IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

**Traitement du Signal**
Vol. 43, No. 2, April, 2026, pp. 971-980
Journal homepage: http://iieta.org/journals/ts

# SportNet: A Lightweight Deep Neural Network Framework for Quantitative Evaluation of Sports Training Movements

Yicen Zhong, Zhuo Liu*

School of Physical Education and Health, Guilin Institute of Information Technology, Guilin 541000, China

Corresponding Author Email: liu28zhuo@163.com



https://doi.org/10.18280/ts.430233

**Received:** 29 September 2025
**Revised:** 26 February 2026
**Accepted:** 10 March 2026
**Available online:** 30 April 2026

***Keywords:***
*sports action recognition, deep neural networks, quantitative action evaluation, ablation study, Penn Action dataset, temporal modeling*

**ABSTRACT**

In sports training and physical education, the evaluation of movement standardization and accuracy has long relied on manual observation, which is inherently subjective, labor-intensive, and difficult to scale. To address these challenges, this paper proposes SportNet, a lightweight deep neural network–based framework for quantitative evaluation of sports training movements. The framework innovatively integrates temporal features from RGB video sequences with human pose keypoint information, employing ResNet18 as the backbone feature extractor to achieve efficient and accurate automated action recognition and assessment. Extensive experiments were conducted on the Penn Action dataset, which contains 15 categories of sports movements. First, a performance benchmark was established, with the baseline model achieving a test accuracy of 93.94%. Subsequently, systematic ablation studies were performed to investigate the effects of temporal frame length, model capacity, and batch size on performance. Experimental results demonstrate that appropriate adjustment of batch size significantly enhances model generalization, yielding an improvement of 3.44 percentage points over the baseline. Moreover, a moderate temporal input length (16 frames) achieves competitive performance while offering substantially higher computational efficiency compared to longer sequences. Ultimately, the optimized SportNet model attains a recognition accuracy of 97.38%. The proposed method achieves a favorable balance between accuracy and real-time performance, providing an effective technical pathway for the development of intelligent sports training assistance systems.

## 1. INTRODUCTION

Human action understanding based on video has become a key supporting technology in intelligent sports training, sports rehabilitation, and physical fitness assessment. It is not only related to the recognition at the level of "action category," but also closely associated with movement standardization, skill proficiency, and potential injury risk warning during the training process. Compared with motion capture systems under laboratory conditions, which provide high precision but are costly and limited in deployment, monocular video solutions have inherent scalability and application universality. Therefore, they are widely regarded as a practical pathway to extend motion analysis capability from professional venues to daily training scenarios. However, compared with general daily actions, sports actions exhibit more prominent fine-grained characteristics and strong temporal structure. Many action categories are highly similar in static appearance, and their differences are often concentrated in joint coordination at key phases, limb velocity changes, or phase transition order. At the same time, the same action presents significant intra-class variation across different individuals, different skill levels, and different repetitions. Such variation is often highly correlated with action quality itself, leading to the situation where "being able to distinguish categories" is not equivalent to "being able to judge how well the action is performed" [1]. Therefore, vision systems for sports training need to handle three types of contradictions simultaneously: (i) the spatiotemporal sensitivity required for fine-grained discrimination vs. the low computational power required for real-time deployment; (ii) the data scale required for robust generalization vs. the realistic conditions of sports data annotation (medium-to-small scale, noise, and bias); (iii) the necessity of human structural priors vs. the contextual absence of skeleton-only representation. Taking sports datasets such as Penn Action as an example, its action categories (e.g., tennis, baseball, golf, etc.) exhibit clear equipment interaction and phase structure, and it provides both keypoints and action labels, making it a typical benchmark for testing collaborative modeling of "structural information + visual context" [2].

### 1.1 From handcrafted features to deep spatiotemporal representation: The cost of accuracy improvement

Early action recognition mainly relied on manually constructed spatiotemporal features. Methods represented by spatiotemporal interest points (STIP) attempted to capture motion cues through local spatiotemporal saliency [3]. Subsequently, dense trajectories achieved long-term competitiveness on multiple benchmarks through large-scale

sampling and trajectory descriptors (HOG/HOF/MBH) [4], and further improved robustness in real scenes through explicit camera motion compensation [4]. The core advantage of this line of work lies in the explicit modeling of motion cues, but its limitations are also clear: dependence on complex preprocessing (optical flow/tracking), long training and inference pipelines, difficulty in end-to-end optimization, and inability to form a unified representation for occlusion, background interference, and equipment interaction commonly seen in sports scenarios.

Deep learning has promoted the paradigm shift of action recognition from "feature engineering" to "representation learning." Two-stream networks decouple spatiotemporal information modeling through appearance stream (RGB) and motion stream (optical flow), significantly improving recognition accuracy and becoming a classic framework of one generation [5]. However, the additional computational and storage cost brought by optical flow pre-computation makes it unsuitable in training scenarios that require low-latency feedback. Subsequently, 3D convolutional networks directly learn features on spatiotemporal volumes: C3D demonstrated the effectiveness of 3D convolution for action representation [6]; I3D expanded mature 2D convolutional backbones into 3D and combined them with large-scale video pre-training, becoming one of the high-performance baselines [7]; R(2+1)D further decomposed 3D convolution into spatial and temporal steps to improve optimization characteristics [8]; SlowFast strengthened multi-time-scale modeling capability through dual pathways that capture slow semantics and fast motion details [9]. The common problem of these methods is that their performance highly depends on large-scale pre-training (e.g., Kinetics) [10], and the computational and memory costs are significant, which restricts their deployment on edge devices or in real-time training feedback systems. Under the common conditions of sports data—"medium scale + strong intra-class variation + background bias"—direct adoption of heavy spatiotemporal models often faces the dual pressure of training cost and generalization risk.

