Short Paper

One-Shot Action Recognition via Multi-Scale Spatial-Temporal Skeleton Matching

Siyuan Yang [ID], Jun Liu [ID], Shijian Lu [ID], Er Meng Hwa [ID], *Life Fellow, IEEE*, and Alex C. Kot [ID], *Life Fellow, IEEE*

Abstract—**One-shot skeleton action recognition, which aims to learn a skeleton action recognition model with a single training sample, has attracted increasing interest due to the challenge of collecting and annotating large-scale skeleton action data. However, most existing studies match skeleton sequences by comparing their feature vectors directly which neglects spatial structures and temporal orders of skeleton data. This paper presents a novel one-shot skeleton action recognition technique that handles skeleton action recognition via multi-scale spatial-temporal feature matching. We represent skeleton data at multiple spatial and temporal scales and achieve optimal feature matching from two perspectives. The first is multi-scale matching which captures the scale-wise semantic relevance of skeleton data at multiple spatial and temporal scales simultaneously. The second is cross-scale matching which handles different motion magnitudes and speeds by capturing sample-wise relevance across multiple scales. Extensive experiments over three large-scale datasets (NTU RGB+D, NTU RGB+D 120, and PKU-MMD) show that our method achieves superior one-shot skeleton action recognition, and outperforms SOTA consistently by large margins.**

Index Terms—**One-shot learning, optimal matching, skeleton action recognition.**

I. INTRODUCTION

Human action recognition is a fast-developing research area due to its wide applications in human-computer interaction, video surveillance, game control, etc. In recent years, human action recognition with skeleton data has attracted increasing attention as skeleton data encodes high-level representations of human actions and is generally lightweight and robust to variations in appearances, surrounding distractions, viewpoint changes, etc. As of today, most existing studies expect large-scale labeled training data for learning effective human action representations. While facing skeleton data of a new category, they require to collect hundreds of action samples of the new category for adapting or fine-tuning some existing models. How to achieve single-shot recognition for new action categories becomes critically important for circumventing the tedious and laborious data collection and labeling procedure.

One-shot skeleton action recognition is a very challenging task. Beyond the data constraint with one-shot for unseen new classes, the major challenge comes from the very rich variations in human actions. Take the action "put on glasses" as an example. Different persons could perform it by using their left hand, right hand, or both hands. The same person could also perform it at different paces with different motion dynamics. Different approaches have been explored to address this challenging task, and most existing works [1], [2], [3], [4] represent the anchor and target samples with certain pooled feature vectors and compute the sample distance based on the similarity of their pooled feature vectors. However, the adoption of such global feature similarity discards the very useful spatial structures and temporal order of the skeleton sequences. In addition, most existing works learn skeleton action representations at a single scale of the original body joints which tends to lose useful action features under the one-shot scenario. Drawing from [5], [6], [7], [8], [9], human actions are multi-scales in both spatial and temporal spaces. For instance, multiple joints on arms and legs collaborate in walking, and consecutive frames of human actions contain strong temporal correlations. Skeleton action representations should therefore capture the rich semantic correlations at different spatial and temporal scales. In contrast to [5], [6], [7], [8], [9], which employ multi-scale information for recognition and prediction, we leverage the extracted multi-scale features for match and design the innovative matching strategies specifically for one-shot skeleton action recognition challenge.

We propose to capture spatial-temporal features by leveraging spatial structures and temporal orders of skeleton sequences as illustrated in Fig. 1. Inspired by the theory of optimal transport [10], [11], we measure the semantic relevance of two skeleton sequences by computing an optimal matching flow between their feature maps. Specifically, we adopt Earth Mover's Distance (EMD) [12] as the optimal matching metric for acquiring the optimal matching flow. EMD is the metric for computing the distance between two representations, enabling us to determine the similarity between the feature representations of two skeleton samples. In our one-shot skeleton action recognition scenario, given the distances between all skeleton joint pairs, EMD maximizes the impact of relevant joints and minimizes the effect of irrelevant joints between two skeleton sequences. In addition, we model skeleton sequences at multiple spatial scales (joint-scale, part-scale, and limb-scale) and temporal scales as illustrated in Fig. 2, and perform multi-scale matching to capture scale-wise skeleton semantic relevance by using EMD. Further, human action could be performed with different motion magnitudes and motion
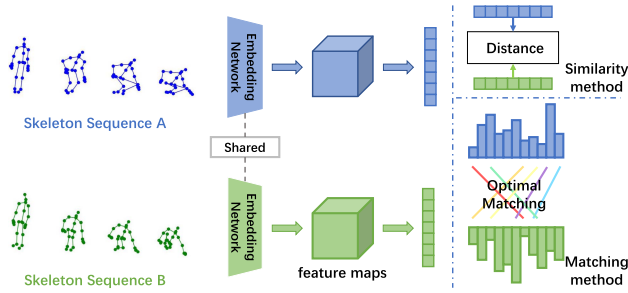
Fig. 1. Skeleton action recognition based on feature similarity or feature matching: Feature similarity computes the distance between feature vectors which discards the very useful spatial skeleton structures and temporal information. The proposed feature matching compares two skeleton sequences by computing a matching flow between their feature distributions which can capture useful spatial and temporal information effectively. The colored line emphasizes channels paired based on their high matching scores.
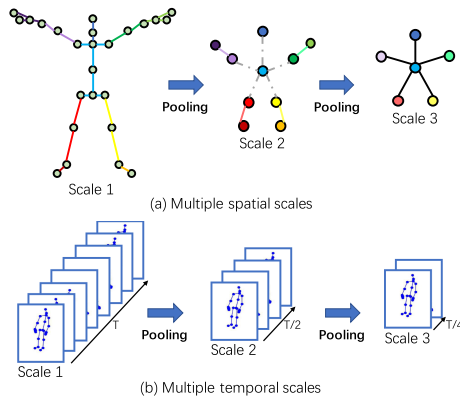

Fig. 2. Proposed multi-scale skeleton modeling at spatial dimension in (a) and temporal dimension in (b) Given the original spatial scale at *Scale 1*, we first divide the skeleton nodes into multiple groups with similar semantic skeleton structures and then perform average pooling to each group to generate skeleton graphs of coarser scales (i.e., *Scale 2* and *Scale 3*.) The nodes whose links are of the same color belong to the same group with similar semantics. Along the temporal dimension, we perform average pooling over features of adjacent frames to obtain temporal features of coarser scales at *Scale 2* and *Scale 3*. (Details of spatial-pooling are available in the appendix.).

paces, e.g., 'hand waving' may be performed by hand (joint-level), forearm (part-level), or the whole arm (limb-level) at different paces. We thus design cross-scale matching that learns semantic relevance by measuring feature consistency across spatial and temporal scales.

