Adaptive Local-Component-aware Graph Convolutional Network for One-shot Skeleton-based Action Recognition

Anqi Zhu[1], Qiuhong Ke[3], Mingming Gong[2], James Bailey[1]
[1]School of Computing and Information Systems, The University of Melbourne
[2]School of Mathematics and Statistics, The University of Melbourne
[3]Department of Data Science & AI, Monash University
azzh1@student.unimelb.edu.au, qiuhong.ke@monash.edu, {mingming.gong, baileyj}@unimelb.edu.au

Abstract

Skeleton-based action recognition receives increasing attention because skeleton sequences reduce training complexity by eliminating visual information irrelevant to actions. To further improve sample efficiency, meta-learning-based one-shot learning solutions were developed for skeleton-based action recognition. These methods predict by finding the nearest neighbors according to the similarity between instance-level global embedding. However, such measurement holds unstable representativity due to inadequate generalized learning on the averaged local invariant and noisy features, while intuitively, steady and fine-grained recognition relies on determining key local body movements. To address this limitation, we present the Adaptive Local-Component-aware Graph Convolutional Network, which replaces the comparison metric with a focused sum of similarity measurements on aligned local embedding of action-critical spatial/temporal segments. Comprehensive one-shot experiments on the public benchmark of NTU-RGB+D 120 indicate that our method provides a stronger representation than the global embedding and helps our model reach state-of-the-art.

1. Introduction

Action recognition is one of the computer vision problems that are practically important for realizing modern applications such as auto surveillance systems [31], video retrieval [7], etc. Past research majorly focuses on RGB-based inputs due to their wide accessibility. Yet, pixel-based inputs have a high risk of information over-richness, making a model easily confused by the task-irrelevant background, brightness, and color changes [28]. A 3-D Skeleton Sequence becomes one of the strong input alternatives [24] because it only records 3-d body joint movements along temporal evolution [3, 13, 26].

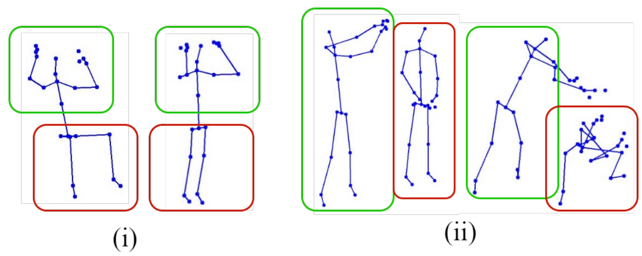

Figure 1. Frame examples for "Putting on a cap" (i) and "Hitting another person with something" (ii) in *NTU-RGB+D 120* [13]. To fastly adapt to a valid recognition, we expect the model to prioritize the similarity discrimination for the movements in the green blocks and suppress the noisy features from the red blocks.

Using deep learning models and vast annotated training samples, existing skeleton-based solutions implemented highly-accurate classification for pre-known activities. But the study on extending prediction to foreign classes is still at the beginning stage. Known as few-shot learning (FSL), it is an active research topic to realize fast adaptation to the classification for new classes with sparse direct supervised examples. Especially, the situation is called a one-shot learning (OSL) problem when only one example is available for each new class. The solutions can effectively help the models overcome their dependency on data-intensive training and enable one-time/rare-case learning that is often more suitable for real-world scenarios [36].

Among current skeleton-based OSL solutions, early studies first transformed skeleton sequences to signal images to solve as unified image-based classification [19, 20]. But the transformation deforms the original skeletal structure and thus causes information loss on the spatial connectivities between neighboring body joints. A more preferable way is to let a model achieve a metric-based OSL by comparing the similarity of the inputs' embeddings by native skeleton-based backbones. The solutions in [13, 25] respectively achieved different implementations for this but their results are still inferior to the signal-based solutions. During their classification process, encoded features are averagely pooled to generate a single global embedding as an input's representation for similarity comparison. With adequate supervised training, such representation could self-accumulate enough distinction learning for latent action-invariant features at a global scope. However, for the sparse examples from few-shot new classes, the generality of their global embedding can be easily biased with local features and fails to robustly focus on the necessary invariant features for refined recognition. On the other hand, we observe that specific local features under body-part-based partitions or consecutive time sections intuitively separate action-critical/irrelevant patterns for valid/invalid recognition clues. As shown in Fig. 1, while generalizing an abstract global pattern across all body joints and frames is difficult, a linear combination of aligned discrimination on the patterns in the green boxes should help quickly determine a good recognition. Similarly, the patterns in the red boxes are apparently non-action relevant and the measurements on their embedding become learning noises that should be directly suppressed to reduce representation bias.

In this paper, we propose Adaptive Local-Component-aware Graph Convolutional Network (ALCA-GCN) as the first metric-based method that relies on local embedding distancing as the main determinant factor for one-shot skeleton-based action recognition. It decomposes the instance-level similarity comparison to a selective sum of local measurements for every body part under each time section. To achieve this, we start with an encoding backbone that extracts hierarchical spatial-temporal features for both body-part-level patterns and skeleton-level contexts. Our embedding function then performs average pooling over the encoded features, which generates the independent comparing unit representation for each segmented local component. When calculating the total similarity, our model sequentially aggregates the embedding distance of each aligned unit between the given support and query inputs, and applies an adaptive emphasis/suppression for the decision impact from the action-critical/noisy units. We evaluate our solution on *NTU-RGB+D 120* [13] and use the official OSL testing protocol to compare with all previous related papers [13, 19, 20, 25]. The result proves that our model achieves state-of-the-art performance. Concretely, our contributions are:

- We propose ALCA-GCN as a novel metric-based OSL solution for skeleton-based action recognition. It models an action as a matrix of local comparable units on both spatial (body parts) and temporal dimensions (average time sections).
- ALCA-GCN determines the total similarity between two skeleton sequences by a selective sum of embedding distances between all aligned comparing units.
- During the similarity aggregation, ALCA-GCN self-learns the emphasis/suppression against the comparison importance from action-critical/irrelevant units. The model presents better results than the prior art under an extensive one-shot learning experiment setup using *NTU-RGB+D 120*.