### 1.2 Efficient video networks and sparse sampling: "Necessary and Sufficient" temporal information in sports scenarios

Under efficiency constraints, researchers have proposed a route centered on 2D backbones, injecting dynamic information through lightweight temporal mechanisms. Temporal Segment Networks (TSN) emphasize sparse segment sampling and video-level aggregation, significantly reducing computational cost while covering long-range structure, and forming a reusable training paradigm [11]. Temporal Shift Module (TSM) realizes temporal information exchange along the channel dimension with nearly zero additional FLOPs, providing a strong trade-off between efficiency and accuracy [12]. X3D systematically explores the accuracy–complexity Pareto frontier of video models by progressively expanding along the axes of temporal length, spatial resolution, and network width/depth [13]. An important implicit insight of these works is that, for many actions, key information is concentrated in limited temporal segments or key phases, and overly dense temporal sampling may introduce significant redundancy. Sports training movements often exhibit periodicity, stage characteristics, and phase-dominant features, which makes "reasonable temporal sampling + lightweight aggregation" more attractive in engineering practice. It can cover key dynamics while reserving computational resources for learning more discriminative spatial/structural features. The experimental design in this paper (systematic ablation on input frame number, model capacity, and batch size) is carried out around this engineering proposition of "necessary and sufficient temporal information."

### 1.3 Transformer-based video models: Rebalancing representation capacity and data dependence

Video Transformers based on the self-attention mechanism further improve long-range dependency modeling capability. TimeSformer provides a relatively clear attention modeling framework through spatial–temporal attention factorization [14]. ViViT alleviates the computational bottleneck caused by the large number of video tokens through multiple factorization strategies and demonstrates the feasibility of transferring from image pre-training to video [15]. However, in the deployment context of sports training, Transformer-based models still face typical challenges: they are more sensitive to data scale, regularization, and pre-training strategies, and under edge-side deployment and low-latency feedback objectives, they often require further structural compression and engineering redesign. Therefore, for system design targeting "fast training, fast deployment, and stable generalization," lightweight Convolutional Neural Network (CNN) backbones combined with structural priors still have practical advantages.

### 1.4 Pose keypoints and skeleton graph modeling: The value and gap of structural priors

The discrimination of sports movements is often strongly correlated with joint coordination and limb kinematics, which provides clear motivation for introducing human pose structure. Methods such as OpenPose achieve relatively usable 2D pose extraction [16], and HRNet improves keypoint localization quality by maintaining high-resolution representations [17]. On this basis, skeleton sequences can be represented as spatiotemporal graphs and learned through graph convolution. Spatio-Temporal Graph Convolutional Network (ST-GCN) proposes a classical spatiotemporal graph convolution framework [18], and two-Stream Adaptive Graph Convolutional Network (2s-AGCN) further enhances representation capability through adaptive graph structure and dual-stream modeling [19]. The advantage of skeleton-based methods lies in their relative robustness to background, clothing, and illumination changes, and their ability to directly focus on the kinematic essence. However, their limitations are also prominent in sports scenarios: keypoint errors may accumulate and amplify over time; more importantly, skeleton-only representation often lacks critical context of human–equipment/object interaction (such as the relative position of racket, dumbbell, or ball), while such context often determines action semantics and quality discrimination [2]. Therefore, the complementarity between RGB (context) and pose (structure) suggests a more feasible direction: under controllable computational cost, incorporating structural priors as auxiliary information to collaboratively learn with visual representations.

### 1.5 From "recognition" to "quality evaluation": The hierarchical complexity of sports tasks

Sports training ultimately concerns not only action categories, but also action quality, skill level, and process structure. Automated Quality Assessment research points out that classification alone is insufficient to characterize "how well the action is performed," and promotes systematic research on action quality regression and process understanding [1]. Subsequent work further explores multi-task learning paradigms, jointly modeling tasks such as recognition, language description, and scoring to enhance quality representation [20]. Meanwhile, fine-grained sports datasets such as FineGym emphasize hierarchical structure and phase annotation of actions, highlighting the essential challenges in sports scenarios: small inter-class differences, large intra-class differences, and strong phase structure [21]. These studies jointly indicate that reliable deployment of sports action understanding requires careful trade-offs among representation capacity, structural priors, training stability, and computational resources. This paper focuses on a practical aspect of such trade-offs: on medium-scale sports datasets, constructing a lightweight yet expressive action recognition/evaluation framework that captures key spatiotemporal and structural cues, thereby laying a stable foundation for further quality evaluation and feedback generation.

### 1.6 Positioning and contributions of this work

Based on the above analysis, this paper proposes a lightweight quantitative evaluation framework for sports training movements, namely SportNet. Different from heavy 3D spatiotemporal networks, SportNet adopts a lightweight 2D residual network as the backbone representation (facilitating the use of mature pre-training and efficient inference) [22], and enhances sensitivity to fine-grained motion differences by integrating pose structural information. At the same time, we incorporate key factors such as temporal sampling length, model capacity, and training batch size into systematic ablation studies to obtain reusable engineering principles. It is worth emphasizing that training stability and generalization are often highly related to the statistical estimation of Batch Normalization and batch size [23], and the generalization behavior of large-batch training has long been discussed in deep learning [24]. Therefore, the experimental analysis of batch size in this paper not only serves performance improvement, but also serves empirical summarization of training dynamics.

The main contributions of this paper are summarized as follows:

(1) Proposing a lightweight action evaluation framework for sports training scenarios, emphasizing the balance between accuracy and deployment efficiency;

(2) Conducting systematic ablation on the Penn Action benchmark to quantitatively analyze the influence of temporal sampling, model capacity, and batch size on generalization performance;

(3) Achieving significant performance improvement on this benchmark (the final reported accuracy reaches 97.38%), verifying the effectiveness and engineering applicability of the method [2, 22].

## 2. METHODS

The proposed SportNet aims to address the contradiction between high-dimensional spatiotemporal feature capture and computational efficiency in quantitative evaluation of sports movements. The core logic of SportNet is based on a collaborative learning paradigm of "lightweight spatial representation + pose structural prior + efficient temporal aggregation." The RGB branch captures contextual information such as equipment, environment, and limb appearance through a lightweight convolutional network. The pose branch injects strong kinematic constraints using frame-wise 2D human skeleton sequences. Subsequently, the two representations are fused at the frame level across modalities, and a global aggregation layer is used to extract a video-level discriminative vector, thereby achieving accurate action classification (Figure 1).