The contributions of this work are threefold. *First*, we formulate one-shot skeleton action recognition as an optimal matching problem and design an effective network framework for one-shot skeleton action recognition. *Second,*, we propose a multi-scale matching strategy that can capture scale-wise skeleton semantic relevance at multiple spatial and temporal scales. On top of that, we design a novel cross-scale matching scheme that can model the within-class variation of human actions in motion magnitudes and motion paces. To the best of our knowledge, this is the first work that exploits multi-scale representations and cross-scale matching to capture multi-scale skeleton semantic relevance and maintain consistency across motion scales in one-shot skeleton action recognition. *Third,* extensive experiments on three public datasets (NTU RGB+D, NTU RGB+D 120, and PKU-MMD) show that our method outperforms the state-of-the-art consistently by large margins.

II. RELATED WORKS

Skeleton-based Action Recognition has attracted increasing attention in recent years. Traditional methods design hand-craft features to represent skeleton sequences [13], [14], [15]. In recent years, deep learning has been widely explored by leveraging the advances in Recurrent Neural Networks [16], [17], [18], Convolutional Neural Networks [19], [20], [21], [22], Graph Convolution Networks (GCNs) [9], [23], [24], [25], Hypergraph Neural Networks [26], [27], [28], and Transformer-based methods [26], [29], [30]. Given the inherent topological graph structure of the human skeleton, GCNs have attracted increasing attention in skeleton-based action recognition. Notably, Yan et al. [25] proposed a spatial-temporal graph convolutional network to learn spatial-temporal patterns from skeleton data. Shi et al. [24] designed an adaptive graph convolutional network, which utilized self-attention with the spatial-temporal GCN. In a more recent development, Wang et al. [31] employed parameterized Koopman pooling, replacing average pooling in supervised skeleton action recognition.

Though the prior studies achieve very impressive performance, most of them are supervised and require large-scale training data which is often laborious to collect. We focus on one-shot skeleton action recognition, aiming to address the scenarios where only a few labeled skeleton samples are available for an unseen new class.

Few-Shot Learning: Motivated by the human capability in learning new concepts from just a few samples, few-shot learning, which aims at recognizing unseen concepts with only a few labeled training samples, has received increasing attention and witnessed significant advances in recent years. It has also been widely explored in the computer vision research community. For example, Snell et al. [32] presented the Prototypical Networks that compute distances between a datapoint and class-wise prototypes. Ye et al. [33] defined set-to-set transformations to learn a task-specific feature embedding for few-shot learning. Simon et al. [34] presented the DSN that employs a few-shot learning model via affine subspaces. Xu et al. [35] proposed to learn a novel dynamic meta-filter for few-shot learning.

One-shot Skeleton Action Recognition has attracted increasing interest in recent years. Leveraging the NTU RGB+D 120 dataset, Liu et al. [1] first presented an Action-Part Semantic-Relevance aware (APSR) approach for one-shot skeleton action recognition. Sabater et al. [4] presented a one-shot action recognition approach based on a Temporal Convolutional Network (TCN). Memmesheimer et al. [3] proposed to formulate the one-shot skeleton action learning problem as a deep metric learning problem. Additionally, Memmesheimer et al. [2] presented an image-based skeleton representation, which performs well in the deep metric learning manner. Ma et al. [36] proved that maximally preserving disentangled joint-level spatial features are beneficial to increase representation diversity and recognizability for few-shot classes in one-shot skeleton action recognition. Wang et al. [37] proposed JEANIE, which performs the joint alignment of temporal blocks and simulated viewpoint indexes of skeletons between support-query sequences to select the smoothest path without abrupt jumps in matching temporal locations and view indexes. In a more recent study, Wang et al. [38] introduced the uncertainty-DTW, which take into account the uncertainty of in frame-wise (or block-wise) features by selecting the path which maximizes the Maximum Likelihood Estimation (MLE).

Most existing methods take the skeleton representation as a whole and measure the skeleton similarity globally which often misses useful structure and temporal information. We propose to treat one-shot skeleton action recognition as an optimal matching problem and design multi-scale matching and cross-scale matching which capture the scale-wise semantic relevance and maintain the spatial and temporal consistency across different scales, respectively
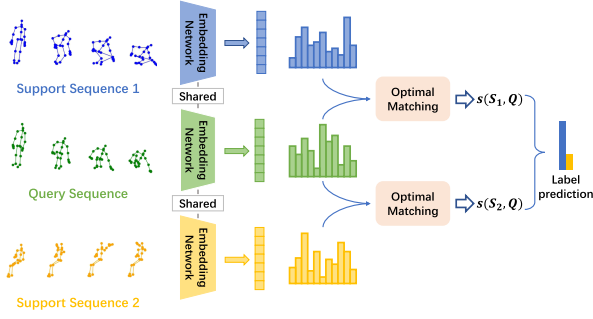
Fig. 3. Pipeline of the proposed method: Given a *Query Sequence* and two support skeleton sequences *Support Sequence 1* and *Support Sequence 2*, skeleton representations are first extracted with a weight-sharing embedding network which are further aligned progressively with the proposed *Optimal Matching*. The semantic relevance score (denoted by $s(\cdot, \cdot)$) between the query and the two support instances can then be computed for action prediction. The pipeline is illustrated with a 2-way 1-shot task. Different *Optimal Matching* strategies are provided in Fig. 4.

