Robust 2D Skeleton Action Recognition via Decoupling and Distilling 3D Latent Features

Xiangyue Zhang, *Student Member, IEEE*, Yifan Jia, Jiaxu Zhang, *Student Member, IEEE*, Yijie Yang, and Zhigang Tu, *Senior Member, IEEE*

***Abstract*—Human skeletons provide a compact representation for action recognition. Compared to 3D skeletons, 2D skeletons lack view-independence and depth, making them less robust for motion analysis. However, 3D skeleton data requires specialized hardware, limiting its practicality, especially in outdoor or dynamic settings. In contrast, 2D skeletons can be extracted from standard RGB videos, making them more accessible. To address this, we propose 2D^3-SkelAct, a 2D skeleton-based action recognition model. It maps 2D inputs to a 3D latent space, where pose and view features are decoupled. Additionally, 2D^3-SkelAct distills motion cues from 3D models, enhancing motion detail capture while keeping the benefits of 2D data. Specifically, the pipeline of our 2D^3-SkelAct consists of two steps: *pose-view decoupling* and *pose-view distilling*. First, we use a spatio-temporal transformer to decouple 2D skeleton sequences into latent pose and view features, enhancing the model's ability to learn motion dynamics. Next, these decoupled features are separately integrated into the 2D skeleton model through two cross-attention modules, allowing it to extract discriminative motion features while mitigating uncertainties in 3D viewpoint and depth. Additionally, we distill motion cues from 3D models to compensate for the limitations of 2D skeletons. Remarkably, our model can seamless integrate with various skeleton feature extractors. We validate the proposed 2D^3-SkelAct through extensive experiments, demonstrating its adaptability across different model architectures as where consistent improvement achieving. When combined with advanced skeleton feature extractors, 2D^3-SkelAct achieves state-of-the-art performance in 2D skeleton-based action recognition.**

***Index Terms*—Action recognition, skeleton representation learning, decoupling, distillation.**

I. Introduction

HUMAN action recognition plays a crucial role in various applications, such as intelligent video surveillance [1], [2], human-computer interaction [3], and sports analysis [4]. Skeleton-based action recognition has emerged as an effective approach due to its efficiency in capturing critical human movement information, significantly reducing computational demands compared to traditional video-based methods. Therefore, skeleton-based human action recognition has become a key research area in computer vision.

Traditionally, skeleton-based action recognition relies on 3D skeleton data, which provides depth-aware and view-independent representations captured by specialized equipment like depth cameras [5]. This approach enhances robustness and accuracy by offering detailed structural information about human movement. However, using 3D skeletons for action recognition presents logistical and technical challenges. Acquiring 3D data requires expensive hardware and has limited spatial coverage, making it impractical for large-scale or outdoor applications. The challenge grows when tracking multiple subjects in dynamic or crowded environments, where occlusions and interactions degrade data quality. These limitations make 3D-based methods costly and difficult to deploy in real-world scenarios.

In response to these challenges, there has been a growing interest in methods that estimate 3D skeletons from 2D poses captured in standard RGB video footage [6], [7], [8], [9]. These approaches typically leverage deep learning models to predict 3D joint locations from monocular images or videos, tackling the fundamental difficulty of inferring 3D structures from 2D projections. While this strategy removes the need for expensive and complex 3D capture equipment, making 3D action recognition more accessible, it introduces its own limitations. Converting 2D poses into 3D inherently leads to depth ambiguities, causing inaccuracies in reconstructed poses. Additionally, this process is highly susceptible to errors from occlusions, depth estimation uncertainties, and the natural variability of human movement, making it a persistent challenge [10]. Moreover, these estimation methods rely heavily on the availability and quality of training data. The performance of models trained on datasets captured in controlled environments may degrade when applied to "in-the-wild" scenarios, characterized by diverse and unpredictable settings.

In contrast, 2D skeleton data, which is directly derived from RGB video, offers a more accessible alternative for skeleton-based action recognition [11], [12]. Without the need for specialized hardware, it offers a scalable solution across diverse environments. Since 2D skeletons are obtained from standard video footage, they can adapt to various environmental conditions without requiring controlled lighting or

Received 5 December 2024; revised 21 March 2025; accepted 16 April 2025. Date of publication 21 April 2025; date of current version 6 October 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3015600 and in part by the Fundamental Research Funds for Central Universities under Grant 2042023KF0180 and Grant 2042025KF0053. This article was recommended by Associate Editor Q. Ye. *(Xiangyue Zhang and Yifan Jia are co-first authors.) (Corresponding author: Zhigang Tu.)*

Xiangyue Zhang, Jiaxu Zhang, Yijie Yang, and Zhigang Tu are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: tuzhigang@whu.edu.cn).

Yifan Jia is with the Department of Pain, Renmin Hospital of Wuhan University, Wuhan 430060, China.

Digital Object Identifier 10.1109/TCSVT.2025.3562877

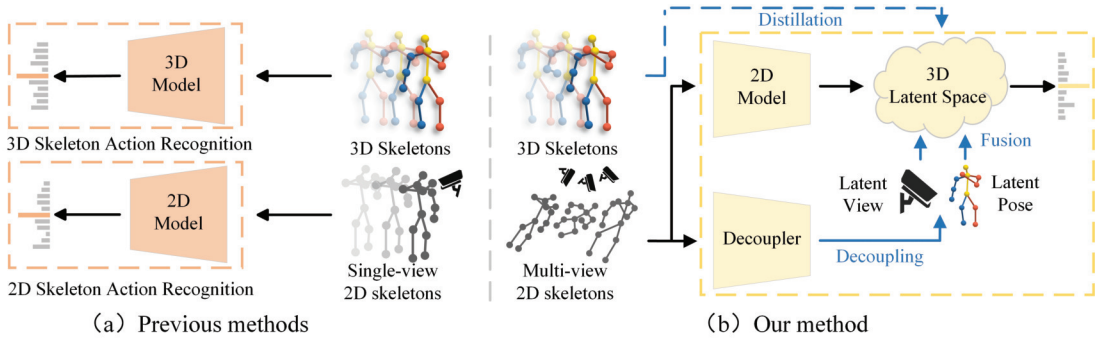

Fig. 1. (a) The previous methods use separate models for recognizing 2D and 3D skeleton actions. Because of variations in viewpoint and pose depth, 2D models often have lower accuracy compared to 3D models. (b) Our 2D^3-SkelAct is designed for 2D skeleton action recognition but achieves competitive performance with the 3D models. The key contribution is decoupling 3D pose and view features using 2D inputs and distilling motion clues from 3D models.

backgrounds. This flexibility makes them particularly useful in dynamic and unstructured settings where deploying 3D capture technology is impractical. However, 2D skeletons are inherently limited by their sensitivity to viewpoint changes and their inability to fully capture depth and fine-grained motion details, reducing their robustness in action recognition tasks.

The discrepancy between the depth-rich accuracy of 3D data and the pragmatic flexibility of 2D data drives the exploration of more advanced solutions. To address the contrast issue of *2D vs 3D* skeleton data, we propose a novel model 2D^3-SkelAct for robust 2D skeleton human action recognition. Our 2D^3-SkelAct effectively converts 2D skeleton data into a 3D latent feature space through the strategic implementation of *decoupling* and *distillation* techniques, thus capitalizing on the advantages offered by 3D skeleton data without directly using it as input. In brief, as shown in Figure 1, central to our pipeline is the fantastic decoupling of pose and view features within the 3D latent space, and utilizing 2D inputs to distill enriched motion information from 3D models.

Specifically, our 2D^3-SkelAct involves two key steps, i.e., *pose-view decoupling* and *pose-view distilling*. In the first step, we use a spatio-temporal transformer [13] to decouple 2D skeleton sequences into 3D latent pose and view features, enhancing their ability to capture fine-grained motion and view-dependent variations. To achieve this, we introduce a 2D-to-3D supervision strategy, leveraging 3D ground truth to guide the learning of latent poses and views. It reconstructs 3D skeletons from latent poses and generates view-dependent 2D skeletons by mixing latent pose and view features.

In the second step, we integrate the decoupled pose and view features into a 2D skeleton feature extractor using pose-aware and view-aware cross-attention modules. The latent pose feature refines depth perception for distinct motion representation, while the latent view feature helps handle camera viewpoint variations. Additionally, we apply knowledge distillation in the pose-aware module to incorporate extra 3D motion cues during training. This strategy enhances motion feature extraction, mitigates viewpoint and depth uncertainties, and overcomes the inherent limitations of 2D-based models.

