Parallel Attention Interaction Network for Few-Shot Skeleton-based Action Recognition

Xingyu Liu[1] Sanping Zhou[1*] Le Wang[1] Gang Hua[2]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]Wormpex AI Research

Abstract

Learning discriminative features from very few labeled samples to identify novel classes has received increasing attention in skeleton-based action recognition. Existing works aim to learn action-specific embeddings by exploiting either intra-skeleton or inter-skeleton spatial associations, which may lead to less discriminative representations. To address these issues, we propose a novel Parallel Attention Interaction Network (PAINet) that incorporates two complementary branches to strengthen the match by inter-skeleton and intra-skeleton correlation. Specifically, a topology encoding module utilizing topology and physical information is proposed to enhance the modeling of interactive parts and joint pairs in both branches. In the Cross Spatial Alignment branch, we employ a spatial cross-attention module to establish joint associations across sequences, and a directional Average Symmetric Surface Metric is introduced to locate the closest temporal similarity. In parallel, the Cross Temporal Alignment branch incorporates a spatial self-attention module to aggregate spatial context within sequences as well as applies the temporal cross-attention network to correct misalignment temporally and calculate similarity. Extensive experiments on three skeleton benchmarks, namely NTU-T, NTU-S, and Kinetics, demonstrate the superiority of our framework and consistently outperform state-of-the-art methods.

1. Introduction

Skeleton-based action recognition [48, 2] has attracted increasing attention in recent years, which is a predominant topic in many fields ranging from human-robot interaction to virtual reality, due to its action-focusing nature and compactness [6]. However, how to identify novel actions still remains an open issue. To overcome this problem, more and more works have focused on few-shot action recognition, which can alleviate the resulting performance degradation for the rare category [10, 27, 20, 43, 12]. In particular, these methods explore both the unlabeled query and labeled support sets, so as to learn a discriminative feature representation to match query actions with categories represented by a few support samples [35, 9].

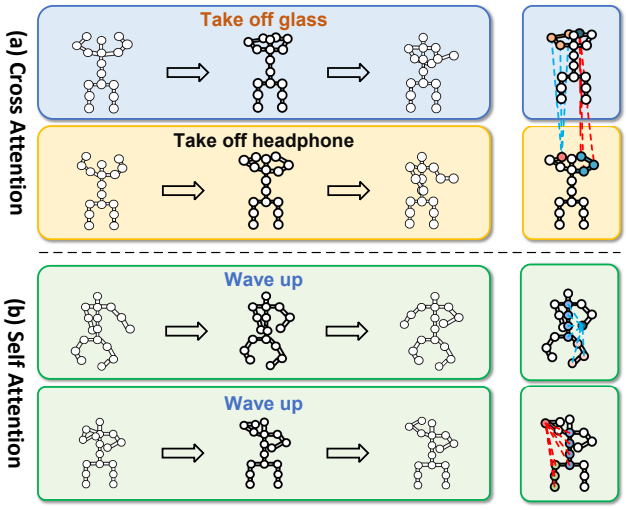


Figure 1: In classifying the samples with similar appearances, the intra-sample representations exhibit body-level comprehensive features, requiring cross-attention to identify subtle differences between joints. In contrast, for samples with inconsistent motion patterns, the inter-sample relationships only reveal differences, necessitating self-attention to enhance the richness of the action.

As shown in Figure 1, current approaches focus on how to exploit intra-skeleton or inter-skeleton relations while ignoring the complementarity between the two paradigms. They are prone to failure in challenging scenarios, such as similar spatial appearances or inconsistent temporal dependencies. On the one hand, some works try to learn discriminative features within sequences. For example, [3, 52, 26, 21, 31] utilize local joint features within the skeleton to capture

*Corresponding author.

distinct patterns. In particular, [3] proposes a part-aware spatial region aggregation, and [52] takes the selective sum of body-part-based local embedding to obtain the individual representation. On the other hand, some works explore adjusting local features across sequences to identify similarities and differences. For example, [25] leverages the cross-attention mechanism to activate their spatial information, and [40] achieves optimal alignment in the temporal and camera viewpoint spaces between query and support.

Even though significant progress has been achieved, we still argue that considering either the dependency within sequences or the association between sequences is inadequate, which necessitates aligning the above samples in parallel. As shown in Figure 1 (a), when classifying samples with similar appearances, such as *take off headphone* and *take off glass*, subtle differences in the spatial position of hand and elbow joints assist in identifying. The interaction between sequences can amplify class-specific information by prioritizing the distinguishing joints, thereby requiring only optimal temporal set matching to identify. For another, when classifying samples with inconsistent motion patterns in Figure 1 (b), such as *wave up*, the elbow and torso joints have distant paths within two graphs but share a solid semantic connection. In such cases, it is necessary to enhance instance-specific information through context aggregation within sequences. With the enrichment of action features, temporally adaptive interactions between sequences can further improve the alignment.

Motivated by this, we propose a novel few-shot skeleton-based action recognition framework, termed as **P**arallel **A**ttention **I**nteraction **Net**work (**PAINet**). We argue that adapting both inter-skeleton and intra-skeleton local joint features are indispensable ways to perfect spatial matching. Compared to previous works, our approach involves the alignment of the spatial and temporal domains within two parallel branches, enabling complementary attention to informative regions in the skeleton sequence. Specifically, we propose a topology encoding module utilizing topology and physical information to enhance the modeling of interactive parts and joint pairs in both branches. In the Cross Spatial Alignment branch, we employ a spatial cross-attention module to build associations of joints across sequences. Afterward, we introduce a directional average symmetric surface metric, which considers all possible pairs of subsequences and selects the pair with maximum similarity. In the Cross Temporal Alignment branch, we propose a spatial self-attention module to aggregate spatial context within sequences. Subsequently, we follow the video-based approach TRX [27] and aggregate the aligned distances by temporal cross-attention matcher. Our contributions can be summarized as follows:

- We propose a novel PAINet for few-shot skeleton-based action recognition, which mitigates challenges posed by similar spatial appearances and inconsistent temporal dependencies during matching.
- We further design a topology encoding module to capture the co-movement between joints and body parts, as well as the intrinsic semantic relations between joints. Also, a directional average symmetric surface metric is proposed to discover the closest temporal relation.
- Extensive experimental results on NTU-T, NTU-S, and Kinetics demonstrate that our model significantly outperforms the state-of-the-art methods.