III. METHOD

We aim to train a model that can recognize human skeleton data of novel classes with only a single labeled sample. It is a very challenging task due to the very rich intra-class spatial-temporal variations in human skeleton action. We address this challenge by proposing a one-shot skeleton action recognition framework as illustrated in Fig. 3. Specifically, we design a novel optimal matching technique to capture the useful spatial structure and temporal order information which is largely neglected in most existing one-shot skeleton action recognition studies. In the following, we first present the problem formulation of the one-shot skeleton action recognition task. We then introduce the embedding network and elaborate on how to construct multi-spatial and multi-temporal scale skeletons. Finally, we describe the proposed optimal matching technique in detail. Key notations and definitions are summarized in Table XIII, available online, in the appendix.

A. Problem Formulation

Inspired by prior studies in few-shot learning [32], [33], [34], [35], [39] and few-shot video action recognition [40], [41], [42], [43], [44], [45], we formulate the one-shot skeleton action recognition task as a meta-learning problem [46] that consists of a meta-training phase and a meta-testing phase. In a *n*-way *1*-shot problem, each episode consists of a support set S and a query set Q, where S contains *1* labeled sample for each of n unseen classes and Q is employed to evaluate the generalization performance. The algorithm aims to determine which support classes each query sample belongs to. Specifically, multiple *n*-way *1*-shot tasks are randomly sampled from the *meta-training set* D_{train} (with seen classes), and employed to train a model in an episodic manner. In *meta-testing phase*, *n*-way and *1*-shot tasks are sampled from the *meta-testing set* D_{test} (with unseen classes) for evaluations.

B. Skeleton Feature Embedding

Following prior studies on few-shot learning [32], [33], [34], [35], [39] and few-shot video action recognition [40], [41], [42], [43], [44], [45], we first pre-train an embedding network on the whole *meta-training set* D_{train} using the cross-entropy loss for standard classification before proceeding to episodic training. We adopt the GCN-based model [24] as the baseline network which has an adaptive spatial-temporal graph for extracting the relation among body joints.

The GCN-based model just processes the features of the original scale. However, such single-scale modeling often misses meaningful skeleton information especially when only a single labeled sample is available as described in Section I. Inspired by [5], [6], [8], [9], [47] which handle multi-scale features, we represent the human skeleton data at multi-spatial and multi-temporal scales.

Multi-Spatial Scale Skeleton: We model skeleton actions at multiple spatial scales. Specifically, we adopt 3 spatial scales including the body-joint scale (s_1), the part-level scale (s_2), and the limb-level scale (s_3) as illustrated in Fig. 2(a). We first build GCN blocks on the first scale to capture joint-wise feature representations and then perform the average pooling (the details of spatial pooing can be found in Appendix D, available online). For skeleton-based representations, the pooling requires meaningful neighborhoods and we simply put joints of the same spatial scale into one group.

Multi-Temporal Scale Skeleton: Recognizing that consecutive frames capture continuous motions and poses reflecting analogous abstract states, we represent skeleton data across multiple temporal scales. After processing several GCN blocks at the original temporal scale, we incorporate two average pooling layers along the temporal dimension to perform temporal pooling as illustrated in Fig. 2(b) and Fig. 7, available online. Specifically, by applying average pooling to the features of consecutive frames from the original scale, we consolidate them into a unified feature to represent a 'new frame' in coarser scales, such as scale 2 or scale 3 As we apply average pooling with strides of 2 and 4, the features of two consecutive frames are pooled to produce the scale 2 features, while those of four consecutive frames are pooled to yield the scale 3 features.

We implement 3 spatial scales and 3 temporal scales for illustration, where each skeleton structure captures unique perspectives of skeleton representations. To extract the multi-scale skeleton representation, each stream within the multi-spatial and multi-temporal scale skeleton structures is individually optimized through the cross-entropy loss.[1]

C. Optimal Matching Strategy

Unlike [1], [2], [3], [4] that compute distances over sequence-level embeddings, we capture discriminative local information of each body joint and design a skeleton optimal matching scheme to compute semantic relevance based on optimal transport theory as discussed in Section I. Specifically, we adopt the Earth Mover's Distance (EMD) [12] as the optimal transport matching metric, which searches for the minimal cost transport plan between two joints' feature distributions by maximizing the impact caused by relevant joints and minimizing the effect between irrelevant joints.

Here, we first use the single-scale model as an example to show how we formulate the one-shot skeleton action recognition as the optimal matching problem by using EMD. The skeleton representation embedded by the single-scale model can be represented as $\mathbf{X} \in \mathbb{R}^{C \times N \times T}$, where C is the number of output channels, N denotes the number of skeleton joints, and T denotes the number of frames. For two feature maps $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{C \times N \times T}$, we first flatten them into two sets of joint's local representations $\mathcal{X} = \{x_i | i = 1, 2, \ldots NT\}$ and $\mathcal{Y} = \{y_j | j = 1, 2, \ldots NT\}$, where x_i and y_j ($x_i, y_j \in \mathbb{R}^C$) denote the joint's local representation at the corresponding spatial and temporal positions. Then we define the EMD between two sets of local representations as the minimum "transport cost" from $\mathcal{X}$ (suppliers) to $\mathcal{Y}$ (demanders). Suppose for each supplier x_i, it has r_i units to transport, and for each demander y_j, it requires c_j units. The overall optimal transport

---

[1]Detailed network structures can be found in the appendix, available online.

matching problem can be formulated as:

$$OT(r,c) = \left\{\pi \in \mathbb{R}^{NT \times NT} | \pi \mathbb{1} = \mathbf{r}, \pi^{\top} \mathbb{1} = \mathbf{c}\right\}, \quad (1)$$

where π is the optimal matching flow between these two distributions, which can also be viewed as the optimal matching plan of two skeleton sequences. r_i and c_j are called the weights of nodes, which control the total matching flows generated by each node, and $\mathbf{r}$ and $\mathbf{c}$ are vectorized representations of $\{r_i\}$ and $\{c_j\}$. EMD seeks an optimal matching flow π between "suppliers" $\mathcal{X}$ and "demanders" $\mathcal{Y}$, such that the overall matching cost can be minimized.