Notably, our 2D^3-SkelAct model requires only wild 2D skeleton inputs during inference. This simplifies the action recognition process in real-world applications while maintaining high accuracy and adaptability. Overall, our contributions can be summarized as follows:

- We propose 2D^3-SkelAct, a novel 2D skeleton action recognition model that bridges the gap between 2D and 3D skeleton features. By strategically leveraging 3D latent feature space, the model enhances the robustness of 2D skeleton-based action recognition, overcoming limitations in depth representation and view-dependence.
- A pose-view decoupler is proposed, supported by a 2D-to-3D supervision strategy. This decoupler explicitly separates pose and view features into a 3D latent space during training, enabling refined motion representation while requiring only 2D skeleton inputs during inference.
- Two cross-attention modules are introduced to distill discriminative motion features from the decoupled latent pose and view features. These modules effectively address uncertainties in depth and viewpoint, allowing the model to extract fine-grained motion features and adapt to real-world variations.
- Extensive experiments are conducted to validate the effectiveness and versatility of 2D^{3}SkelAct, the significant performance gain across various model architectures demonstrates its adaptability.

II. Related Work

A. Skeleton-Based Action Recognition

Skeleton-based action recognition has evolved significantly, from early manual feature design methods [14], [15] to modern deep learning approaches. Initially, the development of action recognition systems was constrained by the limitations of manually designed features, which focused on joint or body part attributes. These early approaches struggled to encapsulate the complex, semantic nuances of skeletal movements, rendering them inadequate for capturing the full scope of human actions.

The advent of deep learning technologies, notably Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), marked a pivotal shift in the field. RNN-based methods [16], [17] capitalized on their ability to process sequential data, addressing the temporal dynamics of variable-length skeleton sequences. Meanwhile, CNN-based methods [18], [19], [20], [21], [22] introduced a novel approach

by encoding these sequences into pseudo-images, facilitating spatial-temporal representation learning. Despite these advancements, both RNNs and CNNs encountered difficulties in fully grasping the intricate spatial-temporal interplay inherent in skeletal movements.

The emergence of Graph Convolutional Networks (GCNs) [23], [24], [25], [26], [27], [28] represented a significant leap forward, treating skeleton data as graphs to capture the relationships between joints. This approach allowed for a more nuanced modeling of joint dynamics over time, significantly improving action recognition accuracy. Several recent methods have expanded on this foundation. VN-GAN [29] leverages Generative Adversarial Networks (GANs) to normalize skeletons to a unified viewpoint, enhancing cross-view consistency. DKE-GCN [30] improves efficiency by employing decoupled knowledge distillation for lightweight models. TranSkeleton [31] incorporates a topology-aware Transformer to better capture spatial-temporal dependencies, while [32] enhances robustness through cross-view learning with multi-scale fusion.

Despite these advancements, both 3D and 2D skeleton-based action recognition systems face inherent limitations. 3D skeleton action recognition systems, while robust and capable of capturing depth information, are often constrained by their reliance on specialized hardware and are limited in their application to environments where such equipment can be feasibly deployed. Furthermore, the processing of 3D data requires significant computational resources, limiting its scalability and flexibility. On the other hand, 2D skeleton action recognition systems, which are more readily derived from widely available video data, suffer from a critical lack of depth information. This absence reduces robustness, particularly in scenarios involving complex actions or varying viewpoints. The 2D approach's inability to account for view independence significantly hampers its effectiveness, making it challenging to achieve high accuracy in diverse conditions.

To address these issues, the proposed 2D^3-SkelAct method enhances mainstream network structures, enabling them to extract rich motion features while mitigating the limitations of existing methods. While other works focus on comprehensive skeleton motion representations through contrastive learning, such as AS-CAL [33], ISC [34], MS^{2}L [35], CrosSCLR [36], and AimCLR [37], these methods are tailored for 3D skeletal data and cannot fully exploit 3D information for 2D skeleton sequences. In contrast, inspired by [38] and [39], which enhance shape and skeleton perception, our 2D^3-SkelAct introduces 3D latent information into 2D models, effectively combining the strengths of both modalities and enhancing robustness in complex action recognition tasks.

B. Knowledge Distillation

Knowledge Distillation (KD) is a critical technique in model compression, facilitating the transfer of knowledge from a larger, more capable teacher model to a smaller, efficient student model. The essence of KD lies in utilizing the soft targets generated by the teacher model as a guiding mechanism for training the student model, significantly enhancing the student's performance by capturing the intricate relationships among categories [40]. Recent advancements in KD have expanded its application to include both intermediate feature distillation and logit distillation, marking a significant shift towards a more comprehensive approach to knowledge transfer. Intermediate feature distillation [41], [42] focuses on leveraging the rich information embedded in the intermediate layers of the teacher model, while logit distillation [43], [44] continues to utilize the traditional method of transferring knowledge through the soft targets of the output layer. This dual approach enables a more nuanced and effective transfer of knowledge, enriching the student model with high-level output and deeper feature-level insights.

Moreover, the advent of online KD [45], [46] has introduced a dynamic aspect to the distillation process, allowing for a simultaneous and mutual learning process among multiple student models. This method eliminates the static teacher-student roles, fostering a more flexible and interactive learning environment. In action recognition tasks, KD has been applied to leverage high-level knowledge from teacher networks, such as depth [47], temporal information [48], and optical flow [49], [50], [51]. In our study, we use online distillation to transfer depth and viewpoint information from a 3D skeleton to a 2D skeleton model. Our 2D^3-SkelAct is the first method to enhance 2D skeleton action recognition by distilling motion cues from 3D skeletons, thereby incorporating complementary 3D information and effectively leveraging the strengths of both modalities to overcome the limitations of traditional 2D action recognition systems.

III. Method

As illustrated in the left side of Figure 2, given a 2D skeleton sequence $X \in \mathbb{R}^{T\times J\times 2}$, where T is the sequence length and J is the number of the joints, our 2D^3-SkelAct decouples it into the latent pose ψ and view ϕ features through a motion decoupler $\mathcal{E}(\cdot)$. This process can be formulated as:

$$\mathcal{E}(X) \mapsto \{\psi, \phi\}. \tag{1}$$

Simultaneously, the 2D motion feature $\boldsymbol{x} \in \mathbb{R}^{T\times J\times C}$ is extracted by a 2D model $\mathcal{F}(\cdot)$ from the input skeleton sequence, where C is the channel dimension. Subsequently, the extracted 2D motion feature is fused with the decoupled latent pose and view features in a 3D latent space through the carefully designed pose-aware cross-attention module $f_p(\cdot)$ and view-aware cross-attention module $f_v(\cdot)$, respectively. These cross-attention modules exploit beneficial 3D information to enrich the 2D motion features. This process is formulated as:

$$\mu(f_p(\psi, \mathbf{x}), f_v(\phi, \mathbf{x})) \mapsto \{\mathbf{x}^{\psi}, \mathbf{x}^{\phi}\}, \tag{2}$$

where $\mu(\cdot)$ is the cross-domain connections between the pose-aware and view-aware branches. Finally, two classifiers are employed to predict the classification score using the pose-aware motion feature $\mathbf{x}^{\psi}$ and the view-aware motion feature $\mathbf{x}^{\phi}$ respectively, and the final result is obtained by averaging the outputs of these two classifiers.

A. Pose-View Decoupling

The pose-view decoupling in 2D^3-SkelAct, illustrated on the right side of Figure 2, involves pre-training a 2D skeleton

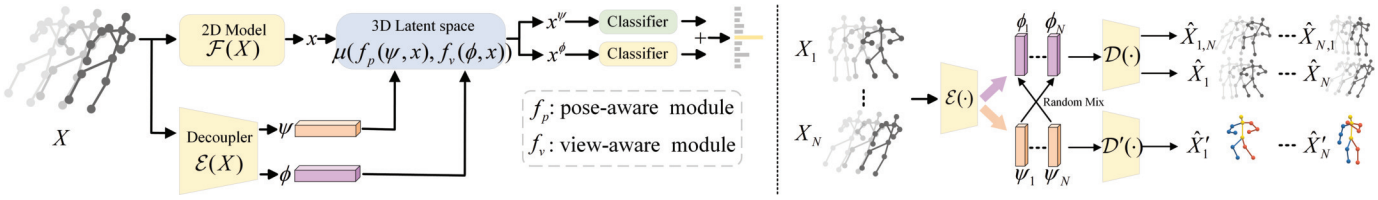

Fig. 2. **On the left** is an overview of the 2D^3-SkelAct pipeline. 2D^3-SkelAct decouples latent pose ψ and view ϕ from the input 2D skeleton sequence X, refines the 2D motion features $\mathbf{x}$ within a 3D latent space, and uses the 3D-enhanced 2D motion features for classifying skeleton actions. **On the right** is the structure of the 2D^3-SkelAct decoupler. The decoupler $\mathcal{E}(\cdot)$ is pre-trained through our 2D-to-3D supervision strategy, which utilizes two decoders $\mathcal{D}(\cdot)$ and $\mathcal{D}'(\cdot)$ to reconstruct the pose-view mixed skeletons.

encoder to separate latent pose and view features from input 2D sequences. These latent features implicitly encode 3D depth and viewpoint information, which are crucial for robust skeleton-based action recognition. To achieve this, inspired by [13] we employ a spatio-temporal transformer to build the 2D skeleton encoder. The spatial transformer block computes self-attention between joints within each frame, capturing spatial relationships across the body. Simultaneously, the temporal transformer block applies self-attention across frames, modeling global temporal dependencies to enhance motion understanding. These two blocks are utilized alternately to achieve enhanced spatio-temporal feature encoding. However, since 2D skeleton data inherently lacks depth and its viewpoint is implicitly included in the joint coordinates, it is challenging to recover this crucial information from 2D input. To address this challenge, as described below, we designed a 2D-to-3D supervision strategy to train our 2D skeleton decoupler.

