Skeleton MixFormer: Multivariate Topology Representation for Skeleton-based Action Recognition

Wentian Xin
Xidian University
Xi'an, China
wtxin@stu.xidian.edu.cn

Qiguang Miao*
Xidian University
Xi'an, China
qgmiao@xidian.edu.cn

Yi Liu
Xidian University
Xi'an, China
ly330@stu.xidian.edu.cn

Ruyi Liu
Xidian University
Xi'an, China
ruyiliu@xidian.edu.cn

Chi-Man Pun
University of Macau
Macau, China
cmpun@umac.mo

Cheng Shi
Xi'an University of Technology
Xi'an, China
C_shi@xaut.edu.cn

ABSTRACT

Vision Transformer, which performs well in various vision tasks, encounters a bottleneck in skeleton-based action recognition and falls short of advanced GCN-based methods. The root cause is that the current skeleton transformer depends on the self-attention mechanism of the complete channel of the global joint, ignoring the highly discriminative differential correlation within the channel, so it is challenging to learn the expression of the multivariate topology dynamically. To tackle this, we present Skeleton MixFormer, an innovative spatio-temporal architecture to effectively represent the physical correlations and temporal interactivity of the compact skeleton data. Two essential components make up the proposed framework: 1) Spatial MixFormer. The channel-grouping and mix-attention are utilized to calculate the dynamic multivariate topological relationships. Compared with the full-channel self-attention method, Spatial MixFormer better highlights the channel groups' discriminative differences and the joint adjacency's interpretable learning. 2) Temporal MixFormer, which consists of Multiscale Convolution, Temporal Transformer and Sequential Holding Module. The multivariate temporal models ensure the richness of global difference expression and realize the discrimination of crucial intervals in the sequence, thereby enabling more effective learning of long and short-term dependencies in actions. Our Skeleton MixFormer demonstrates state-of-the-art (SOTA) performance across seven different settings on four standard datasets, namely NTU-60, NTU-120, NW-UCLA, and UAV-Human. *Related code will be available on Skeleton-MixFormer.*

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

KEYWORDS

video understanding, skeleton action recognition, topology representation, transformer, attention

ACM Reference Format:
Wentian Xin, Qiguang Miao, Yi Liu, Ruyi Liu, Chi-Man Pun, and Cheng Shi. 2023. Skeleton MixFormer: Multivariate Topology Representation for Skeleton-based Action Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611900

1 INTRODUCTION

Human action recognition is a fundamental and significant topic in the field of computer multimedia processing, which provides reliable human-centric action analysis results for automatic driving [19], video surveillance [17], human-computer interaction [35], and end-to-end system [8]. Influenced by multimodal tasks in computer vision, depth and 3D information other than RGB is widely adopted, processed, and fused [32]. In recent years, skeleton-based action recognition has gained great attention and development [12, 38]. Compact skeleton data provide detailed position and motion information of human joints, which facilitates the construction of spatio-temporal motion and pay more attention to the essential characteristics of the action [18, 43].

As Transformer [10, 34] has gradually taken the lead in the performance and efficiency of image and natural language processing, researchers have naturally begun to replace the classical ST-GCN [39] structure with its modules. STTR [27] is the first to apply transformer to skeleton-based spatio-temporal action recognition. It is worth noting that despite the powerful global information abstraction and processing capabilities of transformer-based networks, they have not yet surpassed the accuracy achieved by many outstanding GCN-based works. We believe that there are four main reasons. *Firstly*, the self-attention mechanism, which is at the core of the transformer, has already been incorporated into some GCN-based networks. Additionally, Positional Encoding (PE) performs a similar function to the adjacency matrix of GCN, as it can be seen as equivalent to the positional relationship of patches in Vision Transformer (ViT) [10]. *Secondly*, the GCN-based methods often apply secondary processing to the input of self-attention (equivalent to *Query* and *Key*) to extract spatial characteristics of skeleton data more effectively. In contrast, the transformer relies more on

*Corresponding Author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611900

itself and the global channel information association, which may limit its ability to extract adjacency relations with unity for specific actions with large intra-class differences. *Thirdly*, the design of the adjacency matrix is the core of skeleton action recognition task. GCN-based methods commonly employ the strategy of stacking or sharing adjacency matrix heads to enhance the ability of acquiring multi-level discriminant information. In contrast, transformer-based methods directly utilize the multi-channel adjacency matrix obtained by themselves, often leading to model overfitting. *Fourthly*, the transformer also faces challenges in distinguishing and learning the critical intervals on different action time series robustly, unlike CNNs, due to its reliance on global processing on time series. The standard transformer architecture lacks a keyframe extraction module, which makes it challenging to capture short-term temporal correlation properties and can result in performance degradation.

In order to overcome the issues mentioned above, we have implemented two significant enhancements and some resourceful tricks, which allow for the transformer network to more effectively utilize its global information learning capabilities and surpass the current limitations of existing recognition methods:

1) The first improvement is to present a channel-grouping and mix-attention technique called Spatial MixFormer. The proposed approach enables the creation of a multivariate topological relationship matrix, which comprehensively demonstrates the dynamic differences among adjacent associations. During the final stage of spatial feature learning, a Channel Reforming module is employed to facilitate the exchange of information among joint channel features, thereby mitigating the differential noise generated during grouping learning. This in turn allows the multivariate topological relation matrix constructed by each Spatial MixFormer to have universal applicability.

2) The second improvement is to present a Temporal MixFormer structure that combines Multiscale Convolution, Temporal Transformer, and Sequential Holding Module. The Multiscale Convolution employs a bottleneck design scheme, utilizing varying expansion rates to flexibly facilitate the learning of multiscale global universal features in the temporal domain. The Temporal Transformer has a structure similar to the Spatial MixFormer but applies dimensionality to time. Furthermore, the number of temporal relation matrices is reduced to maintain the temporal channel's feature continuity, ensuring the differentiation of global temporal features while simultaneously realizing the long-term dependence learning of the entire action. The Sequential Holding Module adopts an improved *Query & Key* input self-attention strategy, enabling the identification of essential frame sets in the short time series. By fusing the above three models, a comprehensive and effective update is achieved for the time series of skeleton actions.

3) Leveraging the residual and pyramid structures, we design a skeleton Retrospect Module that can extract spatio-temporal critical features of the shallow layer twice and concatenate them before the classification layer. This enhances structural differentiation and improves the overall discrimination ability.