Additionally, the transporting cost per unit is defined by computing the pairwise distance between supplier node $\mathbf{x}_i$ and demander node $\mathbf{y}_j$ from two skeleton features:

$$d_{ij} = 1 - \frac{\mathbf{x}_i^{\top} \mathbf{y}_j}{\|\mathbf{x}_i\| \|\mathbf{y}_j\|}, \quad (2)$$

where nodes with similar local representations tend to generate small transporting costs between each other. Then we can define the EMD as the optimal transportation problem, which is represented as:

$$D_{emd}(\mathbf{X}, \mathbf{Y}) = \min_{\pi \in OT(r,c)} \sum_{i=1}^{NT} \sum_{j=1}^{NT} d_{ij} \pi_{ij}. \quad (3)$$

The weight of each node (*e.g.*, r_i and c_j) plays an important role in optimal matching problems. Intuitively, the node with a larger weight is more important while matching two sets. Therefore, in order to assign the more important node a higher weight, we follow [39] to generate the weight r_i by a cross-reference mechanism that uses the dot product between a joint representation and the average joint representation in the other skeleton features:

$$r_i = max \left\{ x_i^{\top} \cdot \frac{\sum_{j=1}^{NT} y_j}{NT}, 0 \right\}, \quad (4)$$

where x_i and y_j denote the feature vectors from two skeleton feature maps, and function $max(\cdot)$ ensures the weight is always non-negative. Above we take r_i as an example, and c_j can be calculated in the same manner. Once acquiring the optimal matching flow π, we can compute the semantic relevance score s between two skeleton representations as:

$$s(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{NT} \sum_{j=1}^{NT} (1 - d_{ij}) \pi_{ij}. \quad (5)$$

These semantic relevance scores allow studying the composition of the overall relevance, enabling us to assign high relevance to semantically similar joints no matter whether they are in the same spatial order or temporal frame. We can thus tackle the problem that the semantic relevance of two skeleton sequences can occur at different temporal positions or different spatial joints.

Multi-Scale Matching: As mentioned in Section III-B, human skeleton data can be represented in multi-spatial scales and multi-temporal scales, and each scale's representation contains unique semantic information. Thus, we propose to capture the pair-wise skeleton semantic relevance at multiple scales, including multi-spatial scale matching and multi-temporal scale matching, to acquire optimal matching flow from multiple spatial and multiple temporal scales.

For the multi-spatial scale scenario, there are three pairs of feature embeddings, which can be represented as $\mathbf{X}_{s1} \in \mathbb{R}^{C \times N \times T}$, $\mathbf{X}_{s2} \in \mathbb{R}^{C \times N_2 \times T}$, and $\mathbf{X}_{s3} \in \mathbb{R}^{C \times N_3 \times T}$, respectively. N_2 denotes the number of nodes for the second-scale spatial graph, and N_3 stands for the number of third-scale spatial graph nodes. The semantic relevance score
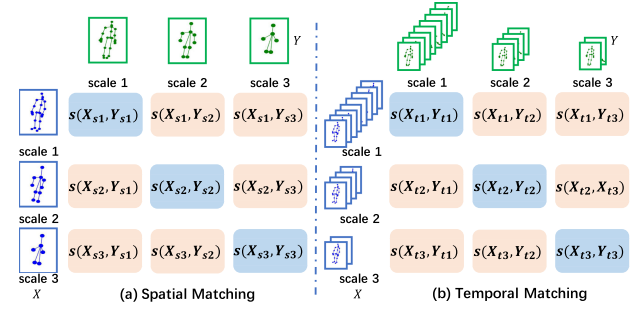

Fig. 4. Illustration of optimal matching in Spatial Matching in (a) and Temporal Matching in (b): In each of the two sub-figures, the three matching strategies along the diagonal (in blue color) illustrate the proposed multi-scale matching, and the rest 6 of the diagonal (in orange color) show the proposed cross-scale matching. Here, $s(\cdot, \cdot)$ represents the semantic relevance score between two skeleton features. X and Y stand for two skeleton sequences.

between two skeleton sequences thus becomes:

$$s_{ms}(\mathbf{X}, \mathbf{Y}) = s(\mathbf{X}_{s1}, \mathbf{Y}_{s1}) + s(\mathbf{X}_{s2}, \mathbf{Y}_{s2}) + s(\mathbf{X}_{s3}, \mathbf{Y}_{s3}). \quad (6)$$

This enables us to seek the optimal matching flow using EMD, and measure the semantic relevance at multiple spatial scales as shown in Fig. 4(a) (highlighted in blue color).

For the multi-temporal scale, there are also three pairs of feature embeddings, which can be represented as $\mathbf{X}_{t1} \in \mathbb{R}^{C \times N \times T}$, $\mathbf{X}_{t2} \in \mathbb{R}^{C \times N \times T/2}$, and $\mathbf{X}_{t3} \in \mathbb{R}^{C \times N \times T/4}$, respectively. The semantic relevance score between two skeleton sequences becomes:

$$s_{mt}(\mathbf{X}, \mathbf{Y}) = s(\mathbf{X}_{t1}, \mathbf{Y}_{t1}) + s(\mathbf{X}_{t2}, \mathbf{Y}_{t2}) + s(\mathbf{X}_{t3}, \mathbf{Y}_{t3}). \quad (7)$$

Using (7), semantic relevance between two skeleton sequences is measured at multiple temporal scales, as shown in Fig. 4(b) (highlighted in blue color).