2D-to-3D Supervision Strategy: Given a batch of 2D skeleton sequences $X = \{X_i\}_{i=1}^{N}$, where i denotes the i-th sequence and N denotes batch size. Each sequence is obtained from distinct camera viewpoints and expresses different actions. The spatial-temporal transformer encodes them into two batches of latent features, i.e., ψ and ϕ. Subsequently, we introduced two skeleton decoders, i.e., a 2D decoder $\mathcal{D}(\cdot)$ and a 3D decoder $\mathcal{D}'(\cdot)$, to reconstruct the corresponding 2D and 3D sequences using the extracted two batches of latent features. This approach, elaborated below, ensures that one batch is dedicated to capturing the pose-aware feature, while the other represents the view-aware feature. These decoders are also constructed with spatial-temporal transformer blocks.

Empirically, similar to the 3D pose estimation process [13], the reconstruction of 3D skeleton sequences can be achieved by decoding the extracted pose-aware features. Therefore, the 3D skeleton decoder takes the latent pose feature as input and generates the corresponding view-independent 3D skeleton sequence, which can be formulated as:

$$\hat{X'}_i = \mathcal{D}'(\psi_i), \tag{3}$$

where $\hat{X'}_i$ is the reconstructed i-th 3D skeleton sequence in a batch. Then, the 3D reconstruction loss $\mathcal{L}_{3D}(\hat{X'}, X')$ can be calculated by aligning the generated 3D sequence and the ground-truth 3D sequence to supervise the learning of the latent pose features, enabling the decoupler to extract pose depth information and construct a reasonable 3D latent space.

The 2D skeleton decoder serves two objectives. Firstly, it aims to utilize the decoupled pose feature ψ_i and the view feature ϕ_i to reconstruct the input 2D skeleton sequence X_i. Secondly, it seeks to estimate new 2D skeleton sequences by randomly combining the pose and view features. For instance, a 2D skeleton sequence $X_{j,k}$ can be estimated by decoding the pose feature ψ_j and the view feature ϕ_k, indicating the projection of the j-th skeleton action from the perspective of the k-th viewpoint. These processes can be formulated as:

$$\hat{X}_{a,b} = \mathcal{D}([\psi_a, \phi_b]), \tag{4}$$

where $a, b \in \{1, \ldots, N\}$ and $[\cdot]$ represents the feature concatenation. This formulation represents the 2D reconstruction process when $a = b$, and denotes the process for estimating mixed 2D sequences when $a \neq b$. Finally, the 2D reconstruction loss $\mathcal{L}_{2D}(\hat{X}, X)$ can be calculated by aligning the reconstructed 2D sequences with the input sequences. The mixed 2D estimation loss $\mathcal{L}_{mix}(\hat{X}, X)$ is calculated by aligning the estimated 2D sequences with the projected 3D ground-truth on the corresponding viewpoints. These two supervisions allow the pose and view features associated with 2D skeleton sequences to be fully decoupled within the constructed 3D latent space.

In our implementation, each of $\mathcal{L}_{3D}$, $\mathcal{L}_{2D}$, and $\mathcal{L}_{mix}$ consists of three components, i.e., the WMPJPE loss, the TC loss, and the MPJVE loss, which is inspired from pose estimation methods [13]. The decoupler and the decoders in our pose-view decoupling process can be optimized by:

$$\min \lambda_{3D}\mathcal{L}_{3D} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{mix}\mathcal{L}_{mix}, \tag{5}$$

where λ_{3D}, λ_{2D}, and λ_{mix} are the loss balancing factors.

B. Pose-View Distilling

The pre-trained decoupler in the pose-view decoupling step encodes pose and view features from 2D skeleton sequences into a 3D latent space. However, these features lack explicit motion information, making them insufficient for direct action classification. To address this, the pose-view distilling step in 2D^3-SkelAct leverages these decoupled features to enhance 2D skeleton models, extracting robust and discriminative motion features. This process involves two key challenges: (1) learning view-aware motion features that remain invariant to viewpoint changes, and (2) learning pose-aware motion features that incorporate 3D depth information. As illustrated in Figure 3, we tackle these challenges by constructing two network branches and designing two cross-attention modules to effectively integrate view and pose features into the motion representation.

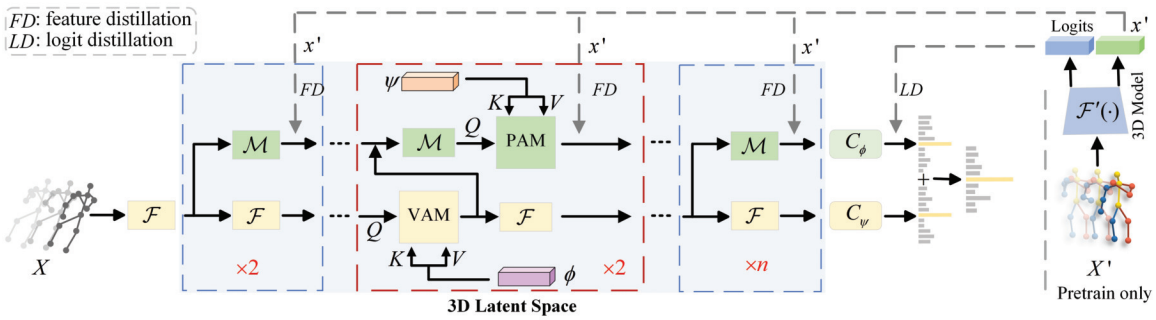

Fig. 3. The detailed structure of 2D^3-SkelAct pipeline. Our pipeline consists of two branches to enhance the 2D motion feature in the 3D latent space and distill rich motion clues from the 3D model. (The 3D model is not utilized during inference.)

1) View-Aware Representation Learning: Let $\mathbf{x}_i$ denote the 2D skeleton feature extracted by the i-th layer of the 2D model. We utilize a cross-attention to fuse the decoupled latent view feature ϕ with $\mathbf{x}_i$, thereby introducing camera viewpoint information to the 2D model. ϕ is used to generate *Key* and *Value* and $\mathbf{x}_i$ serves to generate *Query* in this attention block. This view-aware cross-attention module (VAM) can be formulated as:

$$Q = \mathbf{x}_i W_1, K = \phi W_2, V = \phi W_3,$$
$$x_i^{\phi} = \text{softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V + x_i, \tag{6}$$

where W is the linear mapping matrix and C is the channels. $\sqrt{C}$ represents a scaling factor to normalize the dot products.

2) Pose-Aware Representation Learning: Beyond the implicit viewpoint information, depth information is entirely absent in 2D skeleton sequences. To achieve pose-aware motion feature learning, we employ both cross-attention and knowledge distillation strategies to extract rich motion clues from the decoupled latent pose feature ψ and the 3D skeleton model. Specifically, a multi-layer perception block ($\mathcal{M}$) is utilized to map the 2D skeleton feature $\mathbf{x}_i$ to the 3D latent space, producing a mimetic 3D feature $\tilde{\mathbf{x}}_i$. Next, a cross-attention module is utilized to fuse the decoupled latent pose feature ψ with $\tilde{\mathbf{x}}_i$, thereby introducing pose depth information to the 2D model. Similar to the VAM, ψ is used to generate *Key* and *Value*, and $\tilde{\mathbf{x}}_i$ serves to generate *Query* in the attention block. This pose-aware cross-attention module (PAM) can be formulated as:

$$Q = MLP(\mathbf{x}_i)W_1, K = \psi W_2, V = \psi W_3,$$
$$x_i^{\psi} = \text{softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V + MLP(\mathbf{x}_i). \tag{7}$$

During training, we align the mimetic 3D feature $\tilde{\mathbf{x}}_i$ with the corresponding feature $\mathbf{x}'_i$ of the 3D model, thereby distilling supplemental motion clues absent in the 2D feature. Furthermore, logit knowledge distillation is applied to the classification scores of this branch to transfer pose-aware information from the 3D latent space fully.