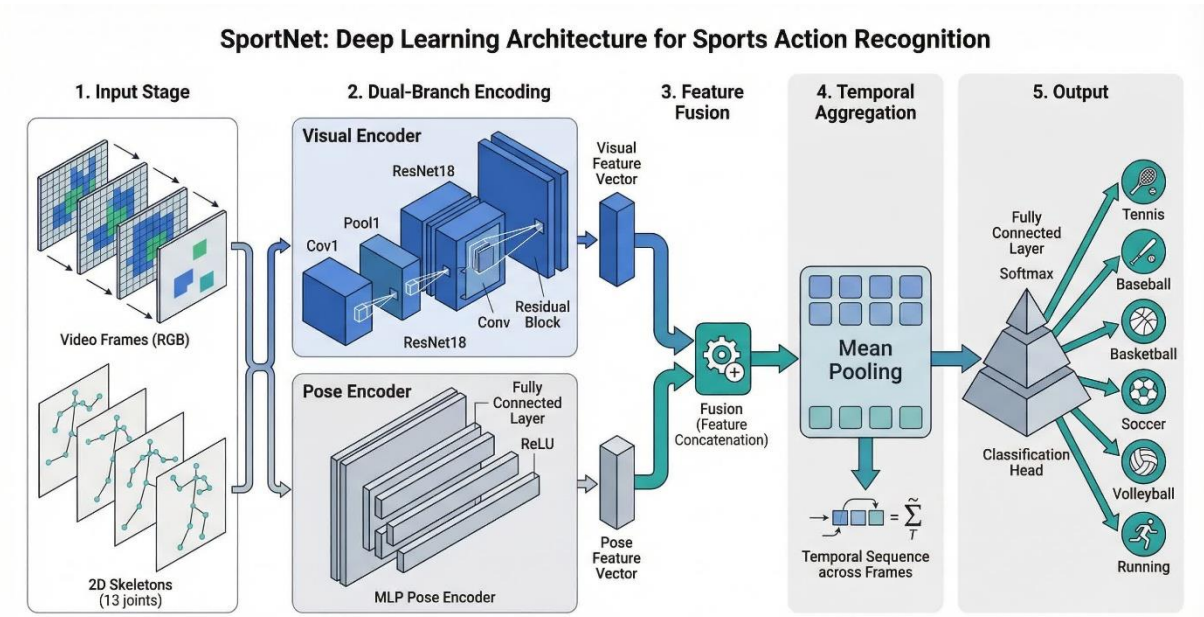


**Figure 1.** Method pipeline

### 2.1 Dataset and task protocol

We validate the proposed framework on the Penn Action benchmark dataset. This dataset is of moderate scale and provides detailed annotations, containing 2,326 sports video clips in natural scenes, covering 15 categories of typical competitive and rehabilitation actions. Penn Action provides frame-wise offsets of 13 human keypoints and visibility indices, along with frame-wise human bounding box annotations. Considering the highly dynamic and fine-grained characteristics of sports movements, this dataset provides an ideal experimental platform for validating fusion models of "visual representation + geometric structure." In this study, we model video-level action recognition as a C-class multi-class classification problem (C = 15), where the classification label is defined as $y \in \{1, \dots, C\}$. To ensure experimental rigor, we strictly follow the official video-level split protocol provided by the dataset, prohibiting frame-level information sharing between the training and testing sets.

### 2.2 Temporal sampling strategy

To address the inconsistency of original video length $L$, this paper introduces a segment-based uniform sampling mechanism to construct a unified spatiotemporal volume with input sequence length $T$. We divide the temporal index space $[1, L]$ of a video into $T$ consecutive and non-overlapping segments. For the $i$-th segment, its temporal boundary $[a_i, b_i]$ is defined as:

$$a_i = \left\lfloor \frac{(i-1)L}{T} \right\rfloor + 1, b_i = \left\lfloor \frac{iL}{T} \right\rfloor$$

During the training stage, one frame index $\tau_i \sim U\{a_i, \dots, b_i\}$ is randomly sampled from the corresponding segment. This random sampling mechanism effectively achieves temporal data augmentation and enhances the model's robustness to action phase drift. During the inference stage, in order to ensure determinism and reproducibility of prediction results, we fix the center frame of each segment as input, i.e., $\tau_i = \lfloor (a_i + b_i)/2 \rfloor$. The constructed input sequence can be represented as a set of modality pairs: $\{(I_{\tau_i}, P_{\tau_i})\}_{i=1}^{T}$.

### 2.3 Spatial normalization and cross-modal alignment

Since there are significant differences in human scale and shooting distance in Penn Action clips, in order to suppress irrelevant background noise and strengthen person-centered features, we adopt a human-centered spatial normalization strategy. For the sampled video frame $I_t$, local cropping is performed using the corresponding bounding box annotation $b_t = (x_t^{(1)}, y_t^{(1)}, x_t^{(2)}, y_t^{(2)})$, and the result is resampled to a unified resolution (224 × 224 pixels). For the normalized frame image, we perform channel-wise mean subtraction and pixel scaling:

$$\hat{I}_t = \frac{I_t - \mu}{\sigma}$$

where, $\mu$ and $\sigma$ are statistics obtained from the large-scale ImageNet dataset. Meanwhile, to maintain spatial topological consistency across modalities, human keypoint coordinates undergo synchronized geometric transformation. During training, data augmentation operations (horizontal flipping, scale scaling) dynamically adjust keypoint coordinate values and semantic indices (logically swapping left and right wrist nodes), ensuring strict geometric alignment between the visual context stream and the pose structural stream at the feature fusion layer.

### 2.4 Kinematic representation

To eliminate the influence of camera intrinsic transformation and initial human position on pose recognition, we perform a local coordinate system transformation on the 2D joint positions $(x_{t,k}, y_{t,k})$. Let the real-time size of the human bounding box be $w_t, h_t$. Then the normalized relative position of the $k$-th joint is defined as:

$$\tilde{x}_{t,k} = \frac{x_{t,k} - x_t^{(1)}}{w_t + \varepsilon}, \tilde{y}_{t,k} = \frac{y_{t,k} - y_t^{(1)}}{h_t + \varepsilon}$$

Considering that discrimination of sports movements often depends on relative joint velocity rather than static position, we introduce first-order spatiotemporal differences as supplementary kinematic features:

$$\Delta\tilde{x}_{t,k} = \tilde{x}_{t,k} - \tilde{x}_{t-1,k}, \Delta\tilde{y}_{t,k} = \tilde{y}_{t,k} - \tilde{y}_{t-1,k}$$

Finally, the complete pose code vector $p_t$ of frame $t$ is formed by concatenating human topological positions, instantaneous motion velocities, and visibility masks:

$$\mathbf{p}_t = [\dots, \tilde{x}_{t,k}, \tilde{y}_{t,k}, \Delta\tilde{x}_{t,k}, \Delta\tilde{y}_{t,k}, m_{t,k}, \dots]^{\top}$$

This compact vectorized encoding effectively compresses the core cues of human kinematics and significantly reduces the processing cost of the subsequent branch.