2. Related Work

2.1. RGB-based Image/Video FSL

Being the primary experimental ground for FSL, many solutions have been developed for image classification, systematically divided into data-based, model-based, and meta-learning-based approaches [36]. Especially, as one of the meta-learning-based methods, metric-based learning is sharply focused due to its simple structure and flexible component scalability. In 2017, Snell *et al*. [27] proposed one of its major frameworks, which classifies according to the nearest Euclidean distances between queries and class prototypes in a well-generalized common embedding space. To align the training feature distribution to true few-shot testing tasks, Vinyals *et al*. [34] devised the episodic learning strategy which trains a model by a multi-tasking learning procedure. During a training epoch, each sub-task simulates the same N-way-K-shot setting from the testing conditions (i.e. having K reference instances for N candidate classes). Based on the above learning framework, various solutions are further designed to improve the few-shot generalization ability for each component, including enriching embedding features with external knowledge (*e.g*. semantic [4]), devising local-descriptor-based similarity matching [11, 39], empowering learning ability to metric functions [9, 21], etc.

FSL for RGB-video-based action recognition requires additional learning on the temporal dimension. Tan and Yang [29] first regarded this as a variant of image classification by compressing the input videos into static dynamic images. It was not until the breakthrough of deep volumetric extraction tailored for video features [8, 32], that many papers [6, 37, 40, 41] started to adopt the new backbones with common FSL frameworks. The difficulty of generalizing a video-based embedding space comes from the exponential increase of sample variance and backbone volume [33] due to the dimension expansion. Thus recent papers started to look for alternatives that figure the total similarity of extracted features in non-parametric ways. In 2021, Ben-Ari *et al*. [1] presented a metric-based solution in which the similarity between a query and generated class prototypes is measured by the sum of feature differences in averagely divided time sections. Cao *et al*. [2] applied Dynamic Temporal Warping (DTW) to orderly aggregate the similarity between the closest embedding match for every frame in two videos. While such local comparison is temporally decomposable according to a sub-action order, pixel-based inputs are hard to define meaningful and stationary geographical partitions for spatial local features of a frame. On the other hand, a performer's skeletal description maintains an invariant graph structure over time, persisting with explicit component meaning for each body part.

2.2. Skeleton-based Action Recognition

Research on the deep feature extraction of skeleton sequences is steadily developed in the past few years. Regarding action recognition as a temporal modeling problem, early solutions adopted RNN/LSTM-based extraction by sequentially feeding frame-level body joint data and predicting according to the accumulated learning status in the last frame [12, 14]. [35] refined spatial encoding with another parallel RNN model by establishing sequential connectivity among body joints according to a pre-defined traversing path. Since recursive networks are not competitive for their spatial modeling ability, [5] replaced the encoder with a CNN that parallelly convolutes the features from linearly-arranged adjacent body joints and frames. In 2018, research reveals that the kinetic dependencies in body joints' native multi-neighbor connections transmit more integral and abundant spatial information. Yan *et al*. [38] proposed a Spatial-Temporal Graph Convolutional Network (ST-GCN), which supports temporal and spatial feature convolution based on an adaptive multi-neighbor sampling scheme. The original scheme only convolutes spatial features from body joints' global relative distances to the body center. In 2019, Li et al. [10] proposed a more diversified dependency learning from the hybrid relations of multi-hop local natural connections and action-based inference connections. Parsa et al. [22] raised a cascaded pyramidal architecture to additionally capture feature correlation at a body-part-level scope and average the predictions at different granularity. In our work, we also designed a hierarchical but parallel convolution to obtain independent and richer representation for local partitions from their own information and skeletal contexts.

2.3. FSL in Skeleton-based Action Recognition

In the existing solutions, the topic is first tackled by Liu *et al*. [13], who implemented a Euclidean-distance-based similarity comparison on a sharing global embedding space by an ST-LSTM backbone [14]. To compensate for the generalization insufficiency, it imports external pre-trained knowledge on the semantic relations between body joint names and instance labels. The relevance scores re-assign the contribution weights of body-joint-level features towards global embedding to emphasize class-related learning. [25] convolutionally extracted features from the normalized body-joint coordinate average for each frame, and regarded the features from the last frame as the representation for instance-level similarity comparison. [19, 20] devised a pre-processing module that transforms skeleton sequences into signal images by lining up body joints as rows and frames as columns. To maintain a matrix format, the compression has to diminish the parallel neighboring relations between multiple body joints into linear connections, which brings information loss. [18] recently proved that maximally preserving disentangled joint-level spatial features is beneficial to increase representation diversity and recognizability for few-shot classes. Trading off between the clustering consistency of pooling embedding and the information richness of local encoding, our method decomposes the similarity metric to a sum of local component measurements. Additionally, it adaptively emphasizes discrimination learning on action-critical areas, while removing comparing diversity and bias brought by strong but action-irrelevant local features.

3. Method

Fig. 2 presents the architecture of ALCA-GCN. Our method follows the basic framework of metric-based solutions, consisting of an encoding backbone, an embedder for modeling the representation matrix, and a linear-metric-based classifier. To remove viewpoint variances, we first unite all inputs to a frontal viewing angle (see Section 4.2). We adopt an ST-GCN [38] network F as the prototype encoding backbone because it allows spatial feature convolution over each body joint's multiple neighbors following a pre-defined (sub-)structure. Maintaining its original skeleton-level convolution, we additionally devise independent kernel matrices for body-part-level convolution. The modified F now captures each joint's surrounding features under its belonging body part and relative global relations at the skeleton scope. We use F to obtain the total feature $\mathbf{f}$ from a pre-processed input $\mathbf{x}$ and pool it as a part-based global representation $\mathbf{G_f}$, which contains a group of local embedding $\mathbf{g_f}$ for 4 body parts (head, hands, torso and legs) under 3 temporal sections. They are regarded as the basic comparing units for local similarity matching. To enhance/suppress the distinction learning on action-critical/irrelevant components, an Adaptive Dependency Learning (ADL) module is attached behind $\mathbf{G_f}$ to adaptively adjust each unit's content influence to the global matching result. An average global embedding is also aggregated into each unit as an instance-level constraint to impose intra-class clustering consistency. Finally, trained with the Euclidean distance sum between all aligned units of the given query and its belonging class support example, the model learns to classify with a strong reliance on the similarity of action-critical local patterns. The differences from the noisy units are amended by supplementing their original embedding with other high-attention contexts. The rest