The VAM and PAM introduced above can be applied to both the joint and temporal dimensions of the 2D skeleton features. This allows for emphasizing salient joints and critical timing in the skeleton sequences, considering both viewpoint and depth information.

3) Training: The 2D model, the 3D model, and the proposed PAM and VAM are trained end-to-end in our 2D^3-SkelAct. Notably, the 3D model does not participate in the inference process. Thus, our 2D^3-SkelAct only requires 2D skeleton input during inference and can obtain discriminative motion features for accurate action classification. The training objective can be formulated as:

$$\begin{aligned}\min\ & \lambda^{3D}\mathcal{L}_{CE}^{3D} + \lambda^{p}\mathcal{L}_{CE}^{p} + \lambda^{v}\mathcal{L}_{CE}^{v} \\ & + \lambda^{FD}\mathcal{L}_{MSE}^{FD} + \lambda^{LD}\mathcal{L}_{DKD}^{LD}\end{aligned} \tag{8}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss between ground truth and predicted labels. $\mathcal{L}_{MSE}$ is the mean squared error for feature distillation. L_{DKD} is the DKD loss [44] for logit distillation. λ^{3D}, λ^{p}, λ^{v}, λ^{LD} and λ^{KD} are loss balancing factors. Moreover, to bolster training stability, we utilize residual connections in the network blocks of our 2D^3-SkelAct.

IV. Experiments

A. Datasets and Evaluation Protocols

We employ four popular skeleton datasets to evaluate our 2D^3-SkelAct comprehensively.

1) NTU-RGB+D 60: NTU-RGB+D 60 (NTU-60), a comprehensive dataset for human action recognition, features 60 diverse action categories performed by 40 subjects, accumulating a total of 56,880 3D skeleton sequences. In our study, we adhere to the evaluation protocols recommended by the dataset authors: cross-subject (X-sub) and cross-view (X-view). The X-sub protocol splits the dataset based on the performers, using sequences from 20 subjects for training and the remaining 20 for testing. Conversely, the X-view protocol segregates the data based on the viewpoint, designating sequences from cameras 2 and 3 for training, with sequences captured by camera 1 reserved for testing.

2) NTU-RGB+D 120: An expansion of NTU-60, NTU-RGB+D 120 (NTU120) doubles the action categories to 120 and extends the dataset to 114,480 skeleton sequences performed by 106 subjects. This version introduces a more demanding evaluation protocol, cross-setup (X-set), which supersedes the X-view protocol from NTU-60. The X-set protocol categorizes sequences into 32 setups, differentiated

by camera distance and background, with half of these setups allocated for training and the other half for testing.

3) PKU-MMD: Following the previous method [34], we segment action instances in PKU-MMD [52] using temporal annotations for 3D action classification, applying the cross-subject protocol for dataset division. PKU-MMD consists of two phases: PKU-MMD I (PKU-I) and PKU-MMD II (PKU-II). PKU-I includes 18,841 training samples and 2,704 test samples. PKU-II is more challenging due to increased noise from wider viewpoint variations, containing 5,332 training samples and 1,613 test samples.

4) UAV-Human: UAV-Human [53] is a large-scale dataset designed for UAV-based human behavior analysis, containing 22,476 high-definition videos captured in both indoor and outdoor environments under diverse lighting and weather conditions. The use of drone footage introduces unique elevated viewpoints, creating additional challenges for action recognition due to variations in perspective and occlusions. To support robust evaluation, the dataset provides two cross-subject protocols: CSv1 and CSv2, which define different subject splits for training and testing.

5) Northwestern-UCLA: The Northwestern-UCLA skeleton dataset, introduced by [54], comprises 1,494 video clips spanning 10 action categories. Each action was recorded using three Kinect cameras from different viewpoints and performed by 10 individuals. Following the standard NW-UCLA evaluation protocol, we use data from two camera angles for training while reserving the third for validation.

B. Experimental Setting on Data

1) Challenging Multi-View Data Experiment Setting: The primary aim of our method is to enhance the robustness of 2D skeleton action models in real-world applications. Hence, in our experiments, we employed random camera views to simulate practical scenarios. To mimic camera view changes, we introduced a method adopted in NTU-60, NTU-120 and PKU-MMD that involves rotating the joint coordinates of a 3D skeleton sequence along three axes using a rotation matrix, thereby generating multi-view 2D skeleton sequences. We randomly select three angles, α, β, and γ, each uniformly distributed between $[-90°, 90°]$ for every sequence, and apply the rotation matrix $\mathbf{R} = \mathbf{R}_X(\alpha)\mathbf{R}_Y(\beta)\mathbf{R}_Z(\gamma)$ to the original coordinates of the skeleton sequence and get the transformed joint coordinates, where $\mathbf{R}_X(\alpha)$,$\mathbf{R}_Y(\beta)$ and $\mathbf{R}_Z(\gamma)$ denote the rotation matrix of the X, Y, and Z axes. Thus, three basic rotation matrices with rotation angles about X, Y, and Z axis are given as follows:

$$\mathbf{R}_X(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \tag{9}$$

$$\mathbf{R}_Y(\beta) = \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \tag{10}$$

$$\mathbf{R}_Z(\gamma) = \begin{bmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

Besides, single-person actions prevail, prompting our focus on the primary individual in each skeleton sequence to mitigate the impact of secondary actors. Following [36], we discard irrelevant frames and standardize each sequence to 50 frames via linear interpolation. Moreover, for a fair comparison, we only use joint stream, as they are adopted by most of the previous methods. In this study, we did not use any other data augmentation to ensure that improvements could be attributed to our model innovations.

2) Vanilla NTU Data Experiment Setting: To provide a fair comparison with previous action recognition models, we also follow the established approach by conducting experiments on the original NTU dataset.

C. Implementation Details

1) Hyperparameters: We integrate VAM and PAM into the third and fourth blocks of the 2D model to enhance robustness by emphasizing key joints and critical time frames while considering viewpoint and depth. The latent pose and view embeddings are 256-dimensional, with a linear layer ensuring consistency before cross-attention. Both modules use four attention heads with a 1024-dimensional hidden layer. For training, we use SGD for both pre-training and the full pipeline. Pre-training follows a batch size of 1024, a dropout rate of 0.1, and GELU activation. In the pose-view decoupling stage, we adopt MixSTE [13] settings, with a 0.001 learning rate, a Multi-Step LR scheduler (milestones at 90 and 130), and 150 training epochs. In the pose-view distilling stage, we follow ST-GCN [23] for consistency across GCN models, using SGD with Nesterov momentum (0.9) and weight decay (0.0001). The learning rate is 0.1, with 70 training epochs and milestones at 40 and 60. For NTU RGB+D [55], NTU RGB+D 120 [56], PKU-MMD [52], and UAV-Human [53], we set the batch size to 256, with T = 50 and M = 1.

2) Detailed Architectures of the Modules: The $2D^3$-SkelAct framework (Table I) includes a decoupler that pre-trains a 2D skeleton encoder to separate pose (ψ) and view (ϕ) features using alternating temporal (TTB) and spatial (STB) transformer blocks. The 3D decoder reconstructs skeletons using ψ, while the 2D decoder reconstructs original views with ψ_i and ϕ_j or generates mixed views by pairing random ψ_i and ϕ_j. The pose-aware (PAM) and view-aware (VAM) cross-attention modules refine features through dimensionality reduction, cross-attention, and residual connections.