### 2.5 SportNet branch architectural design

SportNet adopts a parallel dual-branch architecture. The visual stream branch (Visual Encoder) uses a lightweight ResNet18 as the feature extraction engine. By performing hierarchical convolution and global residual mapping on the normalized frame $\hat{I}_t$, a deep spatial representation vector $v_t = f_{\text{rgb}}(\hat{I}_t; \theta_v) \in \mathbb{R}^{512}$ is extracted. The structural stream branch (Pose Encoder) utilizes a perceptron-based lightweight mapping to transform the pose vector $p_t$ into a high-dimensional hidden feature $u_t \in \mathbb{R}^{d_u}$.

To achieve deep integration of multimodal information, the framework performs fusion mapping at each frame-level temporal point:

$$\mathbf{f}_t = \sigma(W_f[\mathbf{v}_t; \mathbf{u}_t] + \mathbf{b}_f)$$

where, $\sigma$ denotes a nonlinear activation layer. Subsequently, the framework overcomes the temporal locality perspective of short sequences through a Global Temporal Aggregation layer to extract a video-level semantic consistency vector $g$ (default using temporal mean pooling):

$$\mathbf{g} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t$$

Finally, the *Classifier* maps the aggregated feature to the action category space. The scientific motivation of this design lies in utilizing CNN texture perception to recognize

equipment (ball, racket, etc.), while leveraging structural priors learned by Multi-Layer Perceptron (MLP) to overcome human misrecognition under complex background conditions.

### 2.6 Model optimization and inference

The parameter optimization of SportNet is established under the empirical risk minimization (ERM) framework, adopting multi-class cross-entropy loss across category labels:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{p}_c) + \lambda \parallel \theta \parallel_2^2$$

where, $\lambda$ is the regularization weight, used to alleviate the risk of overfitting on medium-scale sports datasets. Our experiments observe that the batch size during training has a significant impact on the accuracy of mean and variance estimation in Batch Normalization layers, which is crucial for the generalization ability of fine-grained tasks such as sports actions. After training is completed, during inference the system performs an arg max selection over the averaged probability distribution generated from the $T$ segments, and finally outputs the predicted action category corresponding to the sample.

## 3. RESULTS AND DISCUSSION

### 3.1 Evaluation metrics and statistical protocol

In this paper, the video-level action recognition task on Penn Action is modeled as a 15-class multi-class classification problem. The primary evaluation metric is Top-1 Accuracy (%). In addition to accuracy, in order to measure the engineering applicability of the method, we record the total training time (min) of different configurations under the same training pipeline and hardware environment, to analyze the trade-off relationship between "accuracy–efficiency".

### 3.2 Comparison with prior methods

To verify the external competitiveness of the proposed model, we compare the final model (the best configuration in this experiment is Larger Batch) with commonly used/representative methods reported in the literature on Penn Action. Table 1 summarizes the action recognition accuracy of the comparison methods (all values except Ours are reported from the literature).

**Table 1.** Comparison of action recognition accuracy on Penn Action (Top-1 Accuracy, %)

| Method | Accuracy (%) |
|---|---|
| AOG-Fine [25] | 73.4 |
| STIP-HoG+HoF [3, 25] | 82.8 |
| C3D [7, 25] | 86.0 |
| JDD [26] | 87.4 |
| MMTSN-RGB+Pose [25] | 91.67 |
| IDT-FV [5, 25] | 92.0 |
| IDT-FV+Pose [5, 25] | 92.9 |
| TSN [12, 25] | 93.8 |
| DPI+att-DTI [27] | 93.9 |
| DPI+att-DTIs [27] | 95.8 |
| SportNet (Ours) | 97.38 |

From the results, SportNet achieves an accuracy of 97.38% on Penn Action, overall outperforming the comparison methods listed in Table 1. Compared with the strongest baseline in the table, DPI+att-DTIs (95.8%), it achieves an improvement of +1.58 percentage points. This result indicates that, for sports actions characterized by strong fine-grained differences and strong contextual dependence, effective fusion of RGB visual context and human pose structural priors can improve the clarity of the decision boundary without significantly increasing temporal burden.

### 3.3 Ablation study: Temporal length, model capacity, and optimization strategy

This section focuses on three types of key factors: temporal sampling length $T$, model capacity (Hidden Dim), and optimization hyperparameter (Batch Size). The overall accuracy comparison is shown in Figure 2 and the changes relative to Baseline are shown in Figure 3.