2. Related Work

Skeleton-based Action Recognition. With the prosperity of skeleton-based action recognition in recent years, early methods such as RNNs or CNNs [22, 34, 16, 8, 50] model skeleton sequences as consecutive vectors or pseudo-images to recognize, ignoring the human body's intrinsic topology. To mitigate this, GCN-based approaches [45, 32] consider the human skeleton as a graph and interleave spatial and temporal modeling separately. Recently, most follow-up works [23, 4, 33, 37, 5] adopt a learnable topology and design high-order or multi-scale adjacency matrices to boost flexibility, limited by the connectivity of handcrafted graphs. Transformer-based approaches [28, 7, 29] utilize spatial and temporal self-attention to attend joint relations, which has sufficient potential to improve modeling capacity but lacks generalizable and intuitive priors. Our method utilizes the common advantages and powerful representation capabilities of theirs to obtain embeddings with rich context information.

Few-shot Action Recognition. Most of the mainstream few-shot methods [41, 44] focus on exploring good metrics to compute the distances between the query and support actions for recognition. Existing works can be divided into video-based and skeleton-based methods. Video-based methods have made significant progress by devising sophisticated matching criteria, and complex multi-level feature associations [40, 38, 43, 18, 12, 24]. In contrast, skeletons have efficient node semantics and a more coherent spatio-temporal motion pattern, free from background clutter. [40] encodes body joints into temporal blocks and then simultaneously performs temporal and view-point alignment by the advanced variant of Dynamic Time Warping (DTW). DASTM [25] proposes a novel spatial matching strategy by adaptively disentangling and activating representations of skeleton joints. Different from previous schemes, we employ parallel attention interaction strategies that complementarily focus on spatially and temporally action-critical regions.

Set Matching. The purpose of set matching in the few-shot setting is to compare the similarity between the feature space of the query and the support. DTW [15] utilizes dynamic programming to calculate the optimal match between

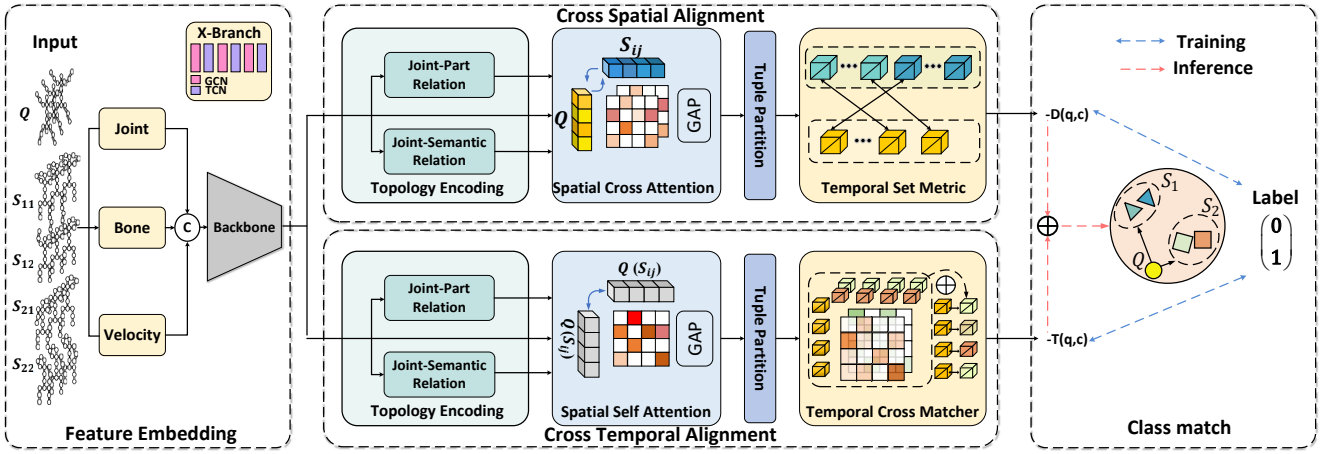


Figure 2: Illustration of our proposed method on a 2-way 2-shot task. X-Branch represents feature extraction from various inputs. The body part pattern and joint semantic associations are obtained by the Joint-Part Relation module and Joint-Semantic Relation module in two parallel branches, respectively. The spatial interactive frame-level feature is obtained by Spatial Cross Attention. The spatial enriched frame-level feature is obtained by Spatial Self Attention. After acquiring all frame-pair-based temporal tuples through tuple partition, the shortest distance $\boldsymbol{D}(q, c)$ and query-specific distance $\boldsymbol{T}(q, c)$ are obtained by Temporal Set Metric and Temporal Cross Matcher, respectively.

two sequences which may vary in speed. Earth Mover's Distance (EMD) [46] generates the optimal matching flows between dense image representations that have the minimum matching cost. Hausdorff distance [13] formulates the longest distance as the similarity between sets, the greatest of all the distances from one set to the closest point in the other set. Among them, average symmetric surface distance [11] is a vital alternative to handle set matching. This paper introduces the directional average symmetric surface distance as a distance match metric between temporal sequences.

3. Method

Our proposed Parallel Attention Interaction Network (PAINet) pipeline is presented in Figure 2. The embedding network with an early-fusion mechanism is employed to extract the features of query and support sequences (Section 3.1). To align spatial and temporal information, we propose the Cross spatial alignment and the Cross temporal alignment branches to exploit inter-skeleton and intra-skeleton semantic information. Specifically, under joint-part relation and joint-semantic relation modules, we obtain body-part patterns and joint semantic associations from the embedded feature (Section 3.2). Subsequently, we first perform the spatial cross-attention and spatial self-attention on the obtained feature to strengthen action-specific spatial relations across and within the skeleton and then obtain frame-level skeleton features by global average pooling (Section 3.3 and 3.4). Next, in the Cross spatial alignment branch, we introduce a directional average symmetric surface metric measuring the closest distance to obtain $\boldsymbol{D}(q, c)$ for class c. In the Cross temporal alignment branch, we follow the temporal cross-attention matcher aggregating query-specific temporal distance to obtain $\boldsymbol{T}(q, c)$ (Section 3.5) for class c. Finally, the distance prediction of the query input and loss $\mathcal{L}$ in two parallel branches are summed to classify (Section 3.6).