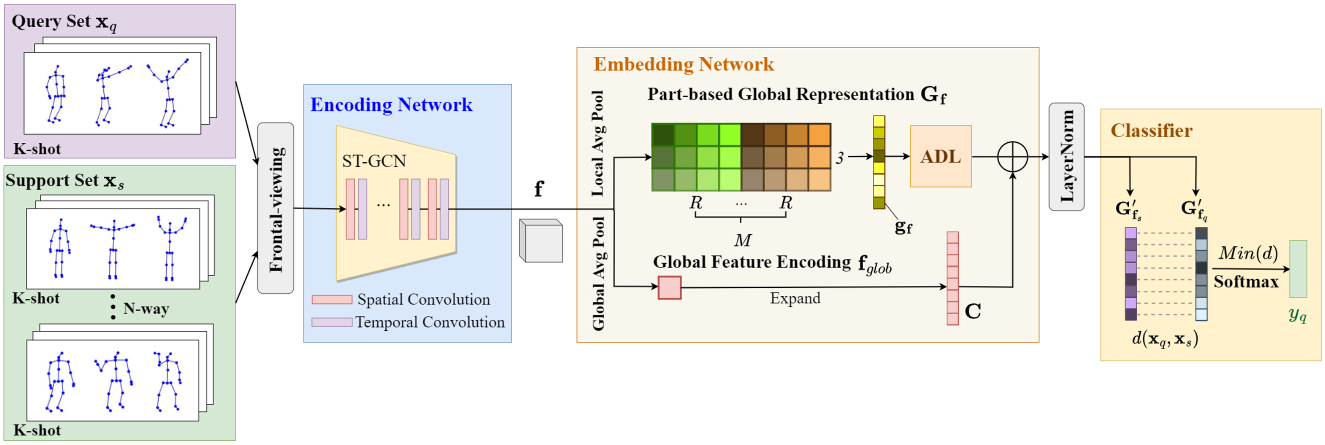

Figure 2. The overview of ALCA-GCN. Each input $\mathbf{x}$ is first pre-processed as frontally viewed. The encoder F applies two types of convolution on $\mathbf{x}$ for body-part-level surrounding and skeleton-level contextual features. The embedding network then partially pools the encoded feature $\mathbf{f}$ to generate the embedding of comparing unit $\mathbf{g_f}$ for R body parts under 3 time sections per each skeleton. Concatenating all $\mathbf{g_f}$ together forms the complete global representation $\mathbf{G_f}$ for describing $\mathbf{x}$. The ADL module highlights the contextual impact from action-critical units based on a self-attention mechanism, and the instance-level constraint $\mathbf{f}_{glob}$ is aggregated to each modified unit to impose intra-class clustering. Eventually, the total similarity between a support sample $\mathbf{x_s}$ and a query sample $\mathbf{x_q}$ is determined by the element-wise local measurements on their representation $\mathbf{G}'_{\mathbf{f}_q}$ and $\mathbf{G}'_{\mathbf{f}_s}$.

of the section further describes the details of each model component.

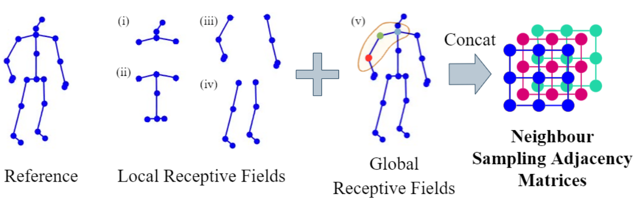

Figure 3. Neighbor sampling scheme for spatial convolution in ALCA-GCN. Local receptive fields cover the feature convolution on 4 body-part-based neighbor areas, including (i) head, (ii) torso, (iii) hands, and (iv) legs. The limbs are symmetrically grouped to avoid handedness disparities. The global receptive field (v) from [38] convolutes on every joint's root, centrifugal, and centripetal neighbors (figured by their global distances to the body center).

3.1. Encoding architecture

We denote an input skeleton sequence as $\mathbf{x}_{orig} \in \mathbb{R}^{C\times T\times U\times M}$, where M refers to the number of performers, U refers to the number of body joints for one performer, T refers to the number of frames, and C refers to the number of body joint coordinate dimension (usually is 3). Obtaining the adjusted input $\mathbf{x} \in \mathbb{R}^{C\times T\times U\times M}$ from the frontal-viewing pre-process, F first encodes it as M individual sequences $\mathbf{x}' \in \mathbb{R}^{C\times T\times U}$ and later concatenates them back before feeding them to the embedding network.

In the original ST-GCN [38], $\mathbf{x}'$ is represented as a graph of $\{\mathcal{V}, \mathcal{E}\}$ that includes every body joint from all frames. Concretely, $\mathcal{V} = \{v | v \in \mathbb{R}^C\}$ and $|\mathcal{V}| = U \times T$. $\mathcal{E}$ consists of the physical and temporal connections as $\mathcal{E} = \{(v_i, v_j)|(v_i, v_j) \in \mathcal{H}\} \cup \{(v_i^t, v_i^{t+1})|t \in (0, T)\}$ where $\mathcal{H}$ includes all naturally adjacent body joint pairs in each frame, and (v_i^t, v_i^{t+1}) refers to the temporal pair connection of the same body joint v_i between the adjacent frame t and $t+1$. Feature extraction is applied via iterative spatial and temporal graph convolutions. A spatial convolution layer contains two groups of learnable matrices $\{\mathbf{W}\}_k^{K_\mathcal{S}}$ and $\{\mathbf{E}\}_k^{K_\mathcal{S}}$. Given an input $\mathbf{f}' \in \mathbb{R}^{C'\times T'\times U}$ from a previous or input layer, for its spatial sub-feature $\mathbf{f}'_{in} \in \mathbb{R}^{C'\times U}$ at each $t' \in T'$, its convolution is calculated as:

$$\mathbf{f}'_{out} = \sum\nolimits_k^{K_\mathcal{S}} \mathbf{W}_k (\mathbf{f}'_{in} \times (\mathbf{A}_k \odot \mathbf{E}_k)) \,, \tag{1}$$

$$\mathbf{A}_k = \mathbf{\Lambda}_k^{-\frac{1}{2}} \bar{\mathbf{A}}_k \mathbf{\Lambda}_k^{-\frac{1}{2}} \,, \tag{2}$$

$$\mathbf{\Lambda}_k^{mn} = \begin{cases} \sum_j^{\mathcal{B}_k^i} (\bar{\mathbf{A}}_k^{mj}), & \text{if } m = n \\ 0, & \text{otherwise} \end{cases} , \tag{3}$$

where $K_\mathcal{S} = L \times R$ is the total number of applied convolutions. $\mathbf{W}_k$ is the k-th convolution matrix with a shape of $C'' \times C' \times 1 \times 1$, where C'' is the layer output dimension. $\mathbf{E}_k$ is a $U \times U$ matrix for re-weighting the neighbor features filtered by $\mathbf{A}_k$. $\odot$ is a dot product operation. $\mathbf{\Lambda}_k$ is the degree matrix of $\bar{\mathbf{A}}_k$ to apply degree normalization in Equation (2). $\bar{\mathbf{A}}_k$ is a pre-defined $U \times U$ adjacency matrix for convolution k, where $\bar{\mathbf{A}}_k^{ij}$ indicates whether v_j belongs to the convoluting area $\mathcal{B}_k^i$ for v_i. ST-GCN regards a skeleton as a single complicated graph that spreads around a body center joint [38]. By measuring the distances to the spine, it categorizes each joint's physical neighbors into three global relation types R, known as Centrifugal, Centripetal and Root (i.e. itself). To apply a one-kernel convolution, $\bar{\mathbf{A}}_r$ records every joint's adjacency status of whether it is a neighbor of type r ($r \in R$) for each body joint. The convolution for each joint is then sampled from its corresponding R neighbors orderly filtered by the matrix multiplication with $\{\bar{\mathbf{A}}_r | r \in R\}$.

Since our model focuses on the class-level representability under local sub-areas (body parts), except for the joint features from the global position relations to the body center, we also value each joint's relative surrounding features under its belonging body part. We get inspiration from [30] and devise extra $\bar{\mathbf{A}}$ to respectively filter a body joint's neighbor features from all local connections under its belonging body part (see Fig. 3). For a certain body part $\mathcal{P}_r$, $\bar{\mathbf{A}}_r$ filters the convoluting area $\mathcal{B}_r^i$ of v_i by:

$$\mathcal{B}_r^i = \{v_j | \mathbf{d}(v_i, v_j) \leq 1, v_i, v_j \in \mathcal{V}_{\mathcal{P}_r}\}\ , \tag{4}$$

$$\bar{\mathbf{A}}_r^{ij} = \begin{cases} 1, & \text{if } v_j \in \mathcal{B}_r^i \\ 0, & \text{otherwise} \end{cases}\ , \tag{5}$$

where $\mathbf{d}(v_i, v_j)$ denotes the minimum path between v_j and v_i. $\mathcal{V}_{\mathcal{P}_r}$ is the set of body joints included in $\mathcal{P}_r$. Edging joints between any two adjacent body parts are overlappingly included in both partitions, so that all $\bar{\mathbf{A}}_r$ for local sampling cover every natural skeletal connection. A spatial convolution layer eventually applies L-kernel groups of convolution on 4 body-part sub-graphs and 1 global skeleton graph. Thus there would be overall $K_{\mathcal{S}} = L \times 5$ parallel convoluting operations, and the weights for each convolution are learned with its own $\mathbf{W_k}$ and $\mathbf{E_k}$.

For a temporal convolution layer, given an input $\mathbf{f}'$ from its previous spatial layer, we remain a 3×1 convolution [38] on its temporal sub-feature $\mathbf{f}'_{in} \in \mathbb{R}^{C' \times T'}$ at each $v_i \in \mathcal{V}$. The convoluting area for $v_i^{t'}$ is the sub-features of v_i at $t'-1$ and $t'+1$.

At the end, F outputs $\mathbf{f} = F(\mathbf{x}) \in \mathbb{R}^{d_{feat} \times T_{feat} \times U \times M}$ after concatenating back M performers' features. d_{feat} and T_{feat} are the encoding sizes on spatial and temporal dimensions for each body joint of every performer.

3.2. Part-based Global Representation

Having $\mathbf{f}$, we apply a segmented mean pooling on the corresponding body joints and temporal dimensions to get the local embedding $\mathbf{g}_{\mathcal{P}_{ri}^m}$ for each body part $\mathcal{P}_r$ of performer m at a temporal section i. Beyond the same body part partitioning in Fig. 3, we averagely divide 3 temporal sections, known as the starting, middle, and ending phases. Therefore:

$$\mathbf{g}_{\mathcal{P}_{ri}^m} = \frac{1}{|\mathcal{V}_{\mathcal{P}_r}||T_i|} \sum\nolimits_v^{\mathcal{V}_{\mathcal{P}_r}} \sum\nolimits_t^{T_i} \mathbf{f}_{vt}^m\ , \tag{6}$$

$$T_i = \{t | t \in ((i-1) \times T_{div}, i \times T_{div}]\}\ , \tag{7}$$

where $R = 4$ is the number of partitioned body parts for a performer, and $T_{div} = T_{feat}/3$ is the length of each temporal section. All $\mathbf{g}_{\mathcal{P}_{ri}^m}$ are then concatenated to generate the matrix $\mathbf{G_f}$ as the instance-level global representation for $\mathbf{x}$.