Cross-Scale Matching: As discussed in Section I, different instances of the same action class may be performed at different magnitudes (spatial scales). For the class, such as 'hand waving', people may perform it by moving their palm only (joint-level), or by moving the forearm (part-level), or even by moving the whole arm (limb-level). Additionally, the same-category samples can also be performed at different speeds (temporal scale). Thus, there is also semantic relevance between different scales' skeleton representations for matching. To address the cross-scale matching challenge, we further investigate how to measure the semantic relevance between skeleton sequences across different scales, including cross-spatial scale and cross-temporal scale matching, considering the possibility of different spatial magnitudes and temporal speeds of the same action.

For cross-spatial scale matching, the three spatial scales' skeleton representations are $\mathbf{X}_{s1} \in \mathbb{R}^{C \times N \times T}$, $\mathbf{X}_{s2} \in \mathbb{R}^{C \times N_2 \times T}$, and $\mathbf{X}_{s3} \in \mathbb{R}^{C \times N_3 \times T}$. It can be seen that all these three scales' representations contain T frame features. Thus, we first perform 1D average pooling ($AvgPool$) on the spatial dimension to generate these three scales' features in the same shape ($\mathbb{R}^{C \times T}$), and then formulate the semantic relevance score as the T frame features optimal matching problem to acquire the optimal matching flow between different spatial-scale representations. The cross-spatial scale semantic relevance score can be represented as:

$$s_{cs}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{3} \sum_{j=1, j \neq i}^{3} s(AvgPool(\mathbf{X}_{si}), AvgPool(\mathbf{Y}_{sj})). \quad (8)$$

This process is shown in Fig. 4(a) (highlighted in orange color). In this way, we solve the problem of matching skeleton sequences with different motion magnitudes.

Furthermore, we also address the problem of matching the skeleton sequences with different motion speeds and design the cross-temporal scale matching. The three temporal scales' skeleton features are represented as $\mathbf{X}_{t1} \in \mathbb{R}^{C \times N \times T}$, $\mathbf{X}_{t2} \in \mathbb{R}^{C \times N \times T/2}$, and $\mathbf{X}_{t3} \in \mathbb{R}^{C \times N \times T/4}$. It can be seen that these three scales' representations all contain N joint features. Similarly, the 1D average pooling ($AvgPool$) can be performed on the temporal dimension to pool $\mathbf{X}_{t1}$, $\mathbf{X}_{t2}$, and $\mathbf{X}_{t3}$ into the same shape ($\mathbb{R}^{C \times N}$). Thus, the Earth Mover's Distance is used to measure the semantic relevance score across different temporal scales as:

$$s_{ct}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{3} \sum_{j=1, j \neq i}^{3} s(AvgPool(\mathbf{X}_{ti}), AvgPool(\mathbf{Y}_{tj})), \quad (9)$$

This process is shown in Fig. 4(b) (highlighted in orange color). Action sequences with different motion speeds can be matched well through (9).

Summary: As above mentioned, we first introduce the single-scale semantic relevance score $s(\mathbf{X}, \mathbf{Y})$ (5) that considers the useful spatial structure and temporal order information during matching. To address the problem that different instances of the same action class samples may be performed at different paces with different motion dynamics, we introduce 4 types of semantic relevance scores including multi-spatial scale ($s_{ms}(\mathbf{X}, \mathbf{Y})$, (6)), multi-temporal scale ($s_{mt}(\mathbf{X}, \mathbf{Y})$, (7)), cross-spatial scale ($s_{cs}(\mathbf{X}, \mathbf{Y})$, (8)), and cross-temporal scale ($s_{ct}(\mathbf{X}, \mathbf{Y})$, (9)). The semantic relevance score for our proposed model is averaging from multi-scale and cross-scale relevance scores, which is then used to predict the action category.

D. Objective Loss, Model Training, and Inference

The majority of current few-shot learning techniques [32], [33], [34], [35], [39] implement a pre-training stage prior to meta-learning. The effectiveness of this stage within the realm of few-shot learning has been validated by [48], [49]. Therefore, the proposed method is trained in two sequential stages: The *First* is the pre-training stage. The embedding network is trained on *meta-training set D_{train}* in a standard supervised learning way (Section III-B). The *Second* is the *meta-training* stage. The embedding network and our optimal matching method (Section III-C) are further optimized in an end-to-end manner, following [12], [50]. Both training stages leverage the Softmax cross-entropy loss as the classification loss L_{cls} for optimization (we experimented with the normalized Softmax cross-entropy loss and Angular margins as objective functions, and the Softmax cross-entropy loss yielded the best performance). Given an unseen query sequence q and its support set S (both sampled from *meta-testing set D_{test}*) at test, the goal is to determine which support set classes q belongs to.

IV. EXPERIMENTS

A. Datasets

NTU RGB+D dataset [51] consists of 56880 skeleton action sequences, which is the most widely-used dataset in skeleton-based action recognition research. In this dataset, action samples are performed by 40 subjects, with three camera views, and categorized into 60 classes. The NTU RGB+D dataset provides two standard evaluation protocols, namely cross-view (CV) and cross-subject (CS).

NTU RGB+D 120 dataset [1] is currently the largest dataset with 3D joints annotations for human action recognition. The dataset contains 114480 action samples in 120 action classes. Action samples are captured by 106 volunteers with three camera views. This dataset contains 32 setups, and each setup denotes a specific location and background. The evaluation protocols of this dataset are (1) cross-subject evaluation and (2) cross-setup evaluation.

PKU-MMD dataset [52] is a large-scale benchmark for continuous multi-modality 3D skeleton action understanding. It contains 21,545 action instances performed by 66 distinct subjects in 51 action categories. The dataset also utilizes CV and CS evaluation protocols. PKU-MMD consists of two subsets, part I and part II, we conduct experiments on the part I subset in this work.