Our main contributions can be summarized as follows:

• We propose a novel Skeleton MixFormer for action recognition. The model is more flexible for building multivariate spatio-temporal representation by relying on the intrinsic correlation of channels to maximize the utilization of highly distinguishable features and optimize the transformer's dependence on global information.

• Spatial MixFormer excavates the discriminative differential association between its own channel groups, realizes the dynamic learning of multivariate topology expression through mix-attention, and enriches the interpretability of skeleton adjacency relations.

• Temporal MixFormer integrates Multiscale Convolution, Temporal Transformer, and Sequential Holding Module to ensure the difference of global temporal features and the learning of long-term and short-term dependence, providing an orderly and effective update for the action sequence.

• On four standard datasets (NTU-60, NTU-120, NW-UCLA, UAV-Human), our Skeleton MixFormer achieves the highest performance both on GCN-based and Transformer-based. Sufficient ablation experiments are demonstrated, providing interpretability and reproducibility for the proposed architecture.

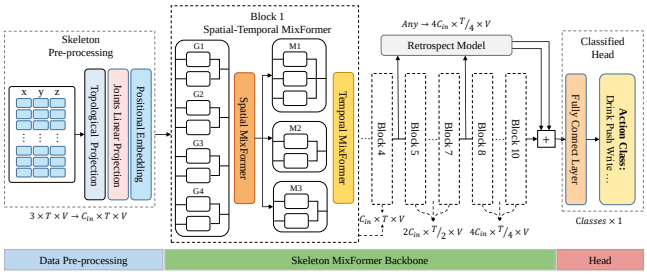


Figure 1: Architecture Overview. The network comprises three main modules: Data Pre-processing, Skeleton MixFormer Backbone, and Classified Head. Skeleton MixFormer Backbone consists of 10 blocks, each containing a Spatial MixFormer and a Temporal MixFormer. Topological Projection is utilized to determine the stream regime.

2 RELATED WORK

2.1 Skeleton Transformer

Transformers possess an inherent advantage in acquiring and processing global information, which is crucial for enhancing the classification ability of skeleton data. *Firstly*, various studies have delved into optimizing self-attention mechanisms in transformer for action recognition. For instance, [40] proposed the UNIK, which utilizes a multi-head attention mechanism to learn an optimal dependency matrix from a uniform distribution. [25] constructed a fully self-attention architecture that leverages spatial or temporal self-attention to replace GCN or temporal convolution in the two-stream network, automatically discovering hidden correlation information relevant to the current action. *Secondly*, some works have focused on enhancing temporal-spatial correlation, a central aspect of action recognition. KA-AGTN [23] was proposed to learn spatio-temporal patterns between joints accurately. TranSkeleton [21] unified spatial and temporal modeling within the transformer via different-aware temporal aggregation and physical connectivity constraints. Bai *et al.* [2] proposed the HGCT, which improves the spatio-temporal feature representation of entanglement. *Thirdly*, some methods have explored the human body's local and global scale correlations. IG-Former [26] developed a distation-based graph that measures

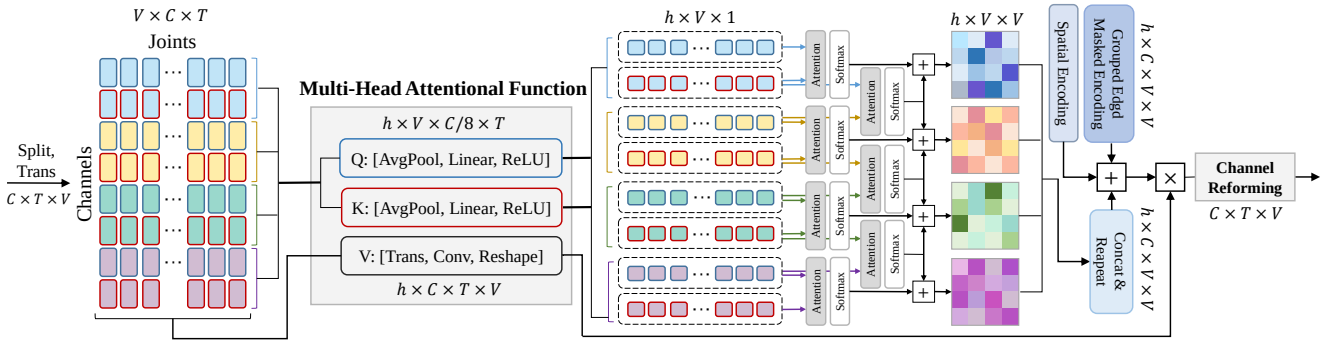


Figure 2: The diagram of Spatial MixFormer. We first perform transpose and split on the input data. Then, Q, K, and V are computed via the Multi-Head Attentional Function. The weight correlation matrix is obtained through mix-attention (self-&cross-attention). To enrich the expression of the model, two graph learning tricks are utilized. Finally, the channel Reforming module is used to eliminate the feature separation of channel grouping.

the distance between body parts to capture the distance information between interacting parts. FG-STFormer [13] was designed to capture relationships between key local joints and global context information in both spatial and temporal dimensions. STST [45] used diverse joint organization strategies to model the skeleton sequence spatio-temporally. *Lastly*, several methods have incorporated multimodal information to enhance expressiveness. For example, [16] introduced a relative transformation mechanism to learn long-distance dependencies through multiscale dynamic representation that fuses multiscale skeleton features. Ahn *et al.* [1] developed a spatio-temporal cross-transformer, comprising an encoder and a decoder, to learn feature representations for cross-modal data.

Nevertheless, the methods above rely on complete channel information modeling, disregarding the unique information differences between channel groups and resulting in an adjacency matrix lacking intrinsic discrimination support. In contrast, our approach involves the utilization of channel-grouping and mix-attention during skeleton correlation learning, in addition to the incorporation of a diverse range of global and local feature extraction techniques during the time series update, facilitating the resolution of various complex action classification problems. Our proposed Skeleton MixFormer model enhances both interpretability and adaptability, and maximizes the model's spatio-temporal discrimination ability.

