action Recognition Using Attention Based LSTM Network with Dilated CNN Features," *Future Generation Computer Systems*, vol. 125, pp. 820-830, 2021. https://doi.org/10.1016/j.future.2021.06.045

[22] Muttenthaler L., Dippel J., Linhardt L., Vandermeulen R., and Kornblith S., "Human Alignment of Neural Network Representations," *in Proceedings of the Machine Learning Group, Technische Universitat*, Berlin, pp. 1-38, 2023. https://iclr.cc/virtual/2023/poster/11553

[23] Nasaoui H., Bellamine I., and Silkan H., "Action Recognition Using Segmental Action Network," *Revue d'Intelligence Artificielle*, vol. 37, no. 1, pp. 185-190, 2023. DOI:10.18280/ria.370123

[24] Patel S. and Kamdar D., "Survey on Sport Video Analysis and Event Detection," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 18, no. 1, pp. 67-98, 2025. DOI: 10.1504/IJAACS.2025.10059628

[25] Shi C., Han L., Zhang K., Xiang H., and et al., "Improved RepVGG Ground-Based Cloud Image Classification with Attention Convolution," *Atmospheric Measurement Techniques*, vol. 17, no. 3, pp. 979-997, 2024. https://amt.copernicus.org/articles/17/979/2024/

[26] Sowmya M., Balasubramanian M., and Vaidehi K., "Human Behavior Classification Using 2D Convolutional Neural Network, VGG16 and ResNet50," *Indian Journal of Science and Technology*, vol. 16, no. 16, pp. 1221-1229, 2023. DOI:10.17485/IJST/v16i16.199

[27] Wang M., Xing J., Su J., Chen J., and Liu Y., "Learning Spatiotemporal and Motion Features in a Unified 2D Network for Action Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3347-3362, 2023. DOI: 10.1109/TPAMI.2022.3173658

[28] Wang Z. and Jing R., "Optimized Design of Flexible Sensors for Snow and Ice Sports Characterization and Intelligent Recognition of Training Technique Movements," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 3, pp. 1-9, 2024. https://doi.org/10.1016/j.jrras.2024.101028

[29] Wensel J., Ullah H., and Munir A., "ViT-RET: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos," *IEEE Access*, vol. 11, pp. 72227-72249, 2023. DOI: 10.1109/ACCESS.2023.3293813

[30] Wu Q., Zhu A., Cui R., Wang T., and et al., "Pose-Guided Inflated 3D ConvNet for Action Recognition in Videos," *Signal Processing: Image Communication*, vol. 91, pp. 116098, 2021. https://www.sciencedirect.com/science/article/abs/pii/S0923596520302186?via%3Dihub

[31] Yang J., Jia Q., Han S., Du Z., and Liu J., "An Efficient Multi-Scale Attention Two-Stream

Inflated 3D Convnet Network for Cattle Behavior Recognition," *Computers and Electronics in Agriculture*, vol. 232, pp. 110101, 2025. DOI: 10.1016/j.compag.2025.110101

[32] Ye X., Sakurai K., Nair N., and Wang K., "Machine Learning Techniques for Sensor-Based Human Activity Recognition with Data Heterogeneity-A Review," *Sensors*, vol. 24, no. 24, pp. 1-39, 2024. DOI:10.3390/s24247975

[33] Zhao L., Lin Z., Sun R., and Wang A., "A Review of State-of-the-Art Methodologies and Applications in Action Recognition," *Electronics*, vol. 13, no. 23, pp. 1-39, 2024. DOI: 10.3390/electronics13234733

[34] Zhao X., "Research on Athlete Behavior Recognition Technology in Sports Teaching Video Based on Deep Neural Network," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1-13, 2022. https://doi.org/10.1155/2022/7260894

Shu Yu is an Associate Professor, Doctoral Student, Master Supervisor of Harbin Sport University. She is mainly engaged in research related to Sports Economy and Management, Sports Tourism, Sports Industry, Alpine Skiing, etc. She has won 2 first prizes of Heilongjiang Provincial Teaching Achievement Award, 1 second prize of Heilongjiang Provincial Teaching Ability Competition, 1 Heilongjiang Provincial High-quality Construction Project, and 4 utility model patents. She has undertaken and completed 7 provincial-level projects and 11 university-level projects. She guided students to win one third prize in national competitions, one second prize and one third prize in provincial competitions. She has published more than 20 papers in domestic and foreign journals, some of which are core, SCI, SSCI and EI. She has compiled 3 teaching materials and works.

Jianxin Kang is graduated from Wooook University with a Ph.D. in Physical Education and currently serves as an associate professor at the Winter Olympic College of Harbin Sport University. His research areas include Winter Sports, the Winter Olympics, and Sports Training.

Dawei Wu is a Ph.D. Professor and a Master's Supervisor. He serves as Vice Dean of the Winter Olympic College, Harbin Sport University, International Judge of Luge, and Member of the Executive Committee of the Federation of International Bandy (FIB). He has successively held the posts of Head Chinese Coach of the National Luge Team, Deputy Sport Manager for Luge, Sports Department of Beijing Organizing Committee for the 2022 Olympic and Paralympic Winter Games, race director for Luge Events at Beijing Winter Olympics, Technical Delegate for Luge at the 2024 Gangwon Winter Youth Olympic Games, Technical Delegate of the International Luge Federation (FIL) World Cup, Chairman of the Jury for the 2025-2026 FIL Luge World Cup Season, and Jury Member for Luge Events at the 2026 Milan Winter Olympics. He has long engaged in research on training theory and practice of winter sports. He has presided over and participated in a number of national and provincial/ministerial research projects with abundant academic achievements.

Wang Wang received the M.S. degree in Physical Education with a major in Social Sports instruction from Harbin Sport University in 2026. Her core coursework covers Social Sports Science, Sports Management, Sports Culture Studies, the Sports Industry, Sports Event Planning and Organization, as well as Sports Tournament Operation and Management. From 2023 to 2026, she served as a research Supervisor Assistant at the School of Winter Olympic Sports, Harbin Sport University, supporting the implementation of multiple scientific research projects. Her primary research interests include Women's Sports, Social Sports, Ice and Snow Sports, Sports Tourism, and sports Etiquette. During this period, she has participated in publishing numerous academic papers and research outputs. Her representative work is as follows: Wang W. and Yu S., "Dilemmas and Paths for the Integrated Development of Wushu Intangible Cultural Heritage and the Tourism Industry in Liaoning Province," in Proc. 4th Nat. Conf. Wushu Education and Health, Jimei Univ., 2025, presented in thematic sessions. She is currently engaged in interdisciplinary research spanning sports sociology and the sports industry, with a focus on the Social Development of Women's Sports and Industrial Integration in Sports Tourism.