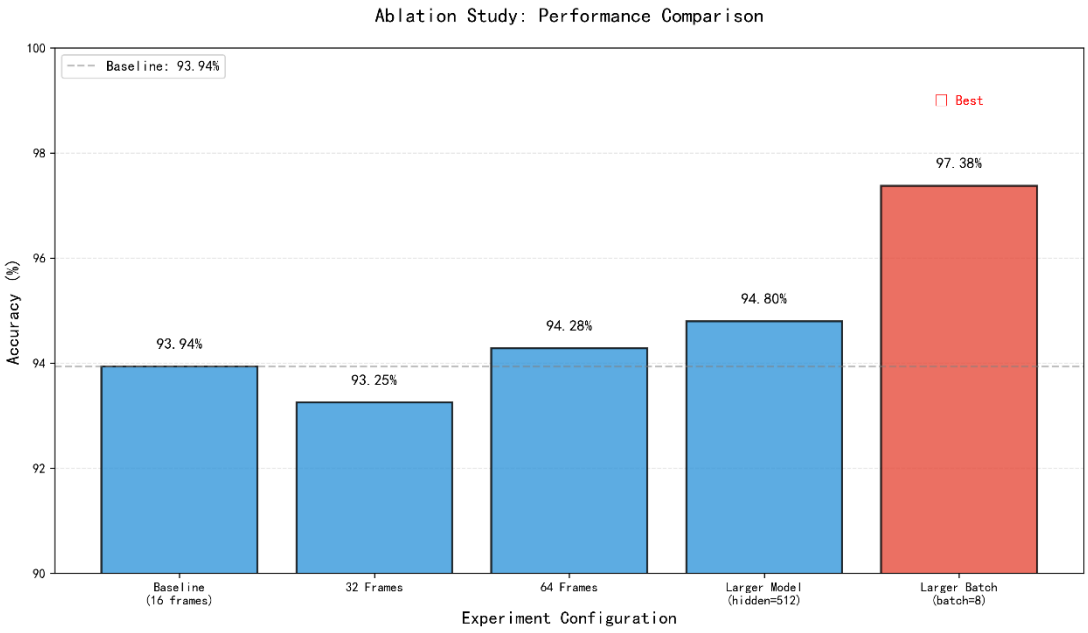


**Figure 2.** Accuracy comparison

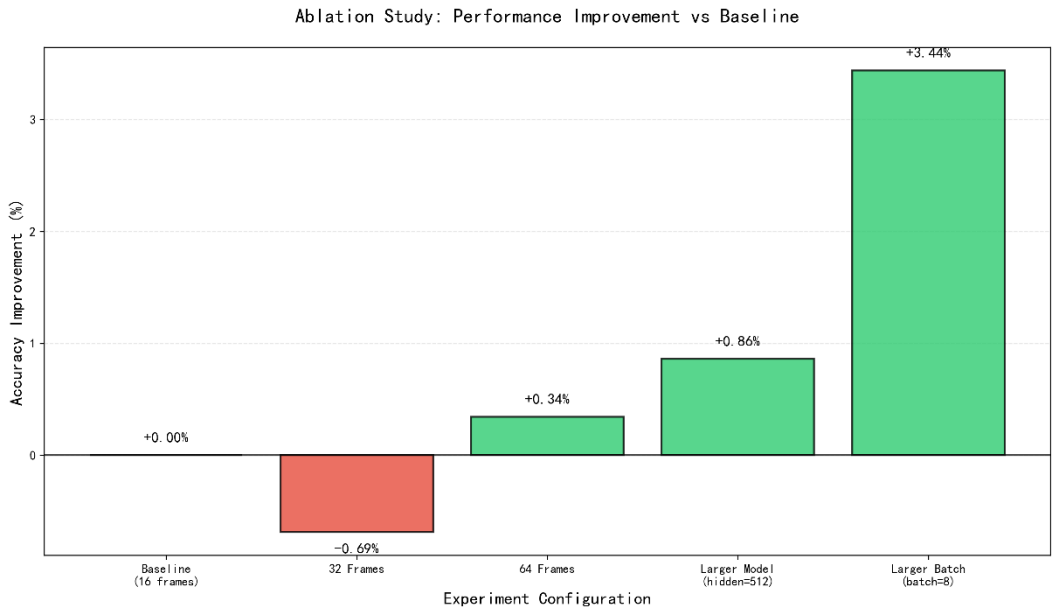


**Figure 3.** Improvement comparison

3.3.1 Influence of temporal sampling length $T$

The experiments show that increasing the number of frames does not necessarily lead to performance improvement. The Baseline ($T = 16$) reaches 93.94%; when $T$ increases to 32, the accuracy decreases to 93.25%; when further increased to 64 frames, it rises to 94.28% (see Figure 2). This non-monotonic phenomenon of "first decreasing and then increasing" indicates that, in the sports action scenario of Penn Action, a length of 32 frames may introduce more action phase redundancy and background interference, leading to "dilution" of key discriminative segments during the temporal aggregation stage. Only after 64 frames cover more complete action cycles does a slight gain reappear.

More importantly, increasing $T$ significantly raises the training cost. In Figure 4, the 64-frame configuration reaches a training time of 16.53 min, which is significantly higher than the 4.72 min of the Baseline, indicating that this gain is accompanied by a relatively high efficiency cost (see the Pareto discussion in Section 3.4).

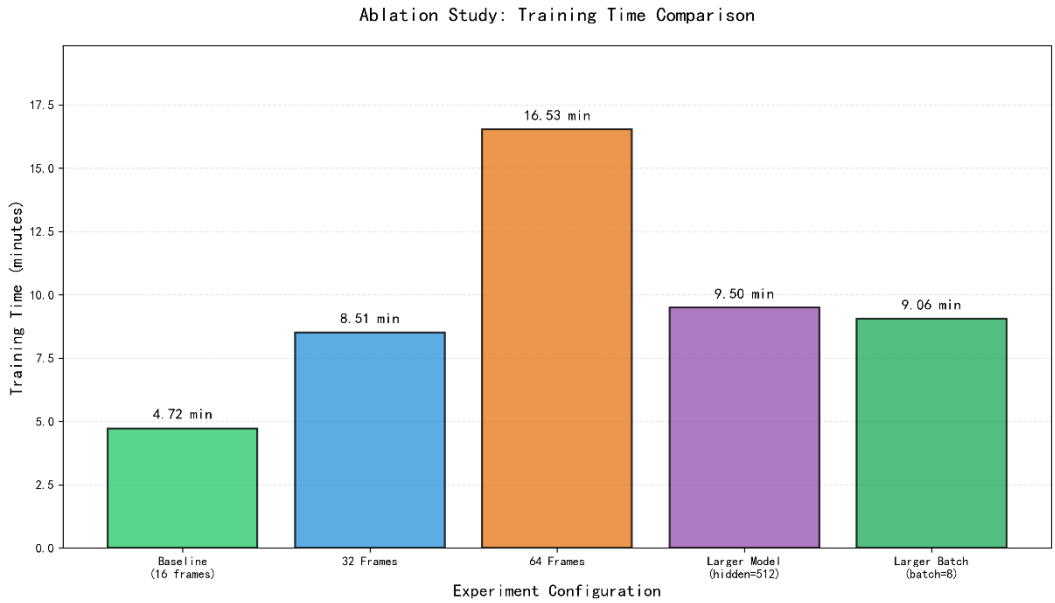


**Figure 4.** Time comparison

3.3.2 Influence of model capacity (hidden dimension)

Under the condition that $T$ remains unchanged, increasing the hidden dimension (Larger Model, hidden = 512) improves the accuracy to 94.80%. Compared with "simply stacking temporal length," moderately increasing representation capacity is more helpful for capturing the composite discriminative patterns of "joint coordination + equipment/scene cues" in sports actions, reflecting the stabilizing effect of capacity expansion on fine-grained classification boundaries.