2.2 Mixer and MixFormer

Evidently, existing transformer-based skeleton action recognition methods still rely on the simplistic self-attention mechanism. Our model draws inspiration from Mlp-mixer [33] and MixFormer [7], particularly in channel mixing, feature extraction, and interactivity representation. Through Mlp-mixer [33], it has been demonstrated that the self-attention layer in ViT can cause some learned function properties to be incompatible with the true underlying distribution. As a result, an excellent dimensional information interaction can only be achieved using channel transpose and MLP structures, which highlight the potential of channel features. MixFormer [7] explores the concept of mixing key and value for template matching, which leverages asymmetric information through cross-attention. With these theoretical foundations, our approach transfers and customizes these techniques to fully harness the potential expression capabilities of each node's channel features, using the small amount and compactness of skeleton data.

We named the model Skeleton MixFormer as it leverages both mixing and transformer techniques for the spatio-temporal layout, which significantly diverges from the above two methods in both implementation processes and ultimate objectives. In particular, our mix-attention structure refers explicitly to grouping channels and crossing computation within groups, which differs from the cross-information in [7].

3 METHODOLOGY

3.1 Preliminaries

The definition of the skeleton in the GCN is consistent with that in the transformer. Given the body joint sequence in 2D or 3D coordinates, the skeleton of the human body can be denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = (v_1, v_2, ..., v_N)$ represents the joint set of N vertices, $\mathcal{E}$ represents the bone set of the edges. In the adjacency matrix $A \in \mathbb{R}^{N \times N}$ (assuming $\mathcal{G}$ is an undirected graph), if the V_i and V_j have a skeleton directly connected, $A_{i,j} = 1$, otherwise, $A_{i,j} = 0$. If the skeleton sequence is represented by X and A, the layer-wise iteration and weights update can be formulated as $X^{l+1} = \sigma(AX^lW^l)$, where $W^l \in \mathbb{R}^{C_l \times C_{l+1}}$ represents the learnable matrix of the network at layer l. To make the adjacency matrix adaptive, some methods [3, 29] convolve the input and use the self-attention method to obtain the adaptive relevance adjacency matrix, as $\mathcal{X}^{l+1} = \sigma(f(\mathcal{X}^l)^{\mathrm{T}} M^{\mathrm{T}} M f(\mathcal{X}^l))$, where M and $f(\cdot)$ represent learnable matrix and mapping operations, respectively.

The interpretability of the relevance matrix in the GNN corresponds with that of the weight relationship matrix in the transformer. Due to their homology, we surmise that tricks might exhibit universality. Consequently, we could adopt both the Fully Learnable Relative Position Embedding (FL-RPE) [9] in transformer, and the Grouped Edge Masked Encoding (G-EME) [4, 42] in GNN to further improve the model performance.

3.2 Spatial MixFormer

In the existing skeleton action recognition, Q and K are computed using the following two methods commonly adopted by transformers in RGB: 1) The $Conv_{1\times1}$ expands the channel by a factor of two,

subsequently dividing it into Q and K. 2) The input undergoes two direct convolutions, with the resulting outputs treated as Q and K. A commonality between these approaches is that they utilize the entire input as the basis for calculating Q and K simultaneously, a process known as full self-attention. We contend that employing the above two RGB-based transformer methods to compute Q and K for deriving the weight association matrix somewhat compromises the compactness or purity of skeleton data. This notion arises because, firstly, these strategies originate from the processing approach for RGB data. If a general clipping strategy is employed, such as 16×16, there would still be 256 patch units, which is tenfold greater than the maximum of 25 joints in skeleton data. Secondly, while the image patch units in RGB data have abstracted numerous surrounding pixel features, the features of skeleton data remain unadulterated coordinate data. Therefore, Spatial MixFormer is proposed to alleviate the self-attention dependence of the global complete channel and the lack of expressiveness of the adjacency matrix, as shown in Fig.2. The details are given in the sub-sections.

3.2.1 Optimization of channel grouping strategy. *Firstly*, rather than doubling the dimension in the computation of Q and K and subsequently splitting it, we directly split it using the original number of channels. This approach reduces the parameters while preserving the inherent characteristics. *Secondly*, we increase the number of split channel groups by directly dividing the input into $2n$ unit groups, which form n combination groups, thereby capturing multivariate interaction association characteristics. If the input is denoted as $\mathcal{X}_S^{in} \in \mathbb{R}^{C_s^{in} \times T \times V}$, the grouping process can be expressed as:

$$\mathcal{X}_S' = split_n(trans_v(\mathcal{X}_S^{in})) = concat[x_s^1, x_s^2, ..., x_s^n], \tag{1}$$

where $x_s^i \in \mathbb{R}^{V \times C_s^{in}/n \times T}$. *Thirdly*, to minimize the computational cost increase associated with the transformer structure, we directly pool the number of channels in the groups to one, achieving joint weight smoothing. Next, full connection and linear activation are applied to ensure that the characteristics obtained by Q and K within each group are global, while the adjacency matrix between each group remains specific, as:

$$\begin{aligned} Q_i, K_i &= \sigma(linear(pool_a(split_2(x_s^i)))), \\ A_s^i &= softmax(atten(Q_i, K_i)), \end{aligned} \tag{2}$$

where $Q_i, K_i \in \mathbb{R}^{V \times 1 \times 1}$, $pool_a(\cdot)$ denotes adaptive average pooling, $linear(\cdot)$ adopts fully connected operation, and $\sigma(\cdot)$ denotes activation operation. *Fourthly*, to further enhance the information capacity contained in the multivariate weighted association matrix, we adopt a cross-group-attention strategy and construct the between-group weighted association matrix, as follows:

$$\begin{aligned} A_c^i &= sotmax(atten(Q_{i+1}, K_i)), \\ A_{sc}^i &= A_s^i + A_c^i + A_c^{i-1}, \\ A_{SC} &= concat[A_{sc}^1, A_{sc}^2, ..., A_{sc}^n], \end{aligned} \tag{3}$$

where the first combination group does not include A_c^{i-1}, and the last does not include A_c^i. Regarding tricks, we utilize adjacency matrix complement strategies of Spatial Encoding (SE) [29, 42] and Grouped Edge Masked Encoding (G-EME) [4, 42]. A_{SE} serves to enhance the physical topological properties, ensuring the proper