3.1. Problem Formulation

Few-shot learning aims to learn a model with strong generalization, which can classify unlabeled query sequences Q into support sets S represented by only a few actions per class. We randomly sample the training eposide $\mathcal{T}$ from the dataset. Each $\mathcal{T}$ includes a support set $\mathbb{S}$ and a query set $\mathbb{Q}$. Notably, the support set $\mathbb{S} = \{\boldsymbol{X}_1^s, \boldsymbol{X}_2^s, \ldots, \boldsymbol{X}_{NK}^s\}$ includes N different classes and K samples for each class, which is called the N-way K-shot problem. $\boldsymbol{X}_i^s \in \mathbb{R}^{T \times V \times C}$ is the i-th sample in support set, where T, V and C denotes its frame length, number of joints, and channel, respectively. For simplicity, we discuss the process for classifying a query sample's $\mathbb{Q} = \{\boldsymbol{X}^q\}$.

As shown in Figure 2, we apply a general spatiotemporal graph convolution network [45, 32, 23] with a multi-model early-fusion mechanism to extract feature representations for each skeleton sequence, obtaining support features $\boldsymbol{F}^s = \{\boldsymbol{F}_{11}^s, \boldsymbol{F}_{12}^s, ..., \boldsymbol{F}_{NK}^s\}$ and the query feature $\boldsymbol{F}^q$.

3.2. Topology Encoding

On the one hand, we consider empirical body grouping and the relation of human parts to facilitate joint-to-part interactions. On the other hand, we incorporate the physical significance of each joint to enhance joint-to-joint semantic association [47].

Joint-Part Relation. Based on the coherence of the body movements, focusing on the movement of corresponding joints is insufficient. The movement of human body parts

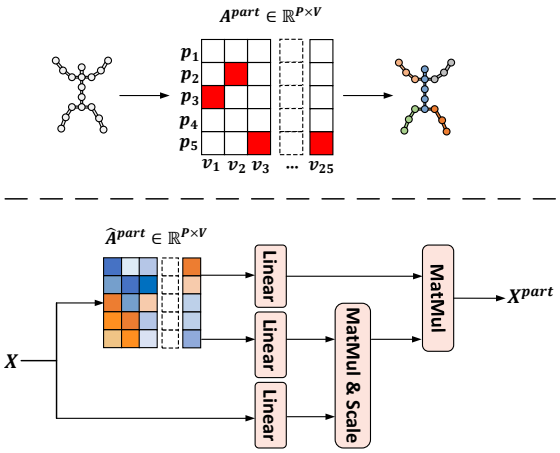


Figure 3: The proposed Joint-Part Relation sub-module used in Topology Encoding block.

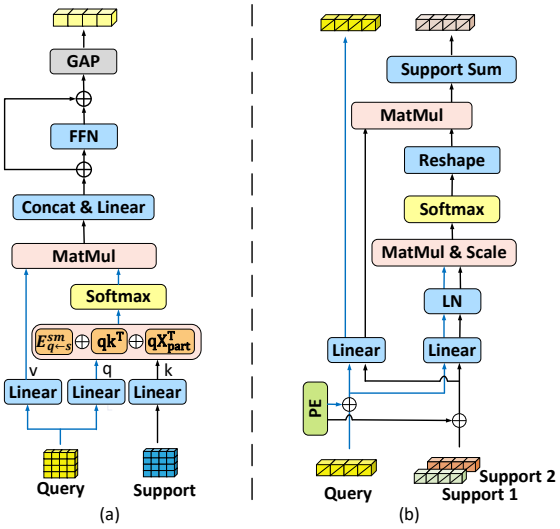


Figure 4: (a) In the pair of query sample and support sample, the spatial cross-attention block per frame. (b) In the comparison of the query sample and all support samples of class c, the temporal cross matcher block on frame pair-based tuples.

carries crucial co-movement information, *i.e.*, the interactions between joints and parts convey rich kinematic information. Based on the physical skeleton structure and human topology prior [36], we first split each skeleton into $P = 5$ body parts, *i.e.*, *left arm, right arm, left leg, right leg, and torso*, and then divide V joints into corresponding body parts. As shown in Figure 3, we propose a part-based allocation matrix $\boldsymbol{A}^{part} \in \mathbb{R}^{P\times V}$, and each column is a one-hot vector representing the part each joint belongs to. To facilitate information interaction between joints and body parts for modeling their co-movement, we devise a Joint-Part Relation sub-block, which is flexible to capture distinct collaborative patterns $\boldsymbol{F}^{\text{part}}$ from $\boldsymbol{F} = \{\boldsymbol{F}^q, \boldsymbol{F}_1^s, \boldsymbol{F}_2^s, \ldots, \boldsymbol{F}_{NK}^s\}$ for skeleton-level action feature and calculated as:

$$\begin{aligned} \boldsymbol{Q} = \boldsymbol{F}\boldsymbol{W}_1, \boldsymbol{K} = \hat{\boldsymbol{A}}^{\text{part}}\boldsymbol{F}\boldsymbol{W}_2, \boldsymbol{V} = \hat{\boldsymbol{A}}^{\text{part}}\boldsymbol{F}\boldsymbol{W}_3, \\ \boldsymbol{F}^{\text{part}} = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{C}})\boldsymbol{V}. \end{aligned} \tag{1}$$

where $\sqrt{C}$ is a scaling factor. $\hat{\boldsymbol{A}}^{\text{part}} = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A}^{\text{part}}+\boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}}$, $\boldsymbol{D}$ is the diagonal degree matrix of $(\boldsymbol{A}^{\text{part}} + \boldsymbol{I})$, $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ denote linear layers.