3.3.3 Influence of batch size (key finding)

Compared with the above factors, the gain brought by Batch Size is the most significant. Larger Batch (batch = 8) achieves

the highest accuracy of 97.38% (Figure 2). The absolute improvement over Baseline is:

$$\Delta = 97.38 - 93.94 = 3.44 \text{ (percentage points)}$$

The relative improvement is approximately:

$$\frac{97.38-93.94}{93.94} \times 100\% \approx 3.66\%$$

As shown in Figure 3, optimization of Batch Size brings a leap in performance, indicating that in this task the stability of optimization dynamics (e.g., more stable gradient estimation and more reliable BN statistics) plays a decisive role in model generalization ability.

For convenience of summary, Table 2 presents the core results of the ablation settings in this paper (corresponding to Figure 2 / Figure 3):

**Table 2.** Summary of ablation experiments (Relative to baseline)

| Parameter Setting | Accuracy (%) | Δ (pp) |
|---|---|---|
| Baseline (16 frames) | 93.94 | 0.00 |
| 32 frames | 93.25 | -0.69 |
| 64 frames | 94.28 | +0.34 |
| Larger Model (hidden=512) | 94.80 | +0.86 |
| Larger Batch (batch=8) | 97.38 | +3.44 |

### 3.4 Computational efficiency and pareto trade-off

While improving accuracy, the training cost of the model must be quantitatively evaluated. Figure 4 shows that the Baseline training time is 4.72 min; the 64-frame configuration takes 16.53 min; the best-performing Larger Batch takes approximately 9.06 min; and the Larger Model takes approximately 9.50 min. This indicates that:

- The accuracy gain brought by extending temporal length is limited, but the cost increases significantly (approximately linear or even faster growth), which is not conducive to rapid iteration and deployment.
- Optimization of Batch Size/model capacity can obtain greater accuracy gains within "minute-level training time."

Further, Figure 5 and Figure 6 jointly depict the two-dimensional relationship between accuracy and training time. From the perspective of Pareto optimality, Larger Batch lies in the advantageous region of "high accuracy–relatively low training time," and can be regarded as a better trade-off point under the current experimental conditions. In contrast, the 64-frame configuration falls into a non-optimal region of "significantly increased time consumption but limited accuracy improvement."

### 3.5 Per-class performance and confusion matrix analysis (per-class and confusion analysis)

To understand the error patterns and interpretability of SportNet, we further report the per-class accuracy in Figure 7 and the confusion matrix in Figure 8.

Overall, the model performs more robustly on categories with strong rhythm and stable geometric structure (some categories can reach near-perfect recognition performance). In contrast, errors are mainly concentrated on category pairs with similar motion trajectories, identical equipment, and key differences occurring in short phases. Taking tennis categories as an example, the recognition accuracy of Tennis_forehand is significantly lower (56.1%), and the confusion matrix shows obvious misclassification aggregation toward Tennis_serve. Such confusion is usually caused by the superposition of the following factors:

- **Phase overlap:** Local postures such as racket swinging or arm lifting are highly similar in certain key frames, making it difficult for temporal pooling to preserve the "decisive moment."
- **Viewpoint and depth variation:** 2D pose is insensitive to motion along the depth direction; under certain viewpoints, the difference between forehand and serve is compressed by projection.
- **Dual-edged effect of equipment and scene priors:** Racket and court background provide "tennis domain" cues, but offer limited help for distinguishing "tennis subcategories," and may instead amplify confusion caused by pose similarity.