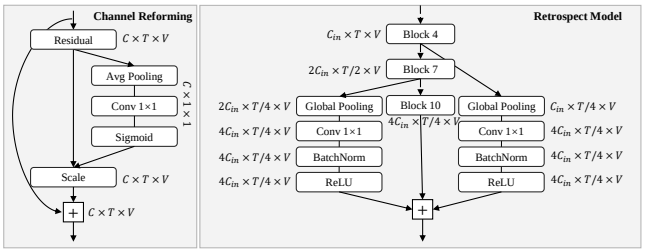


Figure 3: The schema of the Channel Reforming (left) in Sec.3.2.2 and the Retrospect Model (right) in Sec.3.4.

convergence direction of the model. $A_{G\text{-}EME}$ contributes to increasing the autonomy and flexibility of weight learning between joints, while the grouping strategy further reduces the computational cost introduced. Therefore, the final weight incidence matrix of Spatial MixFormer A_{MF} can be expressed as follows:

$$\begin{aligned} A_{MF} &= A_{SC} + A_{SE} + A_{G\text{-}EME}, \\ A_{SE} &= I + A_{in} + A_{out}, \\ A_{G\text{-}EME} &= decoupling(A_m), \end{aligned} \tag{4}$$

where A_{in}, A_{out}, A_m represent centripetal adjacency, centrifugal adjacency, and parameterized adjacency, respectively. We obtain the V_S by unified computation, and the final spatial output can be expressed as follows:

$$\begin{aligned} V_S &= Conv_{1\times1}(Trans_v(\mathcal{X}_S')), \\ \mathcal{X}_S^{out} &= \mathcal{X}_S^{in} + V_S A_{MF}. \end{aligned} \tag{5}$$

3.2.2 Channel Reforming Model. To smooth the feature separation between groups and eliminate noise, the channel relationship of each group needs to be reorganized. We make two improvements to the SE-net [15]. *Firstly*, the objects of average pooling are time and channel, taking the joint as the base dimension. *Secondly*, we remove the *FC* layer, ensuring that the information interaction remains isolated between the joints computed in this module to maintain purity. The specific process is illustrated in Fig.3 (left). The experimental ablation proof refers to Table.4.

3.3 Temporal MixFormer

The Temporal MixFormer is a mixer of Multiscale Convolution, Temporal Transformer, and Sequential Holding Module. In order to maintain the continuity of input with the same timing information for the three sub-modules, the channel grouping strategy is not applied on the input side. Instead, the channel dimension is reduced by $Conv_{1\times1}$ to create multiple input groups, as depicted in Fig.4. If these three modules are denoted as $M1$, $M2$, and $M3$ respectively, the final output can be expressed as follows:

$$X_T^{out} = concat[X_T^{M1}, X_T^{M2}, X_T^{M3}]. \tag{6}$$

3.3.1 M1: Multiscale Convolution. MS-G3D [24] utilizes a strategy of fixing the filter and relaxing dilation to acquire more multivariate multiscale information in temporal domain, while simultaneously reducing the computational cost. When the input is $\mathcal{X}_T^{in} \in \mathbb{R}^{C_t^{in} \times T \times V}$, this strategy can be succinctly expressed as:

$$\begin{aligned} X_T^{m1} &= Conv_{c\to c/5}(\mathcal{X}_T^{in}), \\ X_T^{M1} &= concat[(X_T^{m1})_1, (X_T^{m1})_2, (X_T^{m1})_3], \end{aligned} \tag{7}$$

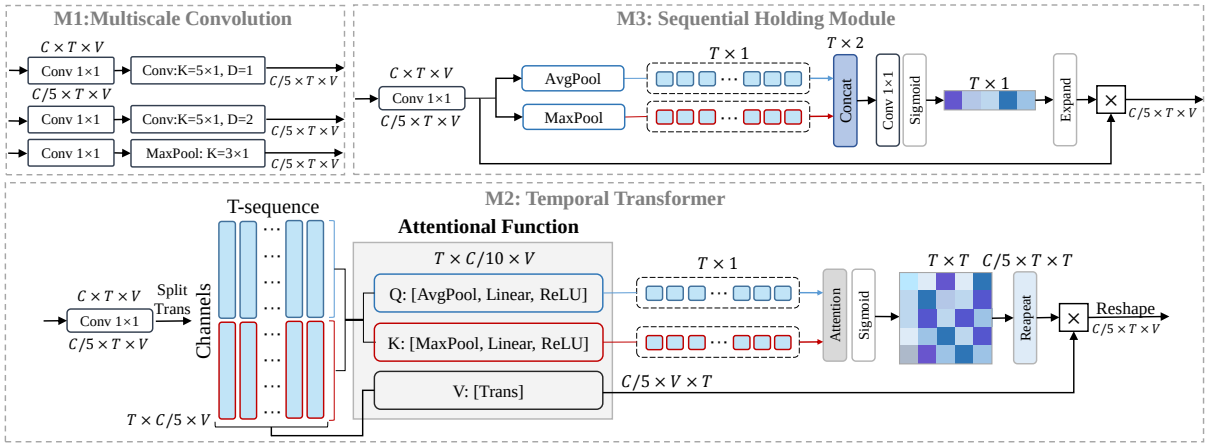


Figure 4: The diagram of Temporal MixFormer. The output consists of three main parts. Multiscale Convolution (M1) is utilized to derive short-term discriminative features. Temporal Transformer (M2) is employed to capture global dissimilarity associations. Sequential Holding Module (M3) is utilized to provide a dynamic benchmark, determining the start, end, and occurrence of actions.

where the $M1$ in Fig.4 shows the process details. We employ a simple optimization by replacing the original weighting with the residual weighting of 2D-TCN [39] to enhance the flexibility of the feature baseline in temporal convolutions. The experimental validation of this optimization is presented in Table.3.