Joint-Semantic Relation. Considering each physical joint plays a unique semantic role and thus has a specific relation to others [47]. Therefore, it is beneficial to consider the human skeleton's physical information. As joint type information is helpful for learning effective adjacent matrices, such as the relation between *head* and *hand*, we expect the relation of joint semantics to participate in the spatial matching process. Each joint pair within or across the skeleton is hence assigned a trainable parameter scalar, *i.e.*, $\boldsymbol{B}^{\text{sm}} \in \mathbb{R}^{V\times V}$. Combined with pre-defined adjacency matrix $\boldsymbol{A}$ in human skeleton topology, we propose $\boldsymbol{E}^{\text{sm}} = \boldsymbol{A} + \boldsymbol{B}^{\text{sm}}$ to represent joint-wise bias within or across skeletons. Such bias aims to model the inherent joint pair relation.

3.3. Spatial Cross Attention

Cross-attention can adjust the importance of self-joints according to the joint representations of the compared skeleton. Compared to directly using Euclidean distance to measure the distance between all joint representations, cross-attention in spatial matching helps to locate the discriminative regions of actions with similar appearances. As shown in Figure 2, query graph representation $\boldsymbol{F}^q \in \mathbb{R}^{T\times V\times C}$ and support graph representation $\boldsymbol{F}^s \in \mathbb{R}^{NK\times T\times V\times C}$ are spatially interacted to each other using topology encoding related cross-attention mechanism, aiming to emphasize the action-critical region. Considering the asymmetry of cross-attention, the same pair of representations $\boldsymbol{F}^{q\leftarrow s}$ and $\boldsymbol{F}^{s\leftarrow q}$ have different joint association positions and strengths, which means more flexibility for adaptive change.

Cross spatial attention from support to query is drawn in Figure 4 (a). Combined with above collaborative joint-part feature $\boldsymbol{X}^{\text{part}}$ and joint-semantic bias $\boldsymbol{E}^{\text{sm}}$ across skeleton, we define the cross-attention process as:

$$\begin{aligned} &\boldsymbol{Q}^q = \boldsymbol{F}^q\boldsymbol{W}_1^q,\ \boldsymbol{K}^s = \boldsymbol{F}^s\boldsymbol{W}_2^q + \boldsymbol{F}_s^{\text{part}},\ \boldsymbol{V}^q = \boldsymbol{F}^q\boldsymbol{W}_3^q, \\ &\boldsymbol{A}^{q\leftarrow s} = \text{Softmax}(\frac{\boldsymbol{Q}^q(\boldsymbol{K}^s)^{\text{T}}}{\sqrt{C}} + \boldsymbol{E}_{q\leftarrow s}^{\text{sm}}), \\ &\boldsymbol{F}^{q\leftarrow s} = \text{GAP}\left(\text{FFN}\left(\boldsymbol{A}^{q\leftarrow s}V^q\right)\right). \end{aligned} \tag{2}$$

where $\boldsymbol{W}_1^q, \boldsymbol{W}_2^q, \boldsymbol{W}_3^q \in \mathbb{R}^{C\times C}$ are linear projection weights and FFN denotes feed forward network. Finally, global average pooling $\text{GAP}(\cdot)$ over the spatial joints is used to obtain frame-level feature $\boldsymbol{F}^{q\leftarrow s} \in \mathbb{R}^{NK\times T\times C}$ of the query. Similarly, the above formula can also calculate the support

representation $\boldsymbol{F}^{s\leftarrow q} \in \mathbb{R}^{NK\times T\times C}$.

3.4. Spatial Self Attention

Although the skeleton features obtained by the embedding module are interacted by graph convolution, the internal associations of joints per frame contain much action information that has not yet been discovered. To this end, we employ topology encoding related self-attention to enable the joint features to aggregate remaining spatial contexts, as shown in Figure 2. We define $\boldsymbol{W}_1^g, \boldsymbol{W}_2^g, \boldsymbol{W}_3^g \in \mathbb{R}^{C\times C}$ represent weight matrix and $\boldsymbol{F} = \{\boldsymbol{F}^q, \boldsymbol{F}^s\} \in \mathbb{R}^{(1+NK)\times T\times V\times C}$. The $\boldsymbol{Q}^g, \boldsymbol{K}^g, \boldsymbol{V}^g$ are generated by different linear projections, which are computed as follows:

$$\boldsymbol{Q}^g = \boldsymbol{F}\boldsymbol{W}_1^g, \boldsymbol{K}^g = \boldsymbol{F}\boldsymbol{W}_2^g + \boldsymbol{F}^{\text{part}}, \boldsymbol{V}^g = \boldsymbol{F}\boldsymbol{W}_3^g. \quad (3)$$

Then we can obtain a self-attention matrix and reweighted embedding as follows:

$$\begin{aligned} \boldsymbol{A}^g &= \text{Softmax}\left(\frac{\boldsymbol{Q}^g(\boldsymbol{K}^g)^{\text{T}}}{\sqrt{C}} + \boldsymbol{E}^{\text{sm}}\right), \\ \boldsymbol{G} &= \text{GAP}\left(\text{FFN}\left(\boldsymbol{A}^g V^g\right)\right). \end{aligned} \quad (4)$$

where FFN here is used to produce a more spatially refined output with three-layer joint-wise convolution. The resulting joint-level representation can highlight important parts and distinguish complete motion patterns. Likewise, we use global average pooling to obtain frame-level feature $\boldsymbol{G} \in \mathbb{R}^{(1+NK)\times T\times C}$ of query and support.

3.5. Temporal Relation Match

We consider that several subsequences can represent the temporal semantic information of action, including various speeds and locations within sequence [27]. Therefore, the temporal similarity between query and support sequences can be characterized by matching all their subsequences. In practice, we use all frame pair-based subsequences to describe the information of the overall sequence. For example, with $\boldsymbol{s}_i \in \mathbb{R}^C$ as the i^{th} frame representation, a frame pair $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in \mathbb{R}^{2C}$ represents a certain aspect of temporal semantics, $1 \le i < j \le T$.