To obtain the similarity metric between two inputs, we regard each $\mathbf{g}_{\mathcal{P}_{ri}^m}$ as the embedding of a local comparing unit, and successively aggregate the Euclidean distances between each aligned unit in the two object sequences. In a one-shot learning scenario and if using the episodic learning algorithm [34], for each epoch, the model meta-trains from a batch of sub-tasks randomly sampled from the auxiliary set. Each sub-task has the same N-way-1-shot setting consistent with the testing task. Having an incoming training/testing query input $\mathbf{x}_q$ and some support instances $\{(\mathbf{x}_{s_1}, s_1), (\mathbf{x}_{s_2}, s_2), ..., (\mathbf{x}_{s_P}, s_P)\}$ for candidate classes $s_1, s_2, ..., s_P$, the classification of $\mathbf{x}_q$ is the same category as its most similar support instance $\mathbf{x}_{s_{min}}$ according to the comparing metric. In other words, the model predicts the probability distribution of $\mathbf{x}_q$ belonging to class s_n via:

$$p_\phi(y_q = s_n | \mathbf{x}_q) = \frac{\exp(-d(\mathbf{x}_q, \mathbf{x}_{s_n}))}{\sum_{p=1}^{P} \exp(-d(\mathbf{x}_q, \mathbf{x}_{s_p}))}\ , \tag{8}$$

$$\begin{aligned} d(\mathbf{x}_q, \mathbf{x}_{s_p}) &= d\langle \mathbf{G}_{\mathbf{f}_q}, \mathbf{G}_{\mathbf{f}_{s_p}} \rangle \\ &= \sum\nolimits_j^{3 \times R \times M} \|\mathbf{g}_{\mathbf{f}_q}^j - \mathbf{g}_{\mathbf{f}_{s_p}}^j\|_2 \end{aligned}\ , \tag{9}$$

for all $n = s_1, s_2, ..., s_P$ with the model parameter ϕ. $d(\cdot, \cdot)$ refers to the similarity distance between the two comparing instances. $d\langle \cdot, \cdot \rangle$ is the actual metric function to calculate the total similarity aggregation between their global representation, which is the sum of Euclidean distances between every pair of aligned comparing unit $\mathbf{g}_{\mathbf{f}_q}^j$ and $\mathbf{g}_{\mathbf{f}_{s_p}}^j$. During training, the model optimizes ϕ by a negative log-probability loss of $\mathcal{L}_\phi = -\log p_\phi(y_{q_{train}} = \mathrm{y}_{q_{train}} | \mathbf{x}_{q_{train}})$ from the predicted probability for the true class $\mathrm{y}_{q_{train}} \in \{s_{1_{train}}, ..., s_{P_{train}}\}$ of $\mathbf{x}_{q_{train}}$. During testing, having a trained model parameter ϕ', the class s_{pred} which meets $p_{\phi'}(y_{q_{test}} = s_{pred} | \mathbf{x}_{q_{test}}) = \max\{p_{\phi'}(y_{q_{test}} = s_n | \mathbf{x}_{q_{test}}) | n \in \{s_{1_{test}}, ..., s_{P_{test}}\}\}$ will become the predicted class for $\mathbf{x}_{q_{test}}$.

3.3. Adaptive Dependency Learning (ADL)

The classification up till now determines the total similarity by an unbiased sum of embedding distances from all comparing units. It equalizes the decision impact of local comparison from action-critical/noisy sub-areas, which hampers the generalization process of correct classification. To emphasize the learning reliance on the former and avoid the negative learning from the latter, we design a self-attention-based module ADL to distribute the contextual significance for each comparing unit. As shown in Fig. 4, we prepare three parametric matrices, known as the value head $\mathbf{V} : \mathbb{R}^{d_{feat}} \times \mathbb{R}^{d_{feat}}$, the key head $\mathbf{K} : \mathbb{R}^{d_{emb}} \times \mathbb{R}^{d_{feat}}$ and the query head $\mathbf{Q} : \mathbb{R}^{d_{emb}} \times \mathbb{R}^{d_{feat}}$, where d_{emb} is the output embedding size. After calculating $\mathbf{K_{G_f}} = \mathbf{K} \cdot \mathbf{G_f}$, $\mathbf{Q_{G_f}} = \mathbf{Q} \cdot \mathbf{G_f}$ and $\mathbf{V_{G_f}} = \mathbf{V} \cdot \mathbf{G_f}$ for a global repre-

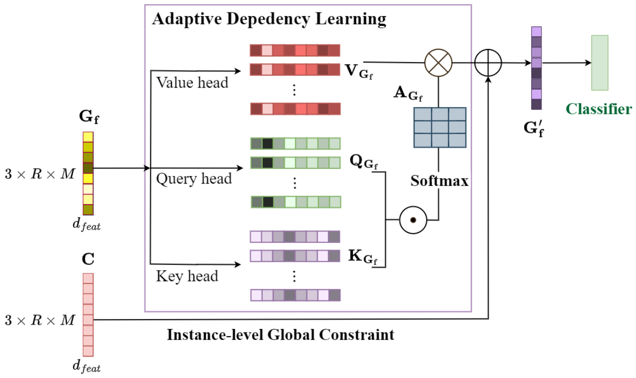

Figure 4. ADL module learns to adaptively distribute contextual comparing focus for each unit on action-related local embedding content.

sentation $\mathbf{G_f}$, we generate a matrix of normalized attention scores $\mathbf{A_{G_f}}$ which captures the action-based contextual dependency for each unit on every other unit in $\mathbf{G_f}$. Using the scores as the weights, the new content of a comparing unit $\mathbf{g'_f}$ is a weighted sum of the value-head output from its original embedding in $\mathbf{g_f}$ and every other comparing unit. To express this as matrix-level operations:

$$\mathbf{A_{G_f}} = \frac{\exp((\mathbf{K_{G_f}} \cdot \mathbf{Q_{G_f}})/\sqrt{d_{emb}})}{\sum_{j=1}^{3\times R\times M} \exp((\mathbf{K_{G_f}} \cdot \mathbf{Q_{G_f}})/\sqrt{d_{emb}})}, \tag{10}$$

$$\mathbf{G'_f} = \mathbf{A_{G_f}} \otimes \mathbf{V_{G_f}} + \mathbf{C}. \tag{11}$$