3.3.2 M2: Temporal Transformer. Multiscale convolution obtains the local diversity representation, but transformer is obviously better at obtaining the global temporal correlation information. We adopt a similar compression strategy as Spatial MixFormer, with three key differences: 1) Regarding the number of groups, only two units are divided. 2) The target dimension is time, that is, the channel and joint dimensions need to be compressed. 3) New compression method, Q and K will adopt average pooling and max pooling respectively. When the input is $\mathcal{X}_T^{in}$, the formula of the Temporal Transformer is expressed as:

$$\begin{aligned} x_t^1, x_t^2 &= split_2(trans_t(Conv_{c\to c/5}(\mathcal{X}_T^{in}))), \\ Q_t = \sigma(linear(pool_a(x_t^1))), &K_t = \sigma(linear(pool_m(x_t^2))), \\ A_T^{m2} &= sigmoid(atten(Q_t, K_t)), \end{aligned} \tag{8}$$

where $x_t^1, x_t^2 \in \mathbb{R}^{T\times C_t^{in}/10\times V}$, $Q_t, K_t \in \mathbb{R}^{T\times 1\times 1}$, and $pool_m(\cdot)$ denotes adaptive maximum pooling. We obtain the V_T by unified computation, and the final spatial output can be expressed as follows:

$$\begin{aligned} V_T &= Conv_{1\times 1}(Trans_t(Conv_{c\to c/5}(\mathcal{X}_T^{in}))), \\ X_T^{M2} &= V_T A_T^{m2}, \end{aligned} \tag{9}$$

3.3.3 M3: Sequential Holding Module. In the Temporal Transformer module, we obtain Q, representing the global average representation of temporal features, and K, representing the temporal features with prominent action performance. In the previous module, Q and K are combined using matrix multiplication to obtain the differential temporal adjacency matrix. In this module, we adopt a linear combination of Q and K to obtain a second representation of the time series weights. The purpose of the Sequential Holding Module is to fine-tune the original timing features, which is beneficial for the identification of data with large intra-class differences. The formulas can be expressed as follows:

$$\begin{aligned} X_T' &= Trans_t(Conv_{c\to c/5}(\mathcal{X}_T^{in})), \\ A_T' &= Conv_{c\to c/2}(concat[pool_a(X_T'), pool_m(X_T')]), \\ A_T^{m3} &= expand(sigmoid(A_T')), \\ X_T^{M3} &= X_T' \cdot A_T^{m3}. \end{aligned} \tag{10}$$

3.4 Retrospect Model and Multi-stream strategy

A simple yet effective Retrospect Model is specifically designed for residual information to extract key information twice for the final classification. As shown in Fig.3 (right), the Retrospect Model adopts an adaptive pyramid structure to pass the shallow features back to the final layer, significantly alleviating the key information loss problem due to the small number of joints in the network iteration process. Ablation experiments in Table.5. compare four different module connection strategies and demonstrate the effectiveness of the proposed method. In addition, we validate the model under the widely used 3-stream fusion (3s), 4-stream fusion (4s), and 6-stream fusion (6s), respectively. Following prior work [6], the input of multiple streams refers to $\tilde{\mathcal{X}}_k = (I - P^k)\mathcal{X}$, where $k = 1, 2, ..., K$, and K depends on different datasets.

4 EXPERIMENTS

4.1 Datasets

NTU-RGB+D 60. NTU RGB+D [28] is a 60-classes action recognition dataset completed by 40 volunteers, containing 56,880 skeletal action sequences. The 3D skeleton data includes the 3D positions of 25 main body joints in the human body. Two Benchmark evaluations named cross-subject (C-sub) and cross-view (C-view) are recommended. The testing set consists of 18,960 samples, including two 45-degree views on the left and right of the action.

Table 1: Classification accuracy comparison with state-of-the-art methods on different datasets.

	Method	Publisher	NTU RGB+D 60		NTU RGB+D 120		UAV-Human		NW-UCLA
			C-Sub (%)	C-View (%)	C-Sub (%)	C-Set (%)	CS-v1 (%)	CS-v2 (%)	Top-1 (%)
GCN	ST-GCN [39]	*AAAI'2018*	81.5	88.2	-	-	30.3	56.1	
	ST-GCN++ [11]	*ACMMM'2022*	92.1	97.0	87.5	89.8	-	-	-
	2s-AGCN [29]	*CVPR'2019*	88.5	95.1	82.9	84.9	34.8	66.7	-
	Shift-GCN [5]	*CVPR'2020*	90.7	96.5	85.9	87.6	38.0	67.0	-
	MS-G3D [24]	*CVPR'2020*	91.5	96.2	86.9	88.4	-	-	-
	CTR-GCN [3]	*ICCV'2021*	92.4	96.8	88.9	90.6	43.4	-	96.5
	MKE-GCN [41]	*ICME'2022*	92.5	96.9	89.7	91.1	44.6	-	-
	EfficientGCN [31]	*TPAMI'2022*	92.1	96.1	88.7	88.9	-	-	-
	Info-GCN [6]	*CVPR'2022*	93.0	97.1	89.8	91.2	-	-	97.0
	SAP-CTR [14]	*ACMMM'2022*	93.0	96.8	89.5	91.1	-	-	-
	ACFL-CTR [37]	*ACMMM'2022*	92.5	97.1	89.7	90.9	45.3	-	-
	FR-GCN [46]	*CVPR'2023*	92.8	96.8	89.5	90.9	-	-	96.8
Trans-former	ST-TR [27]	*ICPR'2021*	89.9	96.1	84.3	86.7	-	-	-
	ST-ST [45]	*ACMMM'2021*	91.9	96.8	-	-	-	-	-
	HG-CT [2]	*ICME'2022*	92.2	96.5	89.2	90.6	-	-	-
	FG-STFormer [13]	*ACCV'2022*	92.6	96.7	89.0	90.6	-	-	97.0
	TranSkeleton [21]	*TCSVT'2023*	92.8	97.0	89.4	90.5	-	-	-
	Skeleton MixFormer (3s)		92.6	96.9	89.6	91.0	47.8	72.8	96.8
	Skeleton MixFormer (4s)		93.0	97.0	90.0	91.3	48.7	73.9	97.2
	Skeleton MixFormer (6s)		**93.2**	**97.2**	**90.2**	**91.5**	**48.9**	**74.2**	**97.6**

[1] According to the setting of [6], the six streams are S1: k=1, w/o motion, S2: k=2, w/o motion, S3: k=K, w/o motion, S4: k=1, w/ motion, S5: k=2, w/ motion, S6: k=K, w/ motion, three-stream(3s)=S1+S2+S3; four streams(4s)=S1+S2+S4+S5; six streams (6s)= S1+S2+S3+S4+S5+S6, where K=8 in NTU 60/120, K=6 in NW-UCLA/UAV-Human.

NTU-RGB+D 120. NTU RGB+D 120 [22] is a 120-classes action recognition dataset completed by 106 volunteers, containing 113,945 skeletal action sequences, which is extended from NTU RGB+D 60. Two Benchmark evaluations named cross-subject (C-sub) and cross-set (C-set) are recommended.