Temporal Set Metric. For each pair of query feature $\boldsymbol{F}^{q\leftarrow s}$ and support feature $\boldsymbol{F}^{s\leftarrow q}$ that are spatially aligned with each other in Cross spatial alignment branch, we can transform them using the above tuple partitioning method. By traversing all frame pair-based combinations, we can get feature sequence tuple $\hat{\boldsymbol{F}}^{q\leftarrow s} \in \mathbb{R}^{NK\times T'\times 2C}$ and $\hat{\boldsymbol{F}}^{s\leftarrow q} \in \mathbb{R}^{NK\times T'\times 2C}$, where $T' = T(T-1)/2$ represents the number of frame pairs. Next, all features in the query set $\mathbb{Q}$ are averaged for sample number K in each class to form $\hat{\boldsymbol{H}}^{q\leftarrow s} \in \mathbb{R}^{(N\times(1\times T')\times 2C}$. The feature in support set $\mathbb{S}$ form representation $\hat{\boldsymbol{H}}^{s\leftarrow q} \in \mathbb{R}^{N\times(K\times T')\times 2C}$, where sample diversity in each class are preserved.

By utilizing spatially aligned features, we can identify the most temporally correlated locations for action matching. To calculate the distance between $\hat{\boldsymbol{H}}_c^{q\leftarrow s}$ and $\hat{\boldsymbol{H}}_c^{s\leftarrow q}$ for class c, we consider performing exhaustive pair-to-pair comparisons in the temporal dimension to seek the optimal local match for each frame pair. And then, we average the optimal local distance for all frame pairs, which is robust to misaligned instances. Motivated by the set matching strategy, we develop a flexible directional average symmetric surface metric [11, 51] into the few-shot action recognition field. In contrast to the bi-directional distance between query and support [42], we only learn the one-way distance from each query to multiple support representations, thereby avoiding the noise introduced from more support subsequences to the query. Based on spatially aligned features and temporal tuple partition, Eq. (5) can automatically find the best temporal correspondencies $\boldsymbol{D}(q, c)$ between query q and class c, which is given by:

$$\boldsymbol{D}(q,c) = \frac{1}{T'} \sum_{\hat{\boldsymbol{h}}_i^q \in \hat{\boldsymbol{H}}^{q\leftarrow s}} \left(\min_{\hat{\boldsymbol{h}}_j^s \in \hat{\boldsymbol{H}}^{s\leftarrow q}} \|\hat{\boldsymbol{h}}_i^q - \hat{\boldsymbol{h}}_j^s\| \right). \quad (5)$$

It is worth noting that directional average symmetric surface metric does not involve parametric design and relies on the feature obtained by the spatial cross-attention module.

Temporal Cross Matcher. Based on enriched frame-level feature $\boldsymbol{G}$, we consider utilizing temporal cross-attention to mitigate temporal misalignment globally. In our work, we followed the [38], which matched each query sub-sequence with all sub-sequences in the support set to construct a query-specific class prototype.

Specifically, we first globally exchange and integrate frame-wise and channel-wise features with Temporal-Channel Mixer [39] as follows:

$$\begin{aligned} \boldsymbol{U} &= \boldsymbol{G}^{\text{T}} + \sigma\left(\text{LN}(\boldsymbol{G}^{\text{T}})\boldsymbol{W}_T\right), \\ \boldsymbol{Z} &= \boldsymbol{U}^{\text{T}} + \sigma\left(\text{LN}(\boldsymbol{U}^{\text{T}})\boldsymbol{W}_C\right), \end{aligned} \quad (6)$$

where $\sigma(\cdot)$ denotes the GELU activation function and $\text{LN}(\cdot)$ denotes layer normlization. $\boldsymbol{W}_T$ and $\boldsymbol{W}_C$ are two-layer MLP in temporal and channel dimensions, respectively. Next, we adopt tuple partition method to obtain $\hat{\boldsymbol{Z}}^q$ and $\hat{\boldsymbol{Z}}^s$. As shown in Figure 4 (b), temporal cross-attention between query and support of class c is used to calculate the query-specific class prototypes $\hat{\boldsymbol{Z}}^{s\leftarrow q,c}$ via an aggregation of all possible sub-sequences in the support set. Similar to spatial cross-attention, the correspondence between query pair and support pair of support sequence k in class c is calculated as:

$$\begin{aligned} \boldsymbol{A}_k^c &= \text{Softmax}\left(\text{LN}\left(\hat{\boldsymbol{Z}}_k^{s,c}\boldsymbol{W}_1^t\right)\text{LN}\left(\hat{\boldsymbol{Z}}^q\boldsymbol{W}_1^t\right)\right), \\ \hat{\boldsymbol{Z}}^{s\leftarrow q,c} &= \sum_k \boldsymbol{A}_k^c(\hat{\boldsymbol{Z}}_k^{s,c}\boldsymbol{W}_2^t), \quad \hat{\boldsymbol{Z}}^{q,c} = \boldsymbol{W}_2^t\hat{\boldsymbol{Z}}^q. \end{aligned} \quad (7)$$

Backbones	ST-GCN			2s-AGCN			MS-G3D		
Methods	NTU-T	NTU-S	Kinetics	NTU-T	NTU-S	Kinetics	NTU-T	NTU-S	Kinetics
ProtoNet [35]	71.2/81.1	73.3/84.3	37.4/46.8	68.1/81.9	72.8/84.2	38.4/50.5	70.1/82.3	73.6/85.3	39.5/50.0
DTW [15]	74.0/81.0	73.5/81.5	39.2/47.9	70.8/81.2	71.5/82.5	40.9/50.8	72.4/81.3	73.9/83.2	40.6/50.0
NGM [10]	71.8/81.4	75.7/84.2	39.1/48.6	72.2/83.2	73.2/85.9	40.9/49.8	73.5/83.1	76.9/86.7	40.8/50.7
DropEdge [30]	67.3/77.9	70.7/78.6	38.9/48.2	70.1/80.5	72.6/83.1	39.9/50.2	68.7/80.9	69.5/80.2	39.4/50.1
PairNorm [49]	72.9/81.8	72.8/81.4	39.3/48.6	70.0/80.0	70.8/80.3	40.9/50.4	71.0/81.6	70.8/82.5	40.7/50.6
DASTM [25]	75.1/83.0	76.2/85.5	39.3/48.9	73.3/83.8	74.0/86.8	40.8/50.9	75.0/84.9	76.3/87.3	41.1/51.1
PAINet	**82.4/90.8**	**84.6/92.7**	**42.4/53.4**	**78.8/89.9**	**82.9/91.5**	**43.2/53.8**	**81.3/90.9**	**84.2/92.3**	**42.5/54.1**

Table 1: Experimental results with three different backbones on 5-way 1-shot and 5-way 5-shot benchmarks of NTU-S, NTU-T, and Kinetics. The best accuracy(%) is highlighted. The left column represents 1-shot accuracy, while the right column represents 5-shot accuracy.