where $\otimes$ refers to the Hadamard multiplication. Pure local embedding representation clusters weakly for intra-class samples. Thus we remain the global average embedding $\mathbf{f}_{glob} = \frac{1}{T_{feat} \times U \times M} \sum_{t=1}^{T_{feat}} \sum_{v=1}^{U} \sum_{m=1}^{M} \mathbf{f}_{vt}^{m}$ from all body joints and temporal feature dimensions as a simple instance-level constraint and add it into each unit by an expansion matrix $\mathbf{C}$. We now jointly train the attention matrices and the feature encoder together in an end-to-end manner. Eventually, when figuring the new embedding for each unit, the original features from action-critical units would not only be persisted in their units but also be transmitted to other units as high-attention contextual supplements because they contain more invariant information for correct classification. On the other hand, the new embedding for the units whose original features are low-attention (i.e. noisy) would suppress their old information and be more amended by the contextual features from their correlated high-attention units or global embedding. Eventually, this promotes a targeted learning direction that emphasizes the decision weights of similarity measurements according to the native and contextual features from action-critical units, and suppresses the impacts from action-noisy units.

4. Experiments

Aligning to [19], we evaluate our model on the *NTU-RGB+D 120* dataset [13] which provides large-scale action recognition scenarios. According to its official protocol [13], the dataset is split into an 100-class auxiliary set and a 20-class evaluation set with non-overlapping classes, and each class in the evaluation set has only one reference sample. Our experiments are developed in two stages. One is the standard performance examination based on its one-shot testing protocol, checking the model's general performance trained from the full auxiliary set and its corresponding learning efficiency under different reduced auxiliary sizes. We compare our outcomes with the results in all previous related papers and analyze the difference between them and our model. Secondly, we carry out the ablation study to determine the exact learning effect brought by each model component.

4.1. Dataset and Evaluation Protocol

The *NTU-RGB+D 120* [13] dataset is a large action recognition dataset that contains 114,480 skeleton sequences of 120 action classes from 106 subjects in 155 different camera views. The action labels range from daily/health-related individual or mutual actions. Obtained by Kinect depth sensors, each sequence provides real-world 3-d coordinates of 25 body joints for each skeleton (of up to 2 attending performers). Our model needs to first get trained on the available auxiliary set to provide a general common embedding space for any newly coming action class. During the testing stage, our model predicts the evaluating samples by finding their nearest class reference neighbors according to the local-component-based comparison between their embedding representation. For the general performance examination, we use the whole 100-class auxiliary set to train our model. For the auxiliary reduction experiment, aligning to the benchmarks in [13], we apply a variable control on the auxiliary class size in a range of 20, 40, 60, 80, and 100. For the ablation study, we maintain the same experiment settings under different auxiliary sizes but apply them to the different versions of our model with respective variable control on each specific component.

4.2. Implementation Details

The model is implemented in PyTorch [23]. To unify the sequence temporal length, we apply an average-frame-sampling/zero-padding for the skeleton sequences longer/shorter than 75 frames (the mode value for the distribution from all original lengths). For the frontal-viewing pre-process, we borrow the algorithm from [17] and regard the first actor's facing direction in the first frame as the standard frontal direction to the camera throughout a sequence. Concretely, the facing is calculated as the orthogonal direction for the direction from the skeleton's left hip to its right

Approach	Accuracy
Attention Network [16]	41.0
Fully Connected [16]	42.1
Average Pooling [15]	42.9
APSR [13]	45.3
TCN [25]	46.5
SL-DML [20]	50.9
Skeleton-DML [19]	54.2
ALCA-GCN (Episodic)	**57.6**
ALCA-GCN (Traditional)	**55.0**

Table 1. General 1-shot action recognition results (%) on *NTU-RGB+D 120* with full training on all 100 auxiliary classes.

hip and the direction from its central hip to its spine. Then the 3-d location of every body joint in all the frames is vertically rotated to transform to the coordinate system under the new viewing angle. Apart from the convolution sampling strategy, our feature encoder is aligned to [38], composed of 10 iterative blocks of spatial and temporal convolution layers. For each spatial convolution layer, L is set to be 1. The output dimension for each block evolves as $64 \times 4 \rightarrow 128 \times 3 \rightarrow 256 \times 3$. The embedding dimension in ADL is 256. We conduct each experiment with a maximum training of 100 epochs on 2 NVIDIA P100 GPUs, and apply early stops when the validating accuracy doesn't improve in the latest 10 epochs. An Adam optimizer and cosine annealing are used to schedule the learning rate with a starting value of 10^{-3} and the weight decay of 10^{-6}. During training, we mainly adopt the episodic learning algorithm (see Section 3.2), in which each training-use sub-task has the same 20-way-1-shot setting from the testing protocol. As a controlled experiment, we also attempted training our model under a traditional style, in which the model is trained by normal batch learning with a batch size of 64. For the encoded feature $\mathbf{f}$ of an input example, we performed a global average pooling on $\mathbf{G}'_{\mathbf{f}}$ to get a 256-dimension feature vector and then connected it with a SoftMax classifier to train the model by the standard cross-entropy loss. During testing, we disconnected the classifier and used the trained encoder, embedder and ADL to perform the same nearest-neighbor-based classification as episodic learning.

4.3. Results

General and Training Set Size Reduction. Table 1 presents our model's general performance for the given 1-shot task, compared to the available solution results in [13, 15, 16, 19, 20, 25] under the same testing protocol. Table 2 and Fig. 5 present our model's corresponding learning efficiency under different auxiliary sizes, compared to the available results in [13, 19, 20]. The solutions in [13, 15, 16, 25] all use certain global average embedding for similarity comparison, while [19, 20] transform skeleton sequences into signal images. The outcomes show that our model learned by either training strategy always performs

# Training Classes	20	40	60	80	100
APSR [13]	29.1	34.8	39.2	42.8	45.3
SL-DML [20]	36.7	42.4	49.0	46.4	50.9
Skeleton-DML [19]	28.6	37.5	48.6	48.0	54.2
ALCA-GCN (Episodic)	38.7	46.6	**51.0**	**53.7**	**57.6**
ALCA-GCN (Traditional)	**45.0**	**49.8**	50.4	50.7	55.0