Northwestern-UCLA. Northwestern-UCLA [36] is a 10-classes action recognition dataset, containing a total of 1494 video clips, which are shot by three Kinect cameras from different directions. We follow the evaluation method suggested by the author: training data from the first two cameras and test data from the other camera.

UAV-Human. UAV-Human [20] is a 155-classes action recognition dataset containing 22,476 video clips. The dataset was collected by a UAV in multiple urban and rural areas during the day and night. Action data are collected from 119 different subjects and 155 different activity categories at 45 different environmental locations. The authors suggest the following evaluation method: 89 subjects for training and 30 subjects for testing.

4.2 Implementation details

All experiments are conducted on the Pytorch with two NVIDIA RTX 3090ti. We follow previous work [3] for data pre-processing. The batch size of NTU-60, NTU-120, NW-UCLA, and UAV-Human are all 128, the training epoch is set to 90, and we use warm-up for the first 5 epochs. The weight decay is set to 0.0005, and the initialized learning rate is set to 0.1 in NTU-60 & NTU-120 and 0.2 in NW-UCLA & UAV-Human, with a 10× reduction in rounds 35th, 55th, and 75th (only once in 50th for NW-UCLA). The multi-stream fusion strategy [6] is adopted to further improve the performance. We notice the lack of open code for the UAV-Human dataset in terms of unified preprocessing and training, and expose the standard skeleton preprocessing method based on CTR-GCN [3] and SGN [44] to provide more dataset reference for this research direction. Please refer to our published code.

4.3 Compared with the state-of-the-art methods

In this section, we compare the proposed method with state-of-the-art methods on four public benchmarks, and the results are presented in Table.1. Consistent with previous work, we fuse six streams into three main categories (3s, 4s, and 6s). In general, the proposed method is significantly better than the existing methods both in the small-scale dataset NW-UCLA and the large-scale dataset NTU-120. Notably, our transformer-based method achieves a comprehensive outperform of state-of-the-art GCN methods for the first time. Specifically, on NTU-60 C-sub, NTU-120, and NW-UCLA, our method matches the performance of state-of-the-art methods with 6s by using only 4s, and surpasses most transformer-based methods with only 3s. Furthermore, we conduct extensive validation on the challenging latest UAV-Human dataset, providing an up-to-date comparison baseline for skeleton-based action recognition algorithms on this dataset, which complements the absence of Transformer-based approaches.

4.4 Ablation Study

In this section, we evaluate the effectiveness of our proposed method. We begin by analyzing the impact of each module in the Spatial MixFormer on spatial processing, followed by a study of the optimal parameter settings. Similarly, we investigate the impact of each module in the Temporal MixFormer on temporal processing and then analyze the differences in pooling techniques. Additionally,

Table 2: Ablation study on the Spatial Processing.

Spatial Processing	Acc(%) ↑	~GFLOPs ↓	#Param ↓
Baseline	93.86	~3.64	2.27M
+ Spatial MixFormer	95.74 ↑ 1.88	~2.36 ↓ 1.28	1.94M ↓ 0.33
w/o pooling	95.68 ↓ 0.06	~2.75 ↑ 0.39	1.94M −0.00
w/o cross-attention	95.60 ↓ 0.14	~2.35 ↓ 0.01	1.92M ↓ 0.02
w/o $A_{G\text{-}EME}$	95.43 ↓ 0.31	~2.36 −0.00	1.94M −0.00
w/o A_{SE}	95.65 ↓ 0.09	~2.36 −0.00	1.94M −0.00
w/o positional encoding	95.45 ↓ 0.29	~2.36 −0.00	1.94M −0.00

Table 3: Ablation study on the Temporal Processing.

Temporal Processing	Acc(%) ↑	~GFLOPs ↓	#Param ↓
Baseline	88.70	~5.70	5.01M
+ Temporal MixFormer	90.67 ↑ 1.97	~2.36 ↓ 3.34	1.94M ↓ 3.07
w/o M1	89.83 ↓ 0.84	~2.10 ↓ 0.26	1.77M ↓ 0.17
w/o M2	90.34 ↓ 0.33	~2.47 ↑ 0.11	1.99M ↑ 0.02
w/o M3	90.41 ↓ 0.26	~2.48 ↑ 0.12	2.03M ↑ 0.09
M2 w/ two avg-pooling	90.60 ↓ 0.07	~2.36 −0.00	1.94M −0.00
M2 w/ two max-pooling	90.58 ↓ 0.09	~2.36 −0.00	1.94M −0.00
M3 w/ two avg-pooling	90.63 ↓ 0.04	~2.36 −0.00	1.94M −0.00
M3 w/ two max-pooling	90.54 ↓ 0.13	~2.36 −0.00	1.94M −0.00
Residual w/o 2D-TCN	90.47 ↓ 0.20	~2.36 −0.00	1.94M −0.00

we examine the substitution verification of the Channel Reforming and Retrospect Model to evaluate the effectiveness of the model structure. Finally, we validate the number of batch size and the parameter settings in the transformer-based method.

Effectiveness of Spatial MixFormer. The results of our experiments are presented in Table.2. The baseline's spatial processing method utilizes the basic transformer structure, but for comparison fairness, the baseline's Temporal processing method applies the Temporal MixFormer. The experimental results show that the proposed method significantly improves the target model, and the accuracy is increased by 1.88%. Additionally, we conduct ablation experiments to assess the performance of each component in the model. Our findings indicate that the $A_{G\text{-}EME}$ has the most significant impact on the final model result, as its removal decreases accuracy by 0.31%. The reason is that the learnable joint weights can meet the flexibility and adaptability requirements of the model to the greatest extent. Interestingly, the pooling method can significantly reduce computational cost, and its removal does not greatly impact the results. It is speculated that the cross-attention module may play an implicit role in balancing feature communication, although its influence is not obvious, and the other modules also enhance the stability of the model.

Effectiveness of Temporal MixFormer. The results of our experiments are presented in Table.3. The baseline's temporal processing method utilizes the basic Temporal Convolution Network, but for comparison fairness, the baseline's spatial processing method applies the Spatial MixFormer. The experimental results show that the proposed method significantly improves the target model, and the accuracy is increased by about 2.0%. Additionally, we conduct ablation experiments to assess the performance of each component

Table 4: Ablation study on the Channel Reforming.