3) Conversion of Joint Settings: The NTU and UAV-Human datasets have distinct keypoint structures, making direct compatibility challenging. The UAV-Human dataset follows the COCO-Keypoint protocol [62], [63], manually labeling 17 major body joints, whereas the NTU dataset [55], [56] adopts a different keypoint configuration. These differences require careful adaptation to ensure effective fine-tuning across datasets. To address this, we converted the UAV-Human dataset's 17 joint annotations to align with the NTU keypoint protocol. As illustrated in Figure 4, gray joints are common to both protocols and can be directly mapped, ensuring consistency in the conversion process. However, orange joints highlight structural differences, requiring additional processing. To bridge these gaps, we inferred the

TABLE I
OVERVIEW OF THE DETAILED ARCHITECTURES OF THE INDIVIDUAL MODULES IN OUR 2D^3-SKELACT MODEL

Name	Layer	Channels	Notes
Decoupler	STB	$2 \rightarrow 512$	input
	TTB	$512 \rightarrow 512$	8 heads
	STB	$512 \rightarrow 512$	8 heads
	TTB	$512 \rightarrow 512$	8 heads
3D Decoder	STB	$256 \rightarrow 256$	ψ
	TTB	$256 \rightarrow 256$	8 heads
	STB	$256 \rightarrow 256$	8 heads
	TTB	$256 \rightarrow 3$	$\hat{X}'$
2D Decoder	STB	$512 \rightarrow 256$	ψ_i & ϕ_j
	TTB	$256 \rightarrow 256$	8 heads
	STB	$256 \rightarrow 256$	8 heads
	TTB	$256 \rightarrow 2$	$\hat{X}_{i,j}$
	STB	$512 \rightarrow 256$	$[\psi_i, \phi_i]$
	TTB	$256 \rightarrow 256$	8 heads
	STB	$256 \rightarrow 256$	8 heads
	TTB	$256 \rightarrow 2$	$\hat{X}$
PAM ‖ VAM	Linear	$256 \rightarrow 128$	-
	Cross-Attention	$128 \rightarrow 128$	joint
	Residual Block	$128 \rightarrow 256$	output
	Linear	$256 \rightarrow 128$	-
	Cross-Attention	$128 \rightarrow 128$	temporal
	Residual Block	$128 \rightarrow 256$	output

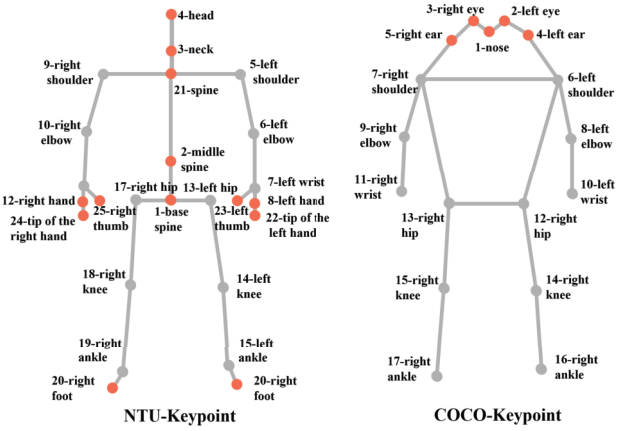

Fig. 4. The joint sets and graph structures of NTU keypoint and COCO keypoint.

missing NTU keypoints using existing COCO keypoints following [64]. In IV-F.1, we initially extract joint points from wild web videos using a pose estimation method, formatted according to the COCO keypoint standard. Subsequently, these joint points are converted via the previously described joint protocol conversion method, enabling category prediction by the pre-trained model.

D. Qualitative Results

1) Comparison on Challenging Multi-View Recognition: We evaluate 2D^3-SkelAct across multiple model architectures to assess its effectiveness in challenging multi-view settings (IV-B). To ensure a fair comparison, we maintain consistent training conditions, including the number of epochs and backbone architecture, and use official implementation codes whenever possible. As shown in Table II, 2D^3-SkelAct significantly improves skeleton action recognition on NTU-60, NTU-120, and PKU-MMD when integrated with baseline models. On NTU-120, the CNN-based HCN model gains an average 5.4% improvement across two benchmarks, while the attention-based DSTA-Net and GCN-based HDGCN models improve by 0.65% and 4.6%, respectively. When applied to BlockGCN, 2D^3-SkelAct achieves state-of-the-art performance across all three datasets. Even on PKU-II, a dataset with extreme viewpoints and noise, 2D^3-SkelAct delivers substantial accuracy improvements, demonstrating its robustness in real-world scenarios.

2) Comparison on Balanced Recognition: We test the effectiveness of our method 2D^3-SkelAct on the (balanced) original NTU datasets by integrating the BlockGCN [61] as the baseline model. We compare 2D^3-SkelAct with state-of-the-arts. As shown in Table IV, following the previous works, the results for multi-stream fusion are reported, including joint and bone (2-stream) as well as joint motion and bone motion (4-stream). Notably, our method employs solely 2D skeletons as input yet consistently demonstrates best performance across multiple datasets, outperforming previous approaches that utilize 3D skeletons. Specifically, 2D^3-SkelAct with 4-stream outperforms the latest methods with 6-stream (or 4-stream) configurations, demonstrating the validity of the 2D^3-SkelAct. On the Northwestern-UCLA dataset, our method 2D^3-SkelAct achieves the best performance by integrating the HDGCN [26] as the baseline model.

3) Comparison to 3D Skeletons: Compared to employing baseline models directly on multi-view 3D skeletons, our 2D^3-SkelAct demonstrates superior performance. Table V illustrates that when integrated with the baselines, 2D^3-SkelAct significantly outperforms the baselines utilizing the estimated 3D skeleton as input, often achieving results comparable to or even surpassing those using ground-truth 3D skeletons as input. Given the prevalence of 2D skeleton data in practical applications, the adoption of 2D^3-SkelAct reduces the computational overhead associated with converting 2D to 3D and mitigates additional noise introduced by explicit 3D recovery.

4) Comparison of Computation Cost: We compared the computational cost and accuracy of our proposed 2D^3-SkelAct against baseline methods (STGCN, CTRGCN) on challenging multi-view. Results indicate our single-stream model achieves higher accuracy with fewer parameters and less computational cost compared to traditional 4-stream models. Although the decoupler adds computational overhead during training, it does not impact inference efficiency, confirming our model's effectiveness for practical applications.

E. Ablation Study

1) Network Architectures: We examined the embedding of PAM and VAM within the model. Table VII indicates the best performance when integrated into the third and fourth blocks, where higher-level features are processed, optimizing the model's ability to capture pose depth and view-independence. We also evaluated the core components of 2D^3-SkelAct using STGCN as the baseline. As shown in Table VIII, PAM, VAM, and KD significantly boosted performance, with PAM

TABLE II

PERFORMANCE COMPARISON OF CHALLENGING MULTI-VIEW 2D SKELETON-BASED ACTION RECOGNITION WITH SINGLE JOINT STREAM AND SINGLE PERSON (SEE IV-B)

Method	NTU-60		NTU-120		PKU-MMD	
	X-sub	X-view	X-sub	X-set	Phase I	Phase II
HCN [57]	58.0	61.8	47.9	47.8	74.8	20.3
HCN (Ours)	$61.3^{\uparrow 3.3}$	$65.4^{\uparrow 3.6}$	$53.2^{\uparrow 5.3}$	$53.3^{\uparrow 5.5}$	$78.2^{\uparrow 3.4}$	$26.7^{\uparrow 6.4}$
STGCN [23]	74.3	80.6	66.0	66.6	86.8	49.3
STGCN (Ours)	$79.4^{\uparrow 5.1}$	$86.7^{\uparrow 6.1}$	$71.3^{\uparrow 5.3}$	$73.0^{\uparrow 6.4}$	$90.4^{\uparrow 3.6}$	$56.7^{\uparrow 7.4}$
AGCN [24]	76.3	82.8	66.7	67.3	88.0	53.9
AGCN (Ours)	$80.2^{\uparrow 3.9}$	$86.8^{\uparrow 4.0}$	$71.9^{\uparrow 5.2}$	$73.6^{\uparrow 6.3}$	$91.5^{\uparrow 3.5}$	$58.3^{\uparrow 4.4}$
DSTA-Net [58]	75.1	82.2	67.1	68.4	87.7	50.6
DSTA-Net (Ours)	$77.4^{\uparrow 2.3}$	$84.6^{\uparrow 2.4}$	$67.9^{\uparrow 0.8}$	$68.9^{\uparrow 0.5}$	$88.9^{\uparrow 1.2}$	$51.9^{\uparrow 1.3}$
CTRGCN [25]	79.0	84.7	71.6	72.9	89.8	54.9
CTRGCN (Ours)	$81.6^{\uparrow 2.6}$	$87.5^{\uparrow 2.8}$	$74.0^{\uparrow 2.4}$	$75.4^{\uparrow 2.5}$	$91.9^{\uparrow 2.1}$	$59.7^{\uparrow 4.8}$
MSG3D [59]	79.2	84.7	70.7	71.3	89.4	54.2
MSG3D (Ours)	$81.2^{\uparrow 2.0}$	$87.0^{\uparrow 2.3}$	$73.1^{\uparrow 2.4}$	$74.4^{\uparrow 3.1}$	$91.1^{\uparrow 1.7}$	$59.7^{\uparrow 5.5}$
HDGCN [26]	77.7	83.8	68.8	70.6	89.7	54.8
HDGCN (Ours)	$81.7^{\uparrow 4.0}$	$85.9^{\uparrow 2.2}$	$73.1^{\uparrow 4.3}$	$75.5^{\uparrow 4.9}$	$92.5^{\uparrow 2.8}$	$58.1^{\uparrow 3.3}$
DS-GCN [60]	80.1	84.9	71.9	73.1	89.9	55.2
DS-GCN (Ours)	$82.0^{\uparrow 1.9}$	$86.1^{\uparrow 2.3}$	$74.1^{\uparrow 2.2}$	$75.9^{\uparrow 2.8}$	$2.5^{\uparrow 2.6}$	$58.3^{\uparrow 3.1}$
BlockGCN [61]	80.4	85.3	72.4	73.6	90.2	55.7
BlockGCN (Ours)	$\mathbf{83.7}^{\uparrow \mathbf{3.4}}$	$\mathbf{87.7}^{\uparrow \mathbf{2.5}}$	$\mathbf{74.9}^{\uparrow \mathbf{2.5}}$	$\mathbf{77.4}^{\uparrow \mathbf{3.8}}$	$\mathbf{92.8}^{\uparrow \mathbf{2.6}}$	$\mathbf{60.0}^{\uparrow \mathbf{4.3}}$