B. Training and Evaluation Protocol

Training Protocol: As described in Section III-A, we formulate the one-shot skeleton action recognition problem as a meta-learning problem [32], [39], [41], [42], [46]. Therefore, we adopt the meta-training phase as our training protocol. All experiments are trained under the *5*-way, *1*-shot setting, and we set the number of query samples for each class to 15.

Evaluation Protocol 1: Similarly, we conduct the evaluation on *5*-way, *1*-shot setting, and adopt this procedure as Evaluation Protocol 1 in the following experiments.

Evaluation Protocol 2: To compare with the existing one-shot skeleton action recognition techniques [1], [2], [3], [4], [37], [38] in a fair way, we also follow the official one-shot protocol described in [1] for the dataset NTU RGB+D 120. Specifically, the testing set consists of 20 novel classes, and we pick one sample from each novel class as the exemplar[2] and leave the rest (except for the 20 exemplars) to test the recognition performance. For datasets NTU RGB+D and PKU-MMD[2], we adopt a similar protocol in experiments.

C. Implementation Details

Pre-training Stage: We use the SGD optimizer with Nesterov momentum (0.9) as the optimizer. The learning rate is set as 0.1 and is divided by 10 at 30_{th} epoch and 40_{th} epoch. The training process is ended at the 50_{th} epoch. For NTU RGB+D and NTU RGB+D 120, the batch size is set as 64. The batch size for the PKU-MMD dataset is 32.

Meta-learning Stage: The learning rate starts at 0.001 and decays every 10 epochs by 0.5. We train for 100 epochs using SGD optimizer, and each epoch consists of 100 episodes from the *meta-training set D_{train}*. For the meta-testing phase, we sample 1000 episodes from *meta-testing set D_{test}*. (D_{train} and D_{test} are defined in Section III-A).

Embedding Network: We use Adaptive Graph Convolutional Network [24] (AGCN) as our single-scale embedding network which has 9 GCN blocks. For multi-spatial and multi-temporal networks, the first 6 blocks are shared to capture the single-scale features. Then each scale feature is processed by the other 3 individually and in parallel.[3]

D. Dataset Splitting

For dataset NTU RGB+D 120, we adopt the one-shot skeleton action setting as described in [1] which splits the full dataset into a training set and a testing set. The action classes of the two sets are distinct which include 100 classes for training and 20 for testing.[2] For dataset NTU RGB+D, the training set and testing set are determined by the selection of 50 classes and 10 classes from the 100 training and 20 testing classes of the NTU RGB+D 120, respectively. Similarly, for

---

[2]Exemplars & dataset splitting details are provided in the appendix, available online.

[3]The detailed network structures can be found in appendix, available online.

TABLE I
ONE-SHOT SKELETON RECOGNITION EXPERIMENT UNDER
THE EVALUATION PROTOCOL 1

Method	NTU	NTU 120	PKU-MMD
ProtoNet [32]	78.3	80.3	84.7
FEAT [33]	77.8	80.0	83.8
Subspace [34]	77.9	80.5	84.2
Dynamic Filter [35]	79.3	80.4	84.9
S-scale (Ours)	80.4	81.2	85.7
M-scale (Ours)	82.6	83.5	88.2
M&C-scale (Ours)	**83.7**	**84.5**	**89.3**

S-scale: single-scale matching; M-scale: multi-scale matching; M&C-scale: multi-scale and cross-scale matching.

TABLE II
ONE-SHOT SKELETON RECOGNITION EXPERIMENTS UNDER
THE EVALUATION PROTOCOL 2

Method	NTU	NTU 120	PKU-MMD
Attention Network [53]	–	41.0	–
Fully Connected [53]	–	42.1	–
Average Pooling [17]	–	42.9	–
APSR [1]	–	45.3	–
TCN [4]	–	46.5	–
SL-DML [3]	–	50.9	–
Skeleton-DML [2]	–	54.2	–
uDTW [38]	72.4	49.0	–
JEANIE [37]	80.0	57.0	–
ALCA-GCN [54]	–	57.6	–
ProtoNet [32]	74.8	60.4	78.1
FEAT [33]	74.3	61.5	75.9
Subspace [34]	75.6	60.9	75.6
Dynamic Filter [35]	75.9	60.6	78.8
S-scale (Ours)	77.4	63.2	82.8
M-scale (Ours)	81.6	67.6	85.7
M&C-scale (Ours)	**82.7**	**68.7**	**86.9**

S-scale: single-scale matching; M-scale: multi-scale matching; M&C-scale: multi-scale and cross-scale matching.

TABLE III
COMPARISON OF DIFFERENT FEATURE EMBEDDING APPROACHES AND
DISTANCE METRICS UNDER EVALUATION PROTOCOL 2. (EMD: EARTH
MOVER'S DISTANCE)

Embedding	Metric	NTU	NTU 120	PKU-MMD
Global	Euclidean	71.6	56.1	78.1
Global	Cosine	74.8	60.4	74.8
Local	Euclidean	74.7	58.0	80.7
Local	Cosine	75.8	61.3	80.8
Local	EMD	**77.4**	**63.2**	**82.8**

EMD: earth mover's distance.

dataset PKU-MMD, we divide the action categories into a training set and a testing set which include 41 classes for training and 10 classes for testing.[2]

As no hold-out validation set is defined in the one-shot skeleton action setting and all these three datasets contain the cross-subject setting for supervised action recognition. Therefore, we divide the training class data into the training set and validation set based on the cross-subject principle for the one-shot skeleton action recognition task. We maintain the testing set of these three datasets as the testing set for one-shot skeleton action recognition.

E. Evaluating on One-Shot Skeleton Action Recognition

We conduct extensive experiments with five optimal matching strategies that include single-scale ('**S-scale**'), multi-spatial scale ('M_s'), multi-temporal scale ('M_t'), cross-spatial scale ('C_s'), and cross-temporal scale ('C_t'). We combine 'M_s' and 'M_t' to form a new multi-scale strategy '**M-scale**'. In addition, we combine 'M_s', 'M_t', 'C_s', and 'C_t' to form another new strategy '**M&C-scale**' that matches spatial and temporal features simultaneously at multiple scales and also cross scales.