Afterward, the distances $\boldsymbol{T}(q, c)$ between sub-sequences of query Q and their corresponding query-specific class prototypes of support are averaged, which obtains the distance of the query to class c as follows:

$$\boldsymbol{T}(q, c) = \frac{1}{T'} \sum_{t \in T'} \|\hat{\boldsymbol{Z}}_t^{q,c} - \hat{\boldsymbol{Z}}_t^{s \leftarrow q,c}\|. \tag{8}$$

3.6. Training

For the distances $\boldsymbol{D}(q, \cdot)$ and $\boldsymbol{T}(q, \cdot)$ obtained by the above two alignment branches, which take the negative distance for each class as logit. Then, given the ground-truth labels $y \in \mathbb{R}^N$, we use the $\mathrm{softmax}$ function to obtain the class probability followed by a standard cross-entropy loss. With $\mathcal{L}_{cs}$ and $\mathcal{L}_{ct}$ representing the cross-spatial and cross-temporal alignment loss respectively, our model is trained by:

$$\mathcal{L} = \mathcal{L}_{cs}(-\boldsymbol{D}(q, \cdot), y) + \lambda \mathcal{L}_{ct}(-\boldsymbol{T}(q, \cdot), y), \tag{9}$$

where λ is a constant weight. We use the weighted sum of the above two negative distances during inference, and the query is assigned to the closest category.

4. Experiment

4.1. Datasets

NTU RGB+D 120 [21] dataset is currently the largest 3D skeleton-based action recognition dataset that contains 114,480 skeleton sequences of 120 action classes. Each skeleton sequence contains the 3D spatial coordinates of 25 joints detected by the depth sensor. Our experiments use 120 action classes, including 80, 20, and 20 classes as training, validation, and test classes. Following, we randomly use 60 samples and 30 samples for each category, denoted as two subsets "**NTU-S**" and "**NTU-T**", respectively.

Kinetics [14] is a large-scale video clip that covers more than 400 human action classes, which includes human-object interactions and human-human interactions. The publicly available Openpose [1] toolbox estimates the location of 2D spatial coordinates on every frame of the clips as the initial joint feature. In this experiment, we only use the first 120 actions with 100 samples per class provided by [25]. The number of training/validation/test partitions is identical to NTU RGB+D 120.

4.2. Evaluation

We evaluate the 5-way 1-shot and 5-way 5-shot action recognition tasks and report the average accuracy over 500 randomly selected episodes from the test stage. We use the variant of DASTM [25] as the baseline, which uses spatial activation as the spatial alignment and the DTW strategy as the temporal alignment. Our method is consistent with the baseline without using any pre-trained model and additional auxiliary datasets for few-shot learning.

4.3. Implementation Details

Spatial-temporal backbones. We utilize typical ST-GCN [45], 2s-AGCN [32], and MS-G3D [23] as the backbones with an early-fusion mechanism to encode skeletal action sequences. The early fusion module utilizes raw skeleton data to generate multi-modality data, i.e., joint, bone, and velocity. Concretely, the input branches for different modalities are implemented by three stacked layers of backbone for complete feature embedding.

Experimental configuration. We adopt the same data-preprocessing procedure as introduced in [25]. During the training and testing stage, the sampled frame number T is set to 50 and 30 per skeleton sequence. We optimize the PAINet model with Adam [17] optimizer, where the initial learning rate is 0.01. For each epoch, we randomly sample 1,000 episodes for training and 500 episodes for validation to ensure sufficient generalization. The model is trained for 100 epochs, and the final performance is reported as the average of 10 epochs in the test stage. To ensure convincing results, each experiment is repeated **3** times to obtain the mean accuracy and standard deviation. All experiments are conducted with one GeForce RTX 3090 GPU.

Method	Early-fusion	SCA^*	TSM^*	$SSA^\dagger$	TCM	$TCA^\dagger$	Accuracy(%)
Baseline							75.1
Cross Spatial Alignment	–	–	✓	–	–	–	78.5
	–	✓	✓	–	–	–	79.2
	✓	✓	✓	–	–	–	**80.1**
	✓	✓	✓	–	✓	–	79.3
Cross Temporal Alignment	–	–	–	–	–	✓	78.6
	✓	–	–	–	–	✓	78.9
	✓	–	–	–	✓	✓	80.8
	✓	–	–	✓	✓	✓	**81.8**
PAINet	✓	✓	✓	✓	✓(T)	✓	**82.4**

Table 2: Ablation study of the effect of the modules in our framework. The baseline is constructed by cross attention activation and temporal alignment DTW. Among them, **SCA** denotes Spatial Cross Attention, **TSM** denotes Temporal Set Metric, **SSA** denotes Spatial Self Attention, **TCM** denotes Temporal-Channel Mixer, **TCA** denotes Temporal Cross Attention Matcher. $*$ and † indicate that the module belongs to Cross spatial alignment and Cross temporal alignment, respectively.

4.4. State-of-the-art Comparison

As shown in Table 1, we conduct experiments using three backbones on three mainstream datasets under a 5-way 1-shot and 5-way 5-shot setting. Specifically, our model significantly outperforms the SOTA approach on the NTU-series datasets and achieves decent improvements on the noisy Kinetics dataset. Furthermore, PAINet achieves considerable performance improvements on the NTU-S dataset with more complex action categories, demonstrating its better generalization. Compared to having the human body structure as the default connection, the multi-layer stacked self-attention structure in 2s-AGCN will lead to over-smooth representations after the message passes through different joints [25]. Besides, MS-G3D achieves comparable results to ST-GCN, indicating that directly adopting complex graph convolution networks will not lead to better generalization, proving the potential of simple spatial-temporal graph convolution.