Table 2. 1-shot action recognition results (%) on *NTU-RGB+D 120* with different auxiliary training set sizes.

better than the existing solutions under any auxiliary condition. Concretely, our model trained by traditional learning outperforms the previous state-of-the-art in [20] with a margin of 8.3% and 7.4% for the auxiliary size of 20 and 40, and our model trained by episodic learning outperforms [19, 20] by a margin of 2.0%, 5.7%, 3.4% for the auxiliary size of 60, 80 and 100. We find that traditional training provides more efficient embedding learning for our model under low auxiliary supports probably because at this moment the training already simulates a similar learning sample distribution to the evaluation task (a 20-way classification), and its larger training batch helps our model more easily get out of local minimum and find the optimal parameters. On the other hand, episodic learning presents a more stable learning increase for a generalized embedding ability by meta-learning from the gradually abundant auxiliary classes. Observing the visualized learning progress under different auxiliary sizes in Fig. 5, we see that both the global-embedding-based method [13] and our model under traditional training reduce their accuracy improvement speed when the auxiliary size raises from 40 to 60 or 60 to 80. More seriously, [19, 20] face temporary learning confusion when the auxiliary size raises from 60 to 80, having a 2.6% and 0.6% accuracy drop. Contrastly, our model under episodic learning demonstrates a steady learning increase, enlarging the advantage gap when the auxiliary size is 80 or 100. We show more performance differences by confusion matrices in **Suppl. Material**.

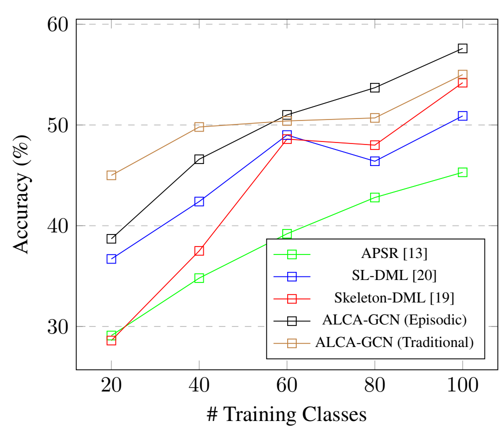

Figure 5. Visualized 1-shot accuracy variation on *NTU-RGB+D 120* with different auxiliary training set sizes.

Ablation Study. Table 3 records the ablation study results on the detailed learning effect brought by each model component (using episodic learning under the same configuration as the full training). We separate the research objects into three types of components: **convolution sampling strategy**, **comparing unit division**, and **instance-level constraints**. For convolution sampling, we examine the influence of spatial feature extraction under different scopes by only using the original convolution scheme in [38] or our body-part-based scheme. For comparing unit division, we consider the learning efficiency under different similarity metrics, including conducting the measurements by global average embedding (labeled as None because there is no local division), pure spatial-wise comparing units (dividing temporally-averaged features according to body-part partitions), or pure temporal-wise comparing units (dividing body-joint-averaged features according to temporal sections). For instance-level constraints, the study examines the performance drop when the ADL module or the global embedding constraint is removed from the original ALCA-GCN.

The outcome indicates that the overall best result under any auxiliary condition is achieved by the full ALCA-GCN. For the spatial convolution scheme, the visual features collected from either skeleton-based or body-part-based neighbor sampling provide comparable distinction validity for classification. The full ALCA-GCN concatenates them to provide a more comprehensive feature description and improves by 1.0%, 2.3%, 4.3%, and 4.2% for the auxiliary size of 40, 60, 80, and 100. A similar situation also appears for dividing comparing units only on spatial or temporal dimensions. We observe that using global average embedding predicts better than using single-dimensional comparing units by 5.4%, 0.8%, and 1.4% when the auxiliary size is 40, 80, and 100. But using double-dimensional comparing units in the full ALCA-GCN outperforms using global embedding under every condition with a respective advantage of 3.8%, 0.8%, 4.8%, 2.6%, 3.3%. Finally, both the ADL module and global embedding constraints are verified as positive regulations for our similarity metric. Especially, the performance boost brought by ADL is the most obvious. Except for the situation under 20 auxiliary classes, it steadily provides an increase of 4.4%, 4.4%, 4.8%, and 7.0% when the auxiliary size grows to 40, 60, 80, and 100 classes. Under smaller auxiliary sizes, the models with only body-part-based spatial convolution or without ADL could achieve similar learning results to the full model, because the embedding discrimination for only 20 training classes is relatively easy. With more abundant and complicated auxiliary classes, the model needs to develop its generalized embedding ability with more explicit and refined pattern recognition on potential class-specific movements, in which our ADL contributes significantly by filtering action-related local features. Despite the benefits, a limitation of our model is that the action is matched in the fixed skeleton order in a video (*i.e.* it assumes that an action is performed by the performers in the constant order). While in real cases, an action could be performed by different people. Global representations average all features from involved performers and thus do not have this issue. This requires further research on the solution of adaptively detecting action referencing order among multiple actors.

	# Training Classes	20	40	60	80	100
Sampling strategy	Body-part-based	**38.7**	45.6	45.7	49.2	53.4
	Skeleton-based	37.5	43.3	48.7	49.4	51.6
Division strategy	None	34.9	45.8	46.2	51.1	54.3
	Spatial-wise	31.4	40.4	47.4	50.3	52.9
	Temporal-wise	35.5	40.2	44.2	46.5	50.0
Constraints	Without ADL	38.6	42.2	46.6	48.9	50.6
	Without global constraints	31.3	45.8	47.1	51.6	55.1
ALCA-GCN		**38.7**	**46.6**	**51.0**	**53.7**	**57.6**

Table 3. Ablation study (%) on *NTU-RGB+D 120* for each component in our proposed model.