Variants	NTU RGB+D 60	
	C-sub(%) ↑	C-view(%) ↑
w/o Channel Reforming	90.58	95.46
w/ Channel Reforming	90.67 ↑ 0.09	95.74 ↑ 0.28
w/ Channel Reforming + FC	90.39 ↓ 0.28	95.23 ↓ 0.51
w/ Channel Reforming + maxpool	90.50 ↓ 0.17	95.55 ↓ 0.19
w/ SE-net [15] (FC-ReLU-FC)	90.50 ↓ 0.17	95.43 ↓ 0.31
w/ Channel atten (AAGCN [30])	90.03 ↓ 0.64	95.34 ↓ 0.40

Table 5: Ablation study on the Retrospect Model.

Variants	NTU RGB+D 60	
	C-sub(%) ↑	C-view(%) ↑
w/o Retrospect Model	90.57	95.62
w/ Retrospect Model	90.67 ↑ 0.10	95.74 ↑ 0.12
w/ Block-Residual (one-order)	87.36 ↓ 3.31	93.42 ↓ 2.32
w/ Block-Residual (three-order)	88.49 ↓ 2.18	94.37 ↓ 1.37
w/ Block-Residual (last-connect)	89.29 ↓ 1.38	95.03 ↓ 0.71

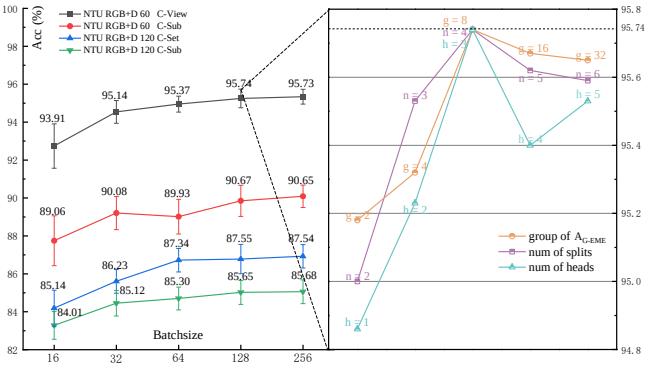
in the model. Our findings indicate that the $M1$ has the most significant impact on the final model result, as its removal decreases accuracy by 0.84%. This is because the multiscale CNN structure has a stronger ability to capture temporal features of short-term sequences, which are crucial for action sequence recognition. Although $M3$ has the least impact on the model's performance, it still improves the final result by 0.26%. Furthermore, Our pooling validation of $M2$ and $M3$ indicates that the averaging and maximization strategies are complementary, and the absence of either strategy slightly affects the final experimental accuracy. Finally, we validate the small optimization proposed in Sec.3.3, which replaces the residual with 2D-TCN, and observe a 0.2% improvement.

Effectiveness of Channel Reforming. As shown in Table.4, the Channel Reforming generally leads to an improvement of 0.1% to 0.3% on different settings. The Channel Reforming is inspired by SE-net [15], with the primary difference being the removal of the FC layer. This operation mainly stems from the experimental proof that FC reduces the accuracy by 0.5%, which is already lower even than deleting the module. We suspect that the transformer architecture plays an essential role in modeling global features, especially the correlations of channels, and that fully connected layers may break this potential connection. Additionally, we test replacing average pooling with max pooling, but this clearly do not yield optimal results, with a reduction about 0.2% in accuracy.

Effectiveness of Retrospect Model. As shown in Table.5, the Retrospect Model generally leads to an improvement about 0.1% on different settings. The Retrospect Model is utilized to extract the key temporal features of dimension difference for a second time before the block of time dimension feature halving. This model can be seen as a simplified pyramid connection. We also test three other commonly used skip connections, but none of them are effective (resulting in a decrease of 0.7% to 3.0% in accuracy).

Parameter and batch size settings. Table.6 (right) illustrates the impact of parameter settings. We observe that increasing the

Table 6: The batch size (left) and parameter (right) settings .



number of channel groups does not always lead to improved performance. In fact, the highest accuracy is achieved when there are four groups, and further increasing the number of groups is counterproductive. We analyze that when the number of groups increases, the total number of features in each group decreases, the effect of average pooling is reduced, and the noise will be more prominent, which is not conducive to the embodiment of discriminative features. Similarly, the optimal number of groups for the $A_{G\text{-}EME}$ component is also around 8, and further increasing the number of groups does not lead to significant improvements in accuracy. This is due to that the group number of $A_{G\text{-}EME}$ is twice the number of channel groups, which aligns with the total number of Q and E. As a result, $A_{G\text{-}EME}$ facilitates the learning of discriminative differences between groups. Finally, the optimal number of multi-heads is three, which may be attributed to the centritic and centrifugal choices in constructing the adjacency matrix [39].

The impact of batch size is illustrated in Table.6 (left), where the source of error bars is the different weight-decay. Existing transformer-based skeleton action recognition lacks validation of the number of batch size, which is a crucial metric that affects the performance of transformer strategies. Our experimental results indicate that a small batch size can be disastrous for transformers, and this holds true in the context of skeleton action recognition. We observe that the model reaches its peak at a batch size of 128, and further increasing it does not have a noticeable impact on accuracy.

4.5 Visualization

Fig.5 (top) presents the grouping visualization of the spatial domain. The darker red color of the incidence matrix indicates the closer connection between the corresponding two joint; the larger the red color of a joint in the skeleton graph indicates the more attention the joint receives. In the none-group setting, the skeleton weights are entangled. In the grouping setting, each group has a clearer division and correlation, with the first and fourth groups focusing more on the whole, and the second and third groups focusing more on the local. And Fig.5 (down) presents the pooling visualization in the temporal domain. We find that the average pooling can clearly determine the climax of the action, while the max pooling can clearly locate the start and end position of the action and small

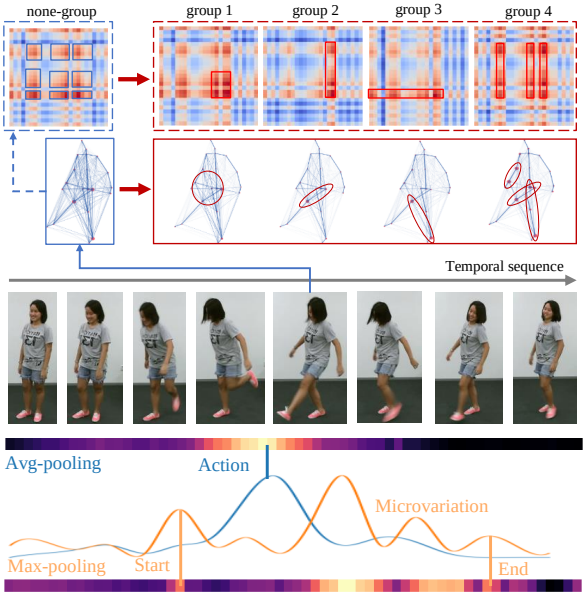


Figure 5: The visualization of weight matrix and skeleton on none-grouped and grouped strategies (top), and the corresponding visualization of different pooling methods on temporal sequence (down).

changes. Both pooling methods are indispensable in the processing of time series, and their effectiveness is also verified in Table.3.