4.5. Ablation Study

For convenience, all ablation experiments are achieved on the NTU-T dataset using STGCN as the backbone.

Impact of proposed contributions. As shown in Table 2, our model outperforms the baseline on separate alignment branches with temporal relation match, demonstrating that temporal tuple partition and align strategy can solve inconsistent temporal dependency. Besides, performance on the parallel branch also dramatically benefits from spatial cross-attention between query and support and spatial self-attention within each input. Secondly, the temporal channel mixer module improves under the cross-temporal alignment branch while negatively impacting the cross-spatial alignment branch. We argue that spatial cross-attention performs frame-by-frame spatial alignment between query and support, and mixing subsequent adjacent frames will cause the previous alignment to fail. Thirdly, merging parallel branches can complementarily focus on temporal and spatial action-critical regions. Additional results on loss functions, different motion modes, computational complexity, and efficiency can be found in the supplementary materials.

Spatial Aggregation	1-shot(%)	5-shot(%)
Self-Att	81.9	90.3
Mask Self-Att	81.6	90.0
Mask ST Self-Att	80.9	89.2
Channel-Specific Att	80.8	89.3
Self-Att + JP	**82.2**	**90.6**
Self-Att + JP + JS	**82.4**	**90.8**

Table 3: Impact of varying spatial aggregation mechanisms within skeleton. **Mask Self-Att** refers to randomly masking several joints on specific frames. **ST Self-Att** denotes the aggregation of spatial joints across some frames with the slide window mechanism. **Channel-specific Att** refers to channel-dependent attention patterns. **JP** and **JS** refers to joint-part and joint-semantic relation.

Impact of spatial joints aggregation. We present the impact of varying spatial aggregation mechanisms in Table 3. The mask strategy enables the model to learn random skeletal associations and improve the generalization ability [19]. However, skeletons have evident sparse distribution and semantic associations, and random masks hence affect meaningful spatial interactions between joints. We consider the spatial-temporal interaction of joints using sliding temporal windows [23]. Although the masking mechanism here is utilized to reduce the spread of redundant messages, which fails to distinguish beneficial messages. In this field, channel-specific attention [4] serves as an effective data-driven way of dynamic joint aggregation. To learn multi-channel motion patterns, however, multiple stacked layers make the joints over-smoothed and less discriminative than before [25]. Thereby, under the topology encoding module, employing self-attention for frame-wise spatial aggregation achieves

Temporal Metric	Temporal Set	1-shot(%)	5-shot(%)
Euclidean	frame-based	79.2	86.9
DTW		79.6	87.5
EMD		80.1	88.1
Hungarian		81.6	89.5
DASSM		81.3	89.4
Hungarian	**pair-based**	80.5	89
DASSM		**82.4**	**90.8**

Table 4: Comparison of various temporal set metrics in Cross spatial alignment branch. Earth Mover's Distance (EMD) and Hungarian algorithm are reimplemented to measure the distance between query and support.

action class	DASTM	Ours
brush hair	0.41	0.63(**+0.22**)
hopping	0.53	0.94(**+0.41**)
take off headphone	0.69	0.91(**+0.22**)
apply cream on hand	0.73	0.93(**+0.2**)
high-five	0.78	0.93(**+0.15**)

Table 5: Major performance gains obtained by our model over DASTM[25] on test categories.

the best generalization performance and sufficient flexibility.

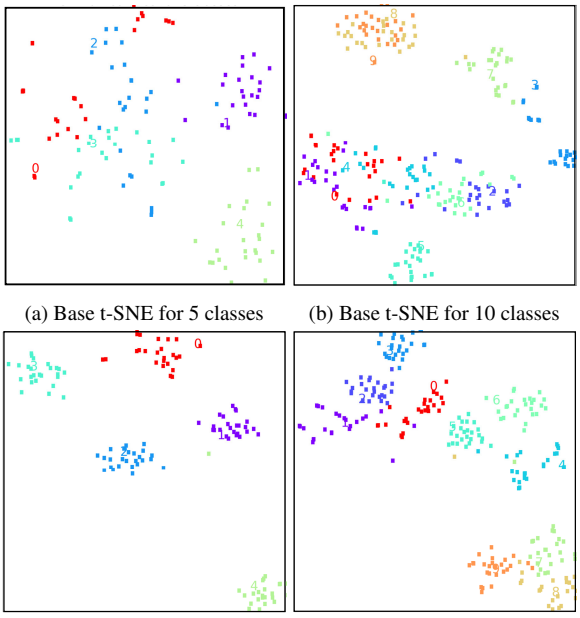
Impact of temporal set metric. Table 4 shows a performance comparison when integrating different temporal match algorithms in our Cross Spatial Alignment branch. For the frame-based partition, the sequence retains the original order information and has less semantic content. Besides, given noisy datasets [14], skeleton information on a single frame can not be effectively distinguished, and the optimal path will be severely disturbed. For the pair-based subsequence, the directional average symmetric surface metric represents the most semantic overlap relationship between each pair of query and all subsequences of the support category. However, the Hungarian algorithm needs a threshold to set the status of the matching object and one-to-one matching is not suitable for matching between subsequences.

4.6. Visualization

In Figure 5, we present a visualization of embeddings before and after adaptation. Generally, due to limited samples in meta-training, it is difficult for few-shot models to form accurate clusters without explicit regularization. We observe that after applying our alignment strategy on based representation, embeddings of the query are clustered compactly relative to the corresponding support representation.

4.7. Performance gains

In Table 5, we observe that our framework achieves gain above 20% for classes such as *brush hair*, *hopping*, *take off headphone* and *apply cream on hand*. Cross spatial alignment enhances the discrimination of important joint features for actions determined by local joint motions, such as *take off headphone*. And for action *hopping*, which depends on the temporal context and temporal subsequences, Cross temporal alignment can align temporal discriminative features.