5. Conclusion

In this paper, we suggest a novel metric-based solution for skeleton-based one-shot action recognition. Our method decomposes the similarity comparison to an adaptive sum of embedding measurements on the local comparing units that contain hierarchical body-part-wise and temporal-wise features. To emphasize/suppress the distinction learning on action-related/noisy units, our ADL module adaptively adjusts each unit's measurement impact according to its instance-level attention. We examined our model's general performance and ablation study under an extensive experiment setup. The results proved that our model outperforms global-embedding-based and signal-based methods by providing a more action-representative similarity comparison. Using episodic learning, the model could steadily develop its embedding ability by meta-learning from the increased auxiliary resources while previous methods face a generalization bottleneck. Our solution reveals that the unique physical properties in skeleton sequences can provide invariant structural meanings as acquiescent prior knowledge to facilitate few-shot learning. Improving our method with adaptive action referencing and extending it to more generalized scenarios such as crowd activity analysis would become interesting directions to explore in the future.

6. Acknowledgement

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. MG was supported by ARC DE210101624.

References

[1] Rami Ben-Ari, Mor Shpigel Nacson, Ophir Azulai, Udi Barzelay, and Daniel Rotman. TAEN: temporal aware embedding network for few-shot action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 2786–2794. Computer Vision Foundation / IEEE, 2021.

[2] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10615–10624. Computer Vision Foundation / IEEE, 2020.

[3] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *IEEE Int. Conf. Image Process.*, pages 168–172. IEEE, 2015.

[4] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Trans. Image Process.*, 28(9):4594–4605, 2019.

[5] Zewei Ding, Pichao Wang, Philip O. Ogunbona, and Wanqing Li. Investigation of different skeleton features for cnn-based 3d action recognition. In *Int. Conf. Multimedia and Expo Worksh.*, pages 617–622. IEEE Computer Society, 2017.

[6] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. In *Int. Conf. Comput. Vis. Worksh.*, pages 1308–1316. IEEE, 2019.

[7] Weiming Hu, Dan Xie, Zhouyu Fu, Wenrong Zeng, and Stephen J. Maybank. Semantic-based surveillance video retrieval. *IEEE Trans. Image Process.*, 16(4):1168–1181, 2007.

[8] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, 2013.

[9] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12573–12581. Computer Vision Foundation / IEEE, 2020.

[10] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3595–3603. Computer Vision Foundation / IEEE, 2019.

[11] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7260–7268. Computer Vision Foundation / IEEE, 2019.

[12] Wenbo Li, Longyin Wen, Ming-Ching Chang, Ser Nam Lim, and Siwei Lyu. Adaptive RNN tree for large-scale human action recognition. In *Int. Conf. Comput. Vis.*, pages 1453–1461. IEEE Computer Society, 2017.

[13] 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 Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, 2020.

[14] Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):3007–3021, 2018.

[15] Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):3007–3021, 2018.

[16] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention LSTM networks for 3d action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3671–3680. IEEE Computer Society, 2017.

[17] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[18] 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. `https://zhoushengisnoob.github.io/papers/DASTM.pdf`, 2022.

[19] Raphael Memmesheimer, Simon Häring, Nick Theisen, and Dietrich Paulus. Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition. In *IEEE Wint. Conf. Appl. Comput. Vis.*, pages 837–845. IEEE, 2022.

[20] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. SL-DML: signal level deep metric learning for multimodal one-shot action recognition. In *Int. Conf. Pattern Recog.*, pages 4573–4580. IEEE, 2020.

[21] Boris N. Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. TADAM: task dependent adaptive metric for improved few-shot learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Adv. Neural Inform. Process. Syst.*, pages 719–729, 2018.

[22] Behnoosh Parsa, Athma Narayanan, and Behzad Dariush. Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment. In *IEEE Wint. Conf. Appl. Comput. Vis.*, pages 1069–1079. IEEE, 2020.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Adv. Neural Inform. Process. Syst.*, pages 8024–8035, 2019.

[24] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A survey on 3d skeleton-based action recognition using learning method. *CoRR*, abs/2002.05907, 2020.

[25] Alberto Sabater, Laura Santos, José Santos-Victor, Alexandre Bernardino, Luis Montesano, and Ana C. Murillo. One-shot action recognition towards novel assistive therapies. *CoRR*, abs/2102.08997, 2021.

[26] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1010–1019. IEEE Computer Society, 2016.

[27] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurIPS*, pages 4077–4087, 2017.

[28] Zehua Sun, Jun Liu, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, and Gang Wang. Human action recognition from various data modalities: A review. *CoRR*, abs/2012.11866, 2020.

[29] Shaoqing Tan and Ruoyu Yang. Learning similarity: Feature-aligning network for few-shot action recognition. In *Int. Joint Conf. Neural Net.*, pages 1–7. IEEE, 2019.

[30] Kalpit C. Thakkar and P. J. Narayanan. Part-based graph convolutional network for action recognition. In *Brit. Mach. Vis. Conf.*, page 270. BMVA Press, 2018.

[31] Theodoros Theodoridis and Huosheng Hu. Action classification of 3d human models using dynamic anns for mobile robot surveillance. In *IEEE Int. Conf. Robot. Biomim.*, pages 371–376. IEEE, 2007.

[32] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6450–6459. Computer Vision Foundation / IEEE Computer Society, 2018.

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6450–6459. Computer Vision Foundation / IEEE Computer Society, 2018.

[34] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Adv. Neural Inform. Process. Syst.*, pages 3630–3638, 2016.

[35] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3633–3642. IEEE Computer Society, 2017.

[36] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.*, 53(3):63:1–63:34, 2020.

[37] Yongqin Xian, Bruno Korbar, Matthijs Douze, Bernt Schiele, Zeynep Akata, and Lorenzo Torresani. Generalized many-way few-shot video classification. In Adrien Bartoli and Andrea Fusiello, editors, *Eur. Conf. Comput. Vis. Worksh.*, volume 12540 of *Lecture Notes in Computer Science*, pages 111–127. Springer, 2020.

[38] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, pages 7444–7452. AAAI Press, 2018.

[39] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12200–12210. Computer Vision Foundation / IEEE, 2020.

[40] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip H. S. Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Eur. Conf. Comput. Vis.*, volume 12350 of *Lecture Notes in Computer Science*, pages 525–542. Springer, 2020.

[41] Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. In Zhi-Hua Zhou, editor, *IJCAI*, pages 1309–1315. ijcai.org, 2021.