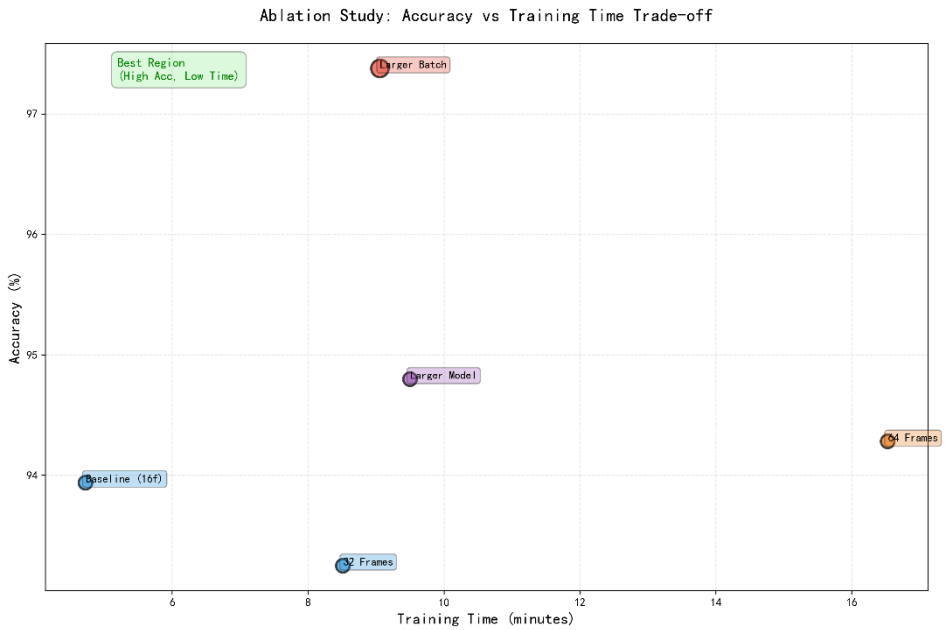


**Figure 5.** Efficiency scatter

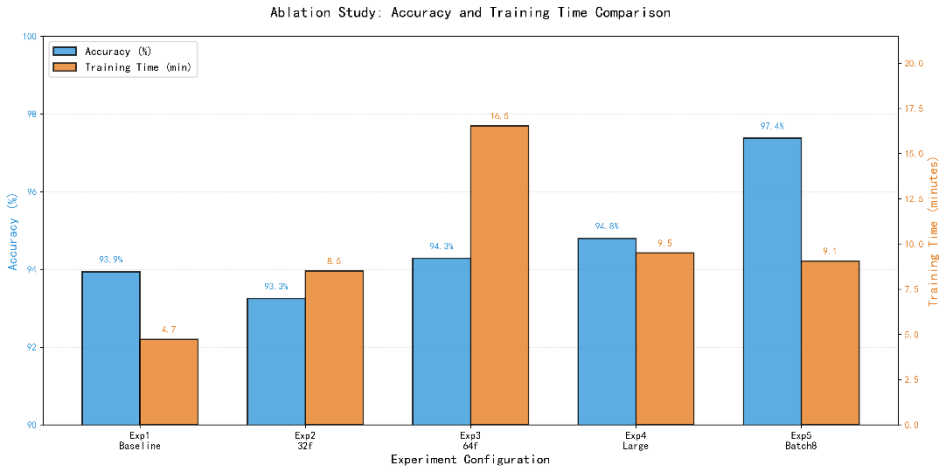


**Figure 6.** Combined comparison

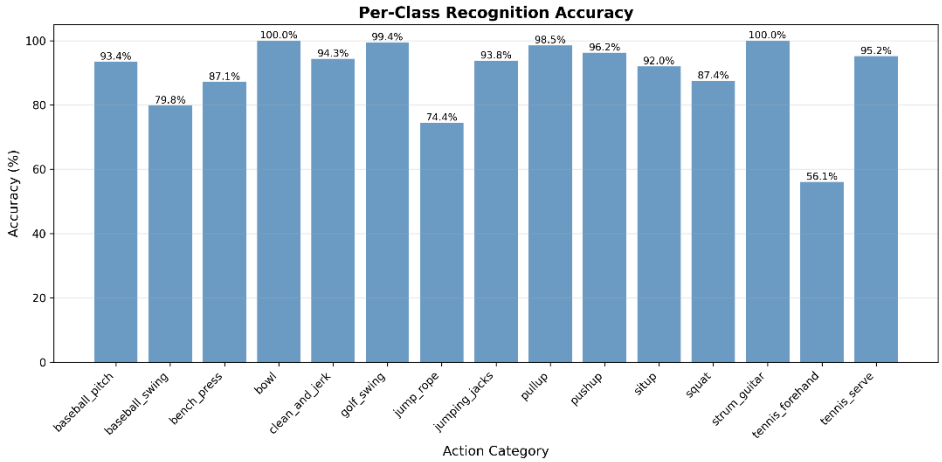


**Figure 7.** Per-class accuracy

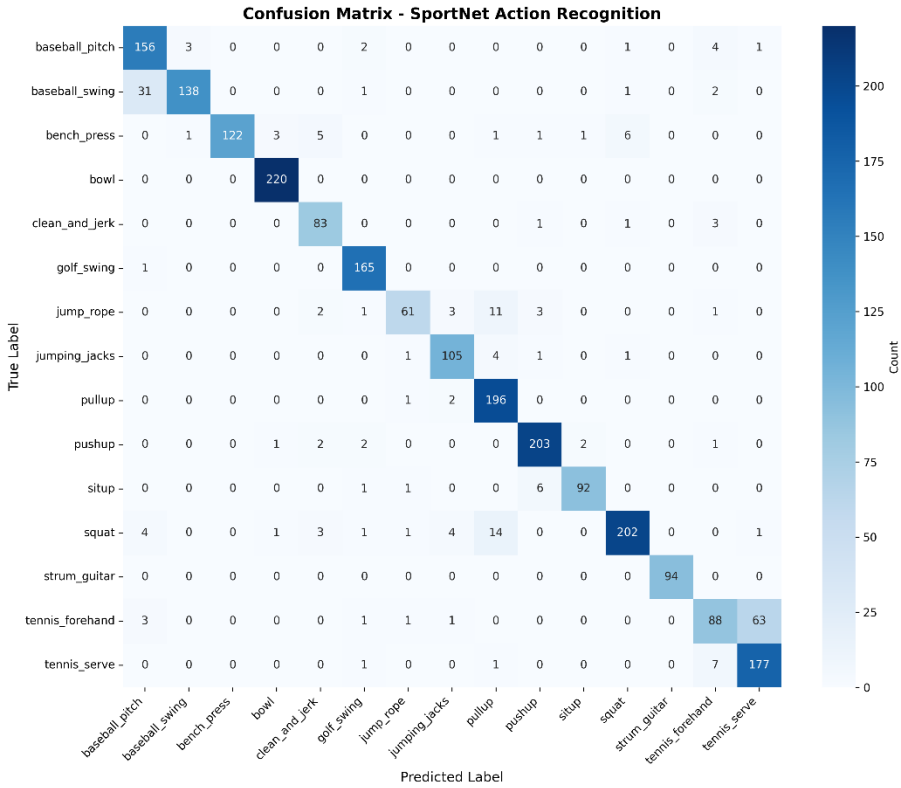


**Figure 8.** Confusion matrix

### 3.6 Discussion

From the comprehensive comparison experiments and ablation results, it can be seen that the performance improvement of SportNet does not mainly rely on "longer temporal sequences," but rather on a "more stable representation learning process." On the one hand, increasing the sampling length $T$ exhibits non-monotonic gains: the slight decrease from $T = 16$ to $T = 32$ and the small rebound at $T = 64$ together indicate that, for sports action data such as Penn Action, long sequences easily introduce phase redundancy and background/equipment interference, thereby diluting key discriminative frames during global temporal pooling. At the same time, the training cost increases significantly, making this path not cost-effective in engineering practice. On the other hand, Larger Batch improves the accuracy to 97.38% ($\Delta = 3.44$ percentage points compared with Baseline) while maintaining minute-level training time, indicating that the stability of optimization dynamics (smoother gradient estimation and more reliable BN statistics) is more critical for the generalization of fine-grained sports actions. Therefore, under the dual objectives of "accuracy–efficiency," the better practical choice under the experimental conditions of this paper is to maintain a relatively short $T$(16 frames) and prioritize tuning Batch Size/training strategies, rather than blindly extending the temporal input length.