(a) Base t-SNE for 5 classes (b) Base t-SNE for 10 classes

(c) Adaptive t-SNE for 5 classes (d) Adaptive t-SNE for 10 classes

Figure 5: t-SNE of features on NTU-T dataset for N-way 1-shot setting. The top row shows **base** features after global average pooling following the backbone whereas the bottom row shows **adaptive** features after the parallel alignment branch. Dots and numbers denote query features and support features, respectively.

5. Conclusion

We propose an effective Parallel Attention Interaction Network (**PAINet**) for few-shot skeleton-based action recognition, aiming to learn discriminative representations for novel actions by intra- and inter-relations learning. Besides, we develop a topology encoding module to improve skeleton embedding and a directional average symmetric surface metric for robust matching. Experimental results on the mainstream benchmark datasets have shown the benefits of our method.

Acknowledgement

This work was supported partly by National Key R&D Program of China under Grant 2021YFB1714700, NSFC under Grants 62088102 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC-41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 6

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[3] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Qian He, Chuanyang Hu, Errui Ding, Yu Guan, and Xuming He. Part-aware prototypical graph network for one-shot skeleton-based action recognition. In *FG*, pages 1–8, 2023. 1, 2

[4] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 2, 7

[5] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. InfoGCN: Representation learning for human skeleton-based action recognition. In *CVPR*, pages 20186–20196, 2022. 2

[6] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. PoTion: Pose motion representation for action recognition. In *CVPR*, pages 7024–7033, 2018. 1

[7] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. DG-STGCN: Dynamic spatial-temporal modeling for skeleton-based action recognition. *arXiv preprint arXiv:2210.05895*, 2022. 2

[8] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *CVPR*, pages 2959–2968, 2021. 2

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICCV*, pages 1126–1135, 2017. 1

[10] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. Neural graph matching networks for fewshot 3d action recognition. In *ECCV*, pages 653–669, 2018. 1, 6

[11] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE T-MI*, 28(8):1251–1265, 2009. 3, 5

[12] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *ECCV*, pages 351–368, 2022. 1, 2

[13] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE T-PAMI*, 15(9):850–863, 1993. 3

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6, 8

[15] Eamonn J Keogh and Michael J Pazzani. Derivative dynamic time warping. In *SDM*, 2001. 2, 6

[16] Tae Soo Kim and Austin Reiter. Interpretable 3D human action analysis with temporal convolutional networks. In *CVPRW*, pages 1623–1631, 2017. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[18] Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. TA2N: Two-stage action alignment network for few-shot action recognition. In *AAAI*, pages 1404–1411, 2022. 2

[19] Chen Liu, Yanwei Fu, Chengming Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. Learning a few-shot embedding model with contrastive learning. In *AAAI*, pages 8635–8643, 2021. 7

[20] Huabin Liu, Weixian Lv, John See, and Weiyao Lin. Task-adaptive spatial-temporal video sampler for few-shot action recognition. In *ACM MM*, pages 6230–6240, 2022. 1

[21] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *IEEE T-PAMI*, 42(10):2684–2701, 2019. 1, 6

[22] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE T-IP*, 27(4):1586–1599, 2017. 2

[23] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 2, 3, 6, 7

[24] Wenyang Luo, Yufan Liu, Bing Li, Weiming Hu, Yanan Miao, and Yangxi Li. Long-short term cross-transformer in compressed domain for few-shot video classification. In *IJCAI*, pages 1247–1253, 2022. 2

[25] Ning Ma, Hongyi Zhang, Xuhui Li, Sheng Zhou, Zhen Zhang, Jun Wen, Haifeng Li, Jingjun Gu, and Jiajun Bu. Learning spatial-preserved skeleton representations for few-shot action recognition. In *ECCV*, pages 174–191, 2022. 2, 6, 7, 8

[26] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Delving deep into one-shot skeleton-based action recognition with diverse occlusions. *IEEE T-MM*, 25(3):1489 – 1504, 2023. 1

[27] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*, pages 475–484, 2021. 1, 2, 5

[28] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In *ICPR*, pages 694–701, 2021. 2

[29] Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv preprint arXiv:2201.02849*, 2022. 2

[30] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. DropEdge: Towards deep graph convolutional networks on node classification. In *ICLR*, 2020. 6

[31] Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In *CVPR*, pages 6471–6480, 2023. 1

[32] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 2, 3, 6

[33] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. AdaSGN: Adapting joint number and model size for efficient skeleton-based action recognition. In *ICCV*, pages 13413–13422, 2021. 2

[34] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *CVPR*, pages 1227–1236, 2019. 2

[35] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4080–4090, 2017. 1, 6

[36] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *ACM MM*, pages 1625–1633, 2020. 4

[37] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *ICCV*, pages 13328–13338, 2021. 2

[38] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *CVPR*, pages 19958–19967, 2022. 2, 5

[39] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. MLP-Mixer: An all-MLP architecture for vision. In *NeurIPS*, pages 24261–24272, 2021. 5

[40] Lei Wang and Piotr Koniusz. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *ACCV*, pages 4176–4193, 2022. 2

[41] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. MoLo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *CVPR*, pages 18011–18021, 2023. 2

[42] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *CVPR*, pages 19948–19957, 2022. 5

[43] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *CVPR*, pages 9151–9160, 2022. 1, 2

[44] Jiazheng Xing, Mengmeng Wang, Yong Liu, and Boyu Mu. Revisiting the spatial and temporal modeling for few-shot action recognition. In *AAAI*, pages 3001–3009, 2023. 2

[45] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018. 2, 3, 6

[46] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. DeepEMD: Differentiable earth mover's distance for few-shot learning. *IEEE T-PAMI*, 45(5):5632–5648, 2022. 3

[47] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pages 1112–1121, 2020. 3, 4

[48] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE T-MM*, 19(2):4–10, 2012. 1

[49] Lingxiao Zhao and Leman Akoglu. Pairnorm: Tackling oversmoothing in gnns. In *ICLR*, 2020. 6

[50] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *CVPR*, pages 10608–10617, 2023. 2

[51] Zhi-Qiang Zhou and Bo Wang. A modified hausdorff distance using edge gradient for robust object matching. In *ICSIP*, pages 250–254, 2009. 5

[52] Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition. In *WACV*, pages 6038–6047, 2023. 1, 2