We compare our method with two groups of state-of-the-art methods on one-shot skeleton action recognition. The first group consists of state-of-the-art few-shot image classification methods including Subspace [34], ProtoNet [32], Dynamic Filter [35], FEAT [33]. All these methods use the same embedding network as our method for fair comparisons. We re-implement [32], [33], [34], [35] based on publicly available codes and conduct experiments on NTU RGB+D, NTU RGB+D 120 and PKU-MMD datasets. The implementation details are available in the appendix, available online. The second group consists of state-of-the-art one-shot skeleton action recognition techniques including APSR [1], TCN [4], SL-DML [3], Skeleton-DML [2], uDTW [38], and JEANIE [37]. For those one-shot skeleton action recognition works [1], [2], [3], [4], [37], [38], the results in Tables I, II, and IV are from the original papers. We compare our method with the first group methods under both evaluation protocols, while the second group with Evaluation Protocol 2 only.

1) Evaluation Protocol 1: We evaluate the *5*-way *1*-shot setting on all three datasets and Table I shows experimental results. It can be seen that our proposed single-scale optimal matching outperforms state-of-the-art few-shot learning methods on all three datasets. In addition, our proposed multi-scale and cross-scale matching strategies further improve one-shot skeleton action recognition by large margins, demonstrating the effectiveness of our proposed method on the one-shot skeleton action recognition task.

2) Evaluation Protocol 2: Following the one-shot setting in [1], we also conduct experiments on NTU RGB+D, NTU RGB+D 120, and PKU-MMD datasets under Evaluation Protocol 2. Similar to the

experiments on Evaluation Protocol 1, our proposed method outperforms the state-of-the-art one-shot skeleton action recognition and few-shot learning methods by large margins (up to 8% on NTU120 and PKU-MMD). Table II shows more details of the experiments.

F. Ablation Studies

Matching Strategies: We compare our proposed optimal matching with the *global matching scheme* that adopts global average pooling to generate feature vector, as well as *a local matching scheme* that computes joint-to-joint distances (either Euclidean or Cosine distance) with local-level representations. For fair comparisons, we adopt the same backbone and training scheme for all compared methods, and Table III reports experiment results. We can observe that models with local-level representations perform better than models that rely on global-level representations (in globally pooled feature vectors). In addition, our method which works with the optimal matching flow between all pairs of joints outperforms all compared methods. Note that all experiments were conducted on 'S-scale' model under evaluation protocol 2.

Reducing Training Classes: While evaluating one-shot action recognition methods, one interesting question is how many training classes are required to achieve fair recognition performance. We examine this issue under Evaluation Protocol 2 by following prior studies on NTU RGB+D 120 dataset [1], [2], [3], [37], [38]. Table IV shows experimental results. It can be seen that our method outperforms the state-of-the-art by large margins under different numbers of training classes. With a training set of 60 classes, our method is on par with

TABLE IV
EXPERIMENTS ON DIFFERENT SIZES OF THE AUXILIARY TRAINING SET FOR
ONE-SHOT SKELETON RECOGNITION ON NTU RGB+D 120 DATASET

Train Classes	20	40	60	80	100
APSR [1]	29.1	34.8	39.2	42.8	45.3
SL-DML [3]	36.7	42.4	49.0	46.4	50.9
Skeleton-DML [2]	28.6	37.5	48.6	48.0	54.2
uDTW [38]	32.2	39.0	41.2	45.3	49.0
JEANIE [37]	38.5	44.1	50.3	51.2	57.0
ALCA-GCN [54]	38.7	46.6	51.0	53.7	57.6
ProtoNet [32]	36.7	45.7	54.4	55.5	60.4
FEAT [33]	37.6	44.2	52.2	55.4	61.5
Subspace [34]	37.7	45.3	54.2	55.5	60.9
Dynamic Filter [35]	34.5	43.0	54.0	52.9	60.6
S-scale (Ours)	37.9	46.5	54.6	58.7	63.2
M-scale (Ours)	41.2	52.6	59.0	62.4	67.6
M&C-scale (Ours)	**44.1**	**55.3**	**60.3**	**64.2**	**68.7**

S-scale:single-scale matching; M-scale: multi-scale matching; M&C-scale: multi-scale and cross-scale matching.

TABLE V
EVALUATION OF DIFFERENT COMBINATIONS OF OPTIMAL MATCHING
MANNERS UNDER THE EVALUATION PROTOCOL 2: THE SECOND LINE SHOWS
THE BASELINE RESULT UNDER THE SINGLE-SCALE MATCHING MANNER

M_t	M_s	C_t	C_s	NTU	NTU 120	PKU-MMD
				77.4	63.2	82.8
✓				79.3	65.0	85.0
	✓			79.6	65.1	84.9
✓	✓			81.6	67.6	85.7
✓		✓		80.7	66.7	85.7
	✓		✓	80.8	67.3	86.0
✓	✓	✓	✓	**82.7**	**68.7**	**86.9**

M_t: multi-temporal scale matching; M_s: multi-spatial scale matching; C_t: cross-temporal scale matching; C_s: cross-spatial scale matching.

TABLE VI
EVALUATION OF OUR PROPOSED MULTI-SCALE MATCHING WITH
DIFFERENT SCALE COMBINATIONS

scales	Multi-Spatial Scale	Multi-Temporal Scale
1	63.2	63.2
1,2	64.3	64.4
1,3	64.6	64.2
1,2,3	**65.1**	**65.0**

The experiments were conducted on NTU RGB+D 120 dataset under the Evaluation Protocol 2.

state-of-the-art methods that are trained by using 100-class training set. This clearly shows the effectiveness of our proposed optimal matching.