TABLE III

COMPARISON OF PERFORMANCE ON VANILLA NTU-60 AND NTU-120 DATASETS. ∗S- MEANS THE FUSION RESULTS OF ∗ STREAMS. NOTABLY, OUR 2D^3-SKELACT ONLY TAKES 2D SKELETONS AS INPUT DURING INFERENCE, WHILE OTHER METHODS USE 3D SKELETONS

Method	Year	NTU 60 (%)		NTU 120 (%)	
		xsub	xview	xsub	xset
2s-SGN [65]	CVPR'20	89.0	94.5	79.2	81.5
4s-Shift-GCN [66]	CVPR'20	90.7	96.5	85.9	87.6
2s-MSG3D [59]	CVPR'20	91.5	96.2	86.9	88.4
4s-SEFN [67]	TCSVT'21	90.7	96.4	86.2	87.8
4s-MSTGCN [68]	AAAI'21	91.5	96.6	87.5	88.8
4s-CTRGCN [25]	ICCV'21	92.4	96.8	88.9	90.6
4s-InfoGCN [69]	CVPR'22	92.7	96.9	89.4	90.7
6s-InfoGCN [69]	CVPR'22	93.0	97.1	89.8	91.2
3s-EfficientGCN [70]	TPAMI'22	91.7	95.7	88.3	89.1
4s-FR-Head [71]	CVPR'23	92.8	96.8	89.5	90.9
6s-StreamGCN [72]	IJCAI'23	92.9	96.9	89.7	91.0
4s-HDGCN [26]	ICCV'23	93.0	97.0	89.8	91.2
4s-SiT-MLP [73]	TCSVT'24	92.3	96.8	89.0	90.5
4s-SelfGCN [74]	TIP'24	93.1	96.6	89.4	91.0
2s-DKE-GCN [30]	TCSVT'24	93.1	97.1	89.9	91.4
3s-HA-GNN [75]	TMM'24	93.4	97.2	89.9	91.5
4s-DS-GCN [60]	AAAI'24	93.1	97.5	89.2	91.1
4s-BlockGNN [61]	'24	93.1	97.0	90.3	91.5
4s-Ours	-	**93.9**	**97.5**	**90.8**	**92.0**

TABLE IV

COMPARISON OF PERFORMANCE ON NORTHWESTERN-UCLA

Method	Year	Accuracy(%)
AGC-LSTM [76]	CVPR'19	93.3
SGN [65]	CVPR'20	92.5
Shift-GCN [66]	CVPR'20	94.6
DC-GCN+ADG [77]	ECCV'20	95.3
CTRGCN [25]	ICCV'21	96.5
4s-InfoGCN [69]	CVPR'22	96.6
6s-InfoGCN [69]	CVPR'22	97.0
4s-HDGCN [26]	ICCV'23	96.9
4s-Ours	-	**97.2**

improving by 0.4% and 0.9%, VAM by 2.6% and 3.4%, and KD by 3.0% and 3.7%. The combined impact of these components enhances the model's robustness, view-independence, and depth of learning.

2) Camera Viewpoints: We examine the impact of varying projection viewpoints (α, β, and γ) on the performance of 2D^3-SkelAct. By uniformly adjusting these angles across different ranges (see IV-B), we find that larger angle ranges improve model performance, as shown in Table IX. Specifically, setting the range to [−90°, 90°], which covers all possible angles, results in the highest performance. This suggests that a broad range of viewpoints enhances pose-view decoupling and strengthens the 2D-to-3D supervision strategy.

3) Comparison to 3D Skeletons: Compared to employing baseline models directly on multi-view 3D skeletons, our

TABLE V

CLASSIFICATION ACCURACY COMPARISON AGAINST BASELINE METHODS ON THE NTU-60, NTU-120 DATASETS BASED ON DATA EXPERIMENT IV-B. OUR 2D^3-SKELACT ONLY TAKES 2D SKELETON SEQUENCES AS INPUT. "3D" IS THE GROUND-TRUTH 3D SKELETON SEQUENCES. "3D*" IS THE ESTIMATED 3D SKELETON SEQUENCES THROUGH 3D DECODER $\mathcal{D}'(\cdot)$

Method	NTU-60		NTU-120	
	X-view	X-sub	X-set	X-sub
STGCN (3D*)	75.5	70.6	64.1	62.4
STGCN (3D)	78.2	73.0	63.3	63.5
STGCN (Ours)	86.7	79.4	73.0	71.3
CTRGCN (3D*)	82.2	75.4	68.0	67.1
CTRGCN (3D)	87.8	82.2	76.8	75.5
CTRGCN (Ours)	87.5	81.6	75.4	74.0
MSG3D (3D*)	82.0	74.8	67.0	66.8
MSG3D (3D)	85.8	79.2	72.4	71.6
MSG3D (Ours)	87.0	81.2	74.4	73.1
HD-GCN (3D*)	81.9	75.4	68.2	67.8
HD-GCN (3D)	86.3	81.8	75.7	74.1
HD-GCN (Ours)	85.9	81.7	75.5	73.1

TABLE VI

COMPARISON OF COMPUTATION COST AND PERFORMANCE

Method	Accuracy(%)		Params	FLOPS
	xview	xsub		
STGCN	74.3	80.6	3.2M	2.7G
4s-STGCN	77.1	83.7	12.8M	10.8G
1s-STGCN(Ours)	79.4	86.7	10.1M	9.4G
CTRGCN	79.0	84.7	5.6M	2.9G
4s-CTRGCN	81.4	87.5	22.4M	11.6G
1s-CTRGCN(Ours)	81.6	87.5	12.7M	11.4G
Decoupler	-	-	12.1M	30.7G

TABLE VII

ABLATION STUDY FOR WHERE PAM AND VAM ARE ADDED IN THE MODEL BLOCK

Blocks	Accuracy(%)	
	xsub	xview
7,8	79.0	85.9
5,6	78.9	86.0
3,4	**79.4**	**86.7**

TABLE VIII

ABLATION STUDY FOR VERIFYING THE EFFECTIVENESS OF COMPONENTS OF 2D^3-SKELACT IN NTU-60

Module	Accuracy(%)	
	xsub	**xview**
STGCN	74.3	80.6
+ PAM	74.7	81.5
+ VAM	76.9	84.0
+ KD	77.3	84.3
STGCN (Ours)	**79.4**	**86.7**

2D^3-SkelAct demonstrates superior performance. Table V illustrates that when integrated with the baselines, 2D^3-SkelAct significantly outperforms the baselines utilizing the estimated 3D skeleton as input, often achieving results comparable to or even surpassing those using ground-truth 3D skeletons as input. Given the prevalence of 2D skeleton data in practical applications, the adoption of 2D^3-SkelAct reduces the computational overhead associated with converting 2D to 3D and mitigates additional noise introduced by explicit 3D recovery.

TABLE IX

ABLATION STUDY FOR THE EFFECT OF THE VIEWPOINT CHANGE RANGE

Viewpoints	Accuracy(%)	
	xview	xsub
[-30°, 30°]	85.4	79.2
[-60°, 60°]	85.8	79.3
[-90°, 90°]	**86.7**	**79.4**

4) Effectiveness of FD and LD: We performed an ablation study to evaluate the contributions of Feature Distillation (FD) and Logit Distillation (LD). The results demonstrate that removing either FD or LD individually leads to a noticeable decline in performance, confirming that both FD and LD substantially improve model accuracy. Combining FD and LD yields the best results, highlighting their complementary roles in enhancing action recognition.