Meanwhile, the per-class accuracy and confusion matrix reveal that the current main bottleneck of the model is concentrated on hard-pairs (Tennis_forehand and Tennis_serve). Such actions often share equipment and scene priors, and their key differences occur in short phases accompanied by significant motion along the depth direction, making it difficult for 2D pose projection to fully express "spatial hierarchy differences." This indicates that the current "global aggregation" still has limited sensitivity to critical moments. Future improvements can be explored along two paths: first, introducing phase-sensitive temporal modeling (e.g., key frame selection or attention mechanisms) to explicitly strengthen the weights of decisive segments; second, integrating more discriminative geometric and interaction cues (3D pose, equipment/hand interaction constraints, or finer-grained local region representations) to reduce the confusion of similar actions in the projection space. Overall, the results of this paper support the effectiveness of "lightweight multimodal fusion + stable training strategies" in sports action recognition, and also clarify the direction for improving fine-grained discrimination capability in the next step.

## 4. CONCLUSION

This paper proposed and validated a lightweight multimodal fusion framework, SportNet, for the Penn Action sports action recognition task. Experimental results show that the proposed method achieved 97.38% Top-1 Accuracy in video-level action classification, and demonstrated stronger overall competitiveness compared with representative existing methods. Systematic ablation further revealed that, under the data distribution and experimental settings of this study, simply extending the temporal sampling length brought limited gains and was accompanied by a significant increase in training cost. In contrast, optimizing training strategies (especially Batch Size) can bring larger generalization improvements within a controllable computational budget, highlighting the importance of "stable optimization dynamics" for fine-grained sports action recognition.

In addition, the per-class accuracy and confusion matrix analysis revealed that the main error source of the current method was concentrated on hard-pairs with highly similar motion trajectories and shared scene/equipment priors (Tennis_forehand and Tennis_serve). This phenomenon suggests that 2D pose and global temporal aggregation still have representation bottlenecks when handling short critical phases and depth-direction differences. Future work can proceed from two aspects: first, introducing more phase-sensitive temporal modeling mechanisms to enhance key frame discrimination capability; second, integrating stronger geometric constraints such as 3D pose or equipment interaction to improve the separability of similar actions. Overall, while ensuring training efficiency, SportNet achieves high-accuracy recognition and provides a practical technical solution and experimental basis for intelligent sports analysis systems oriented toward real-time feedback.

## REFERENCES


[1] Pirsiavash, H., Vondrick, C., Torralba, A. (2014). Assessing the quality of actions. In European Conference on Computer Vision, pp. 556-571. https://doi.org/10.1007/978-3-319-10599-4_36

[2] Zhang, W., Zhu, M., Derpanis, K.G. (2013). From actemes to action: A strongly-supervised representation for detailed action understanding. In 2013 IEEE International Conference on Computer Vision (ICCV), Sydney, Australia. pp. 2248-2255. https://doi.org/10.1109/ICCV.2013.280

[3] Laptev, I. (2005). On space-time interest points. International Journal of Computer Vision, 64(2): 107-123. https://doi.org/10.1007/s11263-005-1838-7

[4] Wang, H., Schmid, C. (2013). Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3551-3558. https://doi.org/10.1109/CVPR.2011.5995407

[5] Simonyan, K., Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27: 1-9.

[6] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, pp. 4489-4497.

[7] Carreira, J., Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 6299-6308. https://doi.org/10.1109/CVPR.2017.502

[8] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 6450-6459. https://doi.org/10.1109/CVPR.2018.00675

[9] Feichtenhofer, C., Fan, H., Malik, J., He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer

Vision, Seoul, Korea (South), pp. 6202-6211. https://doi.org/10.1109/ICCV.2019.00630

[10] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950. https://doi.org/10.48550/arXiv.1705.06950

[11] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European Conference on Computer Vision, pp. 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[12] Lin, J., Gan, C., Han, S. (2019). TSM: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), pp. 7083-7093. https://doi.org/10.1109/ICCV.2019.00718

[13] Feichtenhofer, C. (2020). X3D: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 203-213. https://doi.org/10.1109/CVPR42600.2020.00028

[14] Bertasius, G., Wang, H., Torresani, L. (2021). Is space-time attention all you need for video understanding? In Proceedings of the 38th International Conference on Machine, PMLR 139.

[15] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. (2021). ViVit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, pp. 6836-6846. https://doi.org/10.1109/ICCV48922.2021.00676

[16] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 7291-7299. https://doi.org/10.1109/CVPR.2017.143

[17] Sun, K., Xiao, B., Liu, D., Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 5693-5703. https://doi.org/10.1109/CVPR.2019.00584

[18] Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 7444-7452. https://doi.org/10.1609/aaai.v32i1.12328

[19] Shi, L., Zhang, Y., Cheng, J., Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 12026-12035. https://doi.org/10.1109/CVPR.2019.01230

[20] Parmar, P., Morris, B.T. (2019). What and how well you performed? A multitask learning approach to action quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 304-313. https://doi.org/10.1109/CVPR.2019.00039

[21] Shao, D., Zhao, Y., Dai, B., Lin, D. (2020). Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 2616-2625. https://doi.org/10.1109/CVPR42600.2020.00269

[22] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.

[23] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on Machine Learning, PMLR 37: 448-456.

[24] Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836. https://doi.org/10.48550/arXiv.1609.04836.

[25] Guo, F.Z., Kong, J., Jiang, M. (2020). Action recognition based on adaptive fusion of RGB and skeleton features. Laser & Optoelectronics Progress, 679(20): 310-319. https://doi.org/10.3788/lop57.201506

[26] Cao, C., Zhang, Y., Zhang, C., Lu, H. (2016). Action recognition with joints-pooled 3d deep convolutional descriptors. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp. 3324-3330.

[27] Liu, M., Meng, F., Chen, C., Wu, S. (2019). Joint dynamic pose image and space time reversal for human action recognition from videos. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 8762-8769. https://doi.org/10.1609/aaai.v33i01.33018762