Effect of Multi-Scale and Cross-Scale Matching Manners: We conduct experiments on different combinations of our proposed optimal matching strategies. Table V shows experimental results under evaluation protocol 2. We can see that including any of our proposed matching strategies (M_s, M_t, C_s, and C_t) improves the one-shot skeleton action recognition clearly. Including all four matching strategies perform simply the best over all three datasets, demonstrating the effectiveness of our proposed optimal matching technique. Please refer to the appendix for more ablation studies and visualization of the proposed multi-scale and cross-scale optimal matching strategies.

Effect of Multiple Spatial Scales and Multiple Temporal Scales: We also study how models with different spatial and temporal scales (as the backbone) perform for the one-shot skeleton action recognition task. Table VI shows the one-shot learning performance with different combinations of scales. We can observe that the model performs the best when combining scales 1, 2, and 3. In addition, employing two scales (scales 1 and 2 or scales 1 and 3) also outperforms the model

using scale 1 only, showing the benefits of the proposed multi-scale representations.

More ablation studies about temporal pooling strategies, feature pooling strategies and feature extractors can be found in the Appendix.

V. CONCLUSION AND FUTURE WORKS

In this paper, we address the one-shot skeleton action recognition as a matching problem. We obtain the multi-spatial and multi-temporal scale features by designing a hierarchical pooling that represents the same skeleton sequence at various spatial and temporal scales. Moreover, based on the multi-scale skeleton features, we propose a multi-scale skeleton matching strategy and a cross-scale skeleton matching manner to measure the semantic relevance between two skeleton sequences for one-shot skeleton action recognition. The experiments demonstrate that our proposed method achieves superior one-shot skeleton action recognition performance. At present, our proposed method is applied specifically to the Kinect V2 skeleton format. In our upcoming research, we plan to broaden our scope to include one-shot skeleton action recognition using various skeleton formats, possibly even exploring recognition across different formats. Furthermore, we also consider integrating one-shot skeleton action recognition with the large language models, leveraging their vast reservoir of prior knowledge.

REFERENCES

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

[2] R. Memmesheimer, S. Häring, N. Theisen, and D. Paulus, "Skeleton-DML: Deep metric learning for skeleton-based one-shot action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 837–845.

[3] R. Memmesheimer, N. Theisen, and D. Paulus, "SL-DML: Signal level deep metric learning for multimodal one-shot action recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 4573–4580.

[4] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, and A. C. Murillo, "One-shot action recognition in challenging therapy scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 2777–2785.

[5] Y. Cai et al., "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2272–2281.

[6] S. Yang, J. Liu, S. Lu, M. H. Er, and A. C. Kot, "Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 769–786.

[7] S. Yang, J. Liu, S. Lu, E. Hwa, Y. Hu, and A. C. Kot, "Self-supervised 3D action representation learning with skeleton cloud colorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 509–524, Jan. 2024.

[8] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1113–1122.

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

[10] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2292–2300.

[11] C. Villani, *Optimal Transport: Old and New*. Berlin, Germany: Springer, 2009.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[13] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.

[14] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 588–595.

[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3D human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 914–927, May 2014.

[16] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1110–1118.

[17] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[18] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 816–833.

[19] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. IAPR Asian Conf. Pattern Recognit.*, 2015, pp. 579–583.

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

[21] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2017, pp. 597–600.

[22] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 786–792.

[23] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 183–192.

[24] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12018–12027.

[25] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7444–7452.

[26] L. Wang and P. Koniusz, "3Mformer: Multi-order multi-mode transformer for skeletal action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5620–5631.

[27] Y. Zhu, G. Huang, X. Xu, Y. Ji, and F. Shen, "Selective hypergraph convolutional networks for skeleton-based action recognition," in *Proc. Int. Conf. Multimedia Retrieval*, 2022, pp. 518–526.

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

[29] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "Star-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3330–3339.

[30] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proc. Int. Conf. Mach. Learn.*, 2021, Art. no. 4.

[31] X. Wang, X. Xu, and Y. Mu, "Neural Koopman pooling: Control-inspired temporal dynamics encoding for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10 597–10 607.

[32] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[33] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8808–8817.

[34] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4136–4145.

[35] C. Xu et al., "Learning dynamic alignment via meta-filter for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5182–5191.

[36] N. Ma et al., "Learning spatial-preserved skeleton representations for few-shot action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 174–191.

[37] L. Wang and P. Koniusz, "Temporal-viewpoint transportation plan for skeletal few-shot action recognition," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 307–326.

[38] L. Wang and P. Koniusz, "Uncertainty-DTW for time series and sequences," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 176–195.

[39] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12200–12210.

[40] M. Bishay, G. Zoumpourlis, and I. Patras, "TARN: Temporal attentive relation network for few-shot and zero-shot action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2019, Art. no. 154.

[41] K. Cao, J. Ji, Z. Cao, C.-Y. Chang, and J. C. Niebles, "Few-shot video classification via temporal alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 618–10 627.

[42] H. Zhang, L. Zhang, X. Qi, H. Li, P. H. S. Torr, and P. Koniusz, "Few-shot action recognition with permutation-invariant attention," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 525–542.

[43] X. Wang et al., "Hybrid relation guided set matching for few-shot action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 948–19 957.

[44] A. Thatipelli, S. Narayan, S. Khan, R. M. Anwer, F. S. Khan, and B. Ghanem, "Spatio-temporal relation modeling for few-shot action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 958–19 967.

[45] L. Zhu and Y. Yang, "Compound memory networks for few-shot video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 751–766.

[46] O. Vinyals et al., "Matching networks for one shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3637–3645.

[47] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[48] W. Y. Chen et al., "A closer look at few-shot classification," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HkxLXnAcFQ

[49] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9062–9071.

[50] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12 203–12 213.

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

[52] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 2, pp. 1–24, 2020.

[53] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1647–1656.

[54] A. Zhu, Q. Ke, M. Gong, and J. Bailey, "Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6038–6047.

[55] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13 359–13 368.