5) Visualization of the Attention Maps: Here, we explore how the model's focus on different joints varies with changing viewpoints. By summing the rows of the inward adjacency matrix in the AGCN model [24], we assess joint attention. Figure 5 shows that the model prioritizes different joints depending on the action: "headache" highlights hands and head, "putting palms together" emphasizes hands and arms, "kicking" focuses on legs and hips, and "walking towards" draws attention to hands, legs, and arms. These patterns, consistent with visual analyses, demonstrate AGCN's localized focus when projecting 3D skeletons into 2D, particularly from diverse viewpoints. Additionally, 2D^3-SkelAct enhances this focus diversity, improving attention distribution and strengthening action recognition. These findings suggest that decoupled pose and view features, along with distillation in 2D^3-SkelAct, enhance the model's ability to identify critical joints, improving motion feature extraction and classification robustness.

6) Effectiveness of 2D^3-SkelAct for Specific Actions: Figure 7 presents an accuracy comparison across 60 action categories in the NTU-60 X-view before and after incorporating the STGCN into our 2D^3-SkelAct framework. The data illustrate that our method enhances accuracy for all action categories. Notably, for actions such as "punch," "kicking," and "pointing finger," there is a significant improvement in accuracy, exceeding 20%. These findings demonstrate that 2D^3-SkelAct can seamlessly integrate with various encoders, effectively aiding the skeleton feature extractor in learning the motion representations of skeleton actions. This is particularly beneficial in enhancing recognition accuracy from challenging viewpoints.

7) Visualization of Decoupler's Reconstruction: The reconstruction outcomes of the decoupler, as depicted in Figure 8, reveal that when integrating other viewpoints with reduced

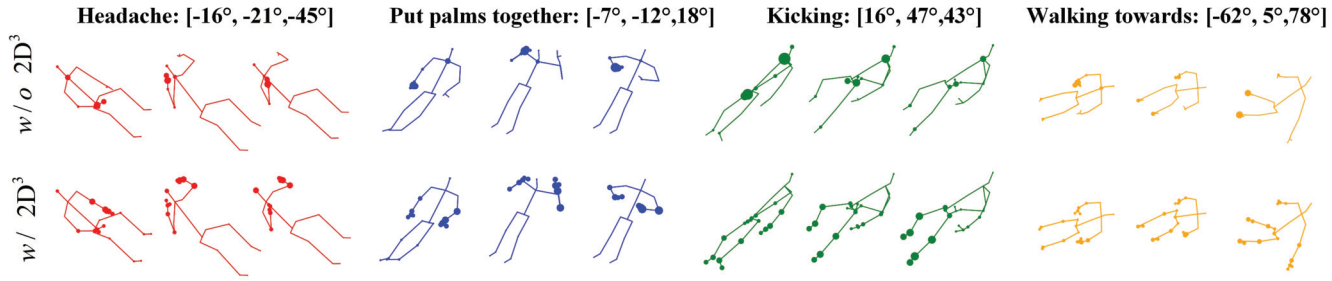

Fig. 5. Illustration of attention intensity in 2D^3-SkelAct. $w/\ 2D^3$ and $w/o\ 2D^3$ respectively mean the graph topology results of AGCN with and without applying 2D^3-SkelAct. Each circle's size denotes a joint corresponding to the attention intensity. The angle means the projection viewpoints [α, β, γ] in the rotation matrix.

Fig. 6. The action classification results of 2D^3-SkelAct on (a) wild web videos and (b) UAV-Human. Marked below each picture is the label and score of the top 3 action categories.

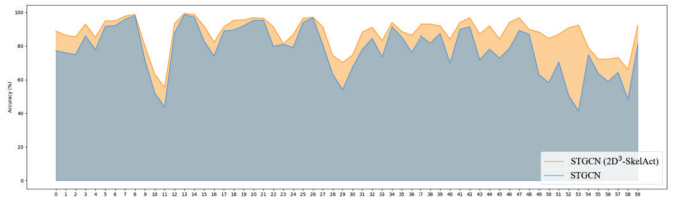

Fig. 7. Accuracy distribution of 2D^3-SkelAct on the NTU-60 X-view benchmark, using STGCN as the baseline model.

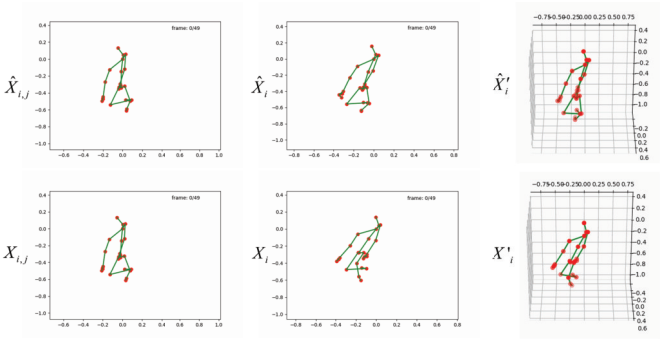

Fig. 8. Visualization of the reconstruction results performed by the decoupler in the 2D ^{3}-SkelAct model.

occlusion, the discrepancy between the synthesized 2D reconstructed skeleton and the actual ground truth is remarkably minimal. Both the 2D and 3D reconstructed skeletons, viewed from the original perspective, align closely with the ground truth. However, a minor deviation is noted at the position of the left-hand joint, which is likely attributable to joint occlusion occurring during movement. This successful reconstruction by the decoupler underscores its capability to effectively separate latent pose and view features.

F. Robustness and Generalizability

1) Finetuned Results on UAV-Human: We pretrain 2D^3-SkelAct on the NTU-120 X-set and then fine-tune 2D^3-SkelAct on the UAV-Human dataset, chosen specifically for its inclusion of aerial vehicle scenes, which pose viewpoint challenges for action recognition. We align the 17 joints of UAV-Human with the 25 joints of the NTU-120 dataset as described in Section IV-C. As shown in Table X, the results in CSv1 and CSv2 demonstrate that 2D^3-SkelAct significantly outperforms the fully supervised ST-GCN and CTRGCN in both benchmarks. This superior performance highlights the efficacy of 2D^3-SkelAct, particularly in challenging viewpoint-variant scenes.

TABLE X
FINETUNED RESULTS ON UAV-HUMAN

Method	Accuracy(%)	
	CSv1	CSv2
STGCN	25.5	43.3
Ours	**30.5**	**53.9**
CTRGCN	29.5	48.2
Ours	**32.5**	**56.2**

TABLE XI
ABLATION STUDY FOR THE EFFECT OF FD AND LD. "W/O" MEANS "WITHOUT"

Method	Accuracy(%)	
	xview	xsub
STGCN	74.3	80.6
STGCN(Ours) w/o FD	78.9	86.1
STGCN(Ours) w/o LD	78.1	85.3
STGCN(Ours)	**79.4**	**86.7**
CTRGCN	79.0	84.7
CTRGCN(Ours) w/o FD	80.9	86.8
CTRGCN(Ours) w/o FD	80.5	86.6
CTRGCN(Ours)	**81.6**	**87.5**

2) Real-World Applications: We analyzed our 2D^3-SkelAct on the wild web videos captured from extreme viewpoints and the UAV-Human videos. Using an off-the-shelf pose estimation model [78], we extracted skeletons and converted them into the NTU-60 joint protocol. We then employed the 2D^3-SkelAct pre-trained by X-view benchmark on NTU-60 with STGCN as the baseline model for action recognition. Figure 6 illustrates action videos presenting challenges like occlusion, extreme viewpoints, and lack of depth, complicating recognition. Despite these hurdles, our model accurately categorizes actions, showcasing the effectiveness of the 2D^3-SkelAct approach in extracting motion cues from 3D skeletons and navigating complexities in challenging scenarios.

V. CONCLUSION

In this work, we propose 2D^3-SkelAct, a novel model for 2D skeleton action recognition that overcomes the challenges associated with the absence of view-independence and depth in 2D data by utilizing the strength of 3D latent space. 2D^3-SkelAct incorporates a pose-view feature decoupler, pre-trained to construct the 3D latent space and decouple pose and view features from input 2D skeleton sequences. Subsequently, pose-aware and view-aware cross-attention modules are presented to facilitate 2D motion feature extraction by fusing latent pose and view features, respectively. Meanwhile, our pipeline distills comprehensive motion clues from the 3D model to enrich motion features. We verify the performance of our 2D^3-SkelAct through extensive experiments, showing its compatibility with various model architectures and consistent improvement in action recognition.

Limitations: Human body occlusion, which results in inaccurate pose estimation, limits the model's action recognition accuracy. Therefore, addressing the challenge of precise pose estimation amidst occlusions will be a priority for enhancing model performance.

ACKNOWLEDGMENT

The numerical calculation was supported by the supercomputing system in the Super-Computing Center of Wuhan University.