5 CONCLUSION

This paper proposes a spatio-temporal skeleton-based action recognition framework, Skeleton MixFormer, which aims to construct more diverse topological representations through mixing, grouping, and attention strategies. Skeleton MixFormer consists of two novel and effective modules, namely Spatial MixFormer and Temporal MixFormer, to improve the discrimination and interpretability of feature learning. Numerous targeted improvements better fit the characteristics of the skeleton data, enhance the performance ability of the correlation matrix by combining self-attention and cross-attention, and more intelligently explore the potential feature correlation in the spatio-temporal channels. The proposed method completely compensates for the shortcomings of the transformer model and is verified on four datasets, comprehensively surpassing GCN-based methods and reaching the state-of-the-art. In addition, this work provides a new baseline for the UAV-Human dataset and exposes the preprocessed data in the hope of validating more work on this new challenging benchmark.

ACKNOWLEDGMENTS

The work was jointly supported by the National Key R&D Program of China under grant No. 2022ZD0117103, the National Natural Science Foundations of China under grant No. 62272364, the Teaching Reform Project of Shaanxi Higher Continuing Education under Grant No. 21XJZ004.

REFERENCES

[1] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. 2023. STAR-Transformer: A Spatio-temporal Cross Attention Transformer for Human Action Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3330–3339.

[2] Ruwen Bai, Min Li, Bo Meng, Fengfa Li, Miao Jiang, Junxing Ren, and Degang Sun. 2022. Hierarchical graph convolutional skeleton transformer for action recognition. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 01–06.

[3] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13359–13368.

[4] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. 2020. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 536–553.

[5] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 183–192.

[6] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20186–20196.

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13608–13618.

[8] Ronghao Dang, Chengju Liu, Ming Liu, and Qijun Chen. 2022. Channel attention and multi-scale graph neural networks for skeleton-based action recognition. *AI Communications* Preprint (2022), 1–19.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[11] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. 2022. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7351–7354.

[12] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2969–2978.

[13] Zhimin Gao, Peitao Wang, Pei Lv, Xiaoheng Jiang, Qidong Liu, Pichao Wang, Mingliang Xu, and Wanqing Li. 2022. Focal and Global Spatial-Temporal Transformer for Skeleton-based Action Recognition. In *Proceedings of the Asian Conference on Computer Vision*. 382–398.

[14] Ruijie Hou, Yanran Li, Ningyu Zhang, Yulin Zhou, Xiaosong Yang, and Zhao Wang. 2022. Shifting Perspective to See Difference: A Novel Multi-View Method for Skeleton based Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4987–4995.

[15] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[16] Momal Ijaz, Renato Diaz, and Chen Chen. 2022. Multimodal transformer for nursing activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2065–2074.

[17] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. 2020. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications* (2020), 1–27.

[18] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoon Lee. 2022. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:2208.10741* (2022).

[19] Peixuan Li and Jieyu Jin. 2022. Time3d: End-to-end joint monocular 3d object detection and tracking for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3885–3894.

[20] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16266–16275.

[21] Haowei Liu, Yongcheng Liu, Yuxin Chen, Chunfeng Yuan, Bing Li, and Weiming Hu. 2023. TranSkeleton: Hierarchical Spatial-Temporal Transformer for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[22] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2684–2701.

[23] Yanan Liu, Hao Zhang, Dan Xu, and Kangjian He. 2022. Graph transformer network with temporal kernel attention for skeleton-based action recognition. *Knowledge-Based Systems* 240 (2022), 108146.

[24] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 143–152.

[25] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. 2022. Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition* 124 (2022), 108487.

[26] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. 2022. IGFormer: Interaction Graph Transformer for Skeleton-based Human Interaction Recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*. Springer, 605–622.

[27] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. 2021. Spatial temporal transformer network for skeleton-based action recognition. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*. Springer, 694–701.

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 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*. 12026–12035.

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing* 29 (2020), 9532–9545.

[31] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2022. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis and machine intelligence* 45, 2 (2022), 1474–1488.

[32] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence* (2022).

[33] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems* 34 (2021), 24261–24272.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[35] Tijana Vuletic, Alex Duffy, Laura Hay, Chris McTeague, Gerard Campbell, and Madeleine Grealy. 2019. Systematic literature review of hand gestures used in human computer interaction interfaces. *International Journal of Human-Computer Studies* 129 (2019), 74–94.

[36] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. 2014. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2649–2656.

[37] Xuanhan Wang, Yan Dai, Lianli Gao, and Jingkuan Song. 2022. Skeleton-based Action Recognition via Adaptive Cross-Form Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1670–1678.

[38] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. 2023. Transformer for Skeleton-based action recognition: A review of recent advances. *Neurocomputing* (2023).

[39] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[40] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. 2021. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580* (2021).

[41] Sen Yang, Xuanhan Wang, Lianli Gao, and Jingkuan Song. 2022. MKE-GCN: Multi-Modal Knowledge Embedded Graph Convolutional Network for Skeleton-Based Action Recognition in the Wild. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 01–06.

[42] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems* 34 (2021), 28877–28888.

[43] Qinyang Zeng, Chengju Liu, Ming Liu, and Qijun Chen. 2023. Contrastive 3D Human Skeleton Action Representation Learning via CrossMoCo with Spatiotemporal Occlusion Mask Data Augmentation. *IEEE Transactions on Multimedia* (2023).

[44] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer*

vision and pattern recognition. 1112–1121.

[45] Yuhan Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. 2021. STST: Spatial-temporal specialized transformer for skeleton-based action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3229–3237.

[46] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. 2023. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10608–10617.