REFERENCES

[1] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2008, pp. 2737–2740.
[2] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.
[3] I. Rodomagoulakis et al., "Multimodal human action recognition in assistive human–robot interaction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2702–2706.
[4] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "StagNet: An attentive semantic RNN for group activity recognition," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 101–117.
[5] P. Elias, J. Sedmidubsky, and P. Zezula, "Understanding the limits of 2D skeletons for action recognition," *Multimedia Syst.*, vol. 27, no. 3, pp. 547–561, Jun. 2021.
[6] B. X. Nie, P. Wei, and S.-C. Zhu, "Monocular 3D human pose estimation by predicting depth on joints," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3467–3475.
[7] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, "Monocular 3D human pose estimation by generation and ordinal ranking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2325–2334.
[8] D. Mehta et al., "Single-shot multi-person 3D pose estimation from monocular RGB," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 120–130.
[9] C. Zheng et al., "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–37, 2023.
[10] M. Cormier, Y. Schmid, and J. Beyerer, "Enhancing skeleton-based action recognition in real-world scenarios through realistic data augmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 300–309.
[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
[12] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3010–3022, Jul. 2016.
[13] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13232–13242.
[14] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.
[15] R. Vemulapalli and R. Chellappa, "Rolling rotations for recognizing human actions from 3D skeletal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4471–4479.
[16] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.
[17] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.
[18] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.
[19] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.

[20] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.

[21] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[22] Q. Nie, J. Wang, X. Wang, and Y. Liu, "View-invariant human action recognition based on a 3D bio-constrained skeleton model," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3959–3972, Aug. 2019.

[23] M. Jiang, J. Dong, D. Ma, J. Sun, J. He, and L. Lang, "Inception spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Int. Symp. Control Eng. Robot. (ISCER)*, Feb. 2022, pp. 208–213.

[24] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[25] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 13359–13368.

[26] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 10444–10453.

[27] X. Hao, J. Li, Y. Guo, T. Jiang, and M. Yu, "Hypergraph neural network for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2263–2275, 2021.

[28] Z. Tu, J. Zhang, H. Li, Y. Chen, and J. Yuan, "Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 1819–1831, 2022.

[29] Q. Pan, Z. Zhao, X. Xie, J. Li, Y. Cao, and G. Shi, "View-normalized and subject-independent skeleton generation for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7398–7412, Dec. 2022.

[30] Y. Liu, Y. Li, H. Zhang, X. Zhang, and D. Xu, "Decoupled knowledge embedded graph convolutional network for skeleton-based human action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9445–9457, Oct. 2024.

[31] H. Liu, Y. Liu, Y. Chen, C. Yuan, B. Li, and W. Hu, "TranSkeleton: Hierarchical spatial–temporal transformer for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4137–4148, Aug. 2023.

[32] H. Zheng and X. Zhang, "A cross view learning approach for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3061–3072, May 2022.

[33] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition," *Inf. Sci.*, vol. 569, pp. 90–109, Aug. 2021.

[34] F. M. Thoker, H. Doughty, and C. G. M. Snoek, "Skeleton-contrastive 3D action representation learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1655–1663.

[35] L. Lin, S. Song, W. Yang, and J. Liu, "MS2L: Multi-task self-supervised learning for skeleton based action recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2490–2498.

[36] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3D human action representation learning via cross-view consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4741–4750.

[37] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding, "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 762–770.

[38] J. Zhang, Z. Tu, J. Weng, J. Yuan, and B. Du, "A modular neural motion retargeting system decoupling skeleton and shape perception," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6889–6904, Oct. 2024.

[39] J. Zhang et al., "TapMo: Shape-aware motion generation of skeleton-free characters," 2023, *arXiv:2310.12678*.

[40] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[41] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 4133–4141.

[42] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. ICLR*, Jan. 2016, pp. 1–12.

[43] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2859–2868.

[44] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11953–11962.

[45] Q. Guo et al., "Online knowledge distillation via collaborative learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11020–11029.

[46] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3430–3437.

[47] N. C. Garcia and P. M. V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 103–118.

[48] J. Xiao et al., "Learning from temporal gradient for semi-supervised action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3252–3262.

[49] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang, "Optical flow guided feature: A fast and robust motion representation for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1390–1399.

[50] J. Stroud, D. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3D: Distilled 3D networks for video action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Mar. 2020, pp. 625–634.

[51] P. Liu, I. King, M. R. Lyu, and J. Xu, "DDFlow: Learning optical flow with unlabeled data distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8770–8777.

[52] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*.

[53] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16266–16275.

[54] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2649–2656.

[55] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[56] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[57] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," 2018, *arXiv:1804.06055*.

[58] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Decoupled spatial–temporal attention network for skeleton-based action-gesture recognition," in *Proc. Asian Conf. Comput. Vis.*, Jan. 2021, pp. 38–53.

[59] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 143–152.

[60] J. Xie, Y. Meng, Y. Zhao, A. Nguyen, X. Yang, and Y. Zheng, "Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 6, pp. 6225–6233.

[61] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "BlockGCN: Redefine topology awareness for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2049–2058.

[62] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[63] T. V. Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 601–617.

[64] H. Choi, G. Moon, and K. M. Lee, "Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Berlin, Germany: Springer, Aug. 2020, pp. 769–787.

[65] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1112–1121.
[66] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 183–192.
[67] J. Kong, H. Deng, and M. Jiang, "Symmetrical enhanced fusion network for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4394–4408, Nov. 2021.
[68] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1113–1122.
[69] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, "InfoGCN: Representation learning for human skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20154–20164.
[70] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1474–1488, Feb. 2023.
[71] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 10608–10617.
[72] Y. Yang et al., "Action recognition with multi-stream motion modeling and mutual information maximization," 2023, *arXiv:2306.07576*.
[73] S. Zhang, J. Yin, Y. Dang, and J. Fu, "SiT-MLP: A simple MLP with point-wise topology feature learning for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8122–8134, Sep. 2024.
[74] Z. Wu et al., "SelfGCN: Graph convolution network with self-attention for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 4391–4403, 2024.
[75] P. Geng, X. Lu, W. Li, and L. Lyu, "Hierarchical aggregated graph neural network for skeleton-based action recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 11003–11017, 2024.
[76] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.
[77] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling GCN with dropgraph module for skeleton-based action recognition," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 536–553.
[78] T. Jiang et al., "RTMPose: Real-time multi-person pose estimation based on MMPose," 2023, *arXiv:2303.07399*.

Xiangyue Zhang (Student Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2023. He is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China. His research interests include computer vision, action recognition, and motion generation.

Yifan Jia received the Ph.D. degree from Huazhong Agricultural University. He is currently the Deputy Director and the Party Branch Secretary of the Pain Department, Renmin Hospital of Wuhan University. He also holds the positions of Associate Professor, Associate Chief Physician, and Master's Supervisor. He has published more than 30 academic journal articles in international and domestic professional journals. He specializes in the minimally invasive treatment of refractory pain, rhinitis, and cancer pain. He has profound expertise in neuromodulation therapy for neuropathic pain. He is also the Chair of the Pain Branch of Hubei Microcirculation Society. He serves as an Expert Reviewer for key pain specialties in Hubei Province and a reviewer for the Natural Science Foundation.

Jiaxu Zhang (Student Member, IEEE) received the B.S. degree from Southeast University, Nanjing, China, in 2020. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China. His research interests include computer vision and computer graphics.

Yijie Yang received the B.S. degree from Northeastern University, Shenyang, China, in 2022. He is currently pursuing the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, China. His research interests include action recognition, 3D human understanding, and computer vision.

Zhigang Tu (Senior Member, IEEE) received the Ph.D. degree from Wuhan University, China, in 2013, and the Ph.D. degree from Utrecht University, The Netherlands, in 2015.

From 2015 to 2016, he was a Post-Doctoral Researcher with Arizona State University, USA. From 2016 to 2018, he was a Research Fellow with Nanyang Technological University, Singapore. He is currently a Professor with Wuhan University and has co-authored more than 70 papers in international SCI-indexed journals and conferences. His research interests include computer vision, image processing, and video analytics, with focusing on motion estimation/retargeting, human behavior (action, pose, gesture) recognition, reconstruction, and generation. He received the Best Student Paper Award at the Fourth Asian Conference on Artificial Intelligence Technology and one of the three best reviewers award for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY in 2022. He is the first organizer of the ACCV2020 Workshop on MMHAU, Japan. He is the Area Chair of AAAI2023/2024/2025 and an Associate Editor of the SCI-indexed journals of *The Visual Computer* (IF = 3.5), the *Journal of Visual Communication and Image Representation* (IF = 2.6), and the *CAAI Transactions on Intelligence Technology* (IF = 8.4).