Temporal-Viewpoint Transportation Plan for Skeletal Few-shot Action Recognition

Lei Wang[†,§] and Piotr Koniusz[§,†]

[†]Australian National University [§]Data61/CSIRO
[§]firstname.lastname@data61.csiro.au

Abstract. We propose a Few-shot Learning pipeline for 3D skeleton-based action recognition by Joint tEmporal and cAmera viewpoiNt alIgnmEnt (JEANIE). To factor out misalignment between query and support sequences of 3D body joints, we propose an advanced variant of Dynamic Time Warping which jointly models each smooth path between the query and support frames to achieve simultaneously the best alignment in the temporal and simulated camera viewpoint spaces for end-to-end learning under the limited few-shot training data. Sequences are encoded with a temporal block encoder based on Simple Spectral Graph Convolution, a lightweight linear Graph Neural Network backbone. We also include a setting with a transformer. Finally, we propose a similarity-based loss which encourages the alignment of sequences of the same class while preventing the alignment of unrelated sequences. We show state-of-the-art results on NTU-60, NTU-120, Kinetics-skeleton and UWA3D Multiview Activity II.

1 Introduction

Action recognition is arguably among key topics in computer vision due to applications in video surveillance [63,65], human-computer interaction, sports analysis, virtual reality and robotics. Many pipelines [59,19,18,7,64,30] perform action classification given the large amount of labeled training data. However, manually collecting and labeling videos for 3D skeleton sequences is laborious, and such pipelines need to be retrained or fine-tuned for new class concepts. Popular action recognition networks include two-stream neural networks [19,18,71] and 3D convolutional networks (3D CNNs) [59,7], which aggregate frame-wise and temporal block representations, respectively. However, such networks indeed must be trained on large-scale datasets such as Kinetics [7,68,66,31] under a fixed set of training class concepts.

Thus, there exists a growing interest in devising effective Few-shot Learning (FSL) for action recognition, termed Few-shot Action Recognition (FSAR), that rapidly adapts to novel classes given a few training samples [47,73,23,14,79,5,67]. However, FSAR for videos is scarce due to the volumetric nature of videos and large intra-class variations.

FSL for image recognition has been widely studied [46,34,20,3,17,33] including contemporary CNN-based FSL methods [29,61,54,21,57,76], which use meta-learning, prototype-based learning and feature representation learning. Just in 2020–2022, many FSL methods [24,13,70,37,42,16,22,35,15,5,58,32,52,78,86,41] have been dedicated to image classification or detection [75,77,82,84,83]. Noteworthy mentioning is the incremental learning paradigm that can also tackle novel classes [51]. In this paper, we aim at advancing few-shot recognition of articulated set of connected 3D body joints.

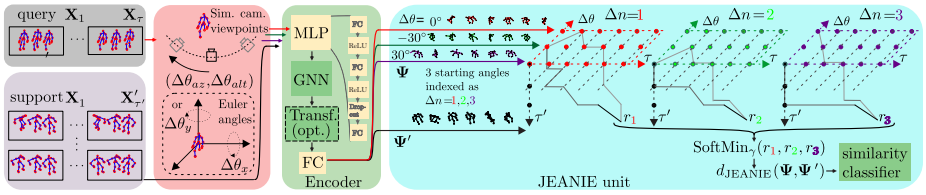


Fig. 1: Our 3D skeleton-based FSAR with JEANIE. Frames from a query sequence and a support sequence are split into short-term temporal blocks $\mathbf{X}_1, ..., \mathbf{X}_\tau$ and $\mathbf{X}'_1, ..., \mathbf{X}'_{\tau'}$ of length M given stride S. Subsequently, we generate (i) multiple rotations by $(\Delta\theta_x, \Delta\theta_y)$ of each query skeleton by either Euler angles (baseline approach) or (ii) simulated camera views (gray cameras) by camera shifts $(\Delta\theta_{az}, \Delta\theta_{alt})$ w.r.t. the assumed average camera location (black camera). We pass all skeletons via Encoding Network (with an optional transformer) to obtain feature tensors $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$, which are directed to JEANIE. We note that the temporal-viewpoint alignment takes place in 4D space (we show a 3D case with three views: $-30°, 0°, 30°$). Temporally-wise, JEANIE starts from the same $t=(1,1)$ and finishes at $t=(\tau, \tau')$ (as in DTW). Viewpoint-wise, JEANIE starts from every possible camera shift $\Delta\theta \in \{-30°, 0°, 30°\}$ (we do not know the true correct pose) and finishes at one of possible camera shifts. At each step, the path may move by no more than $(\pm\Delta\theta_{az}, \pm\Delta\theta_{alt})$ to prevent erroneous alignments. Finally, SoftMin picks up the smallest distance.

With an exception of very recent models [38,39,45,44,67,48], FSAR approaches that learn from skeleton-based 3D body joints are scarce. The above situation prevails despite action recognition from articulated sets of connected body joints, expressed as 3D coordinates, does offer a number of advantages over videos such as (i) the lack of the background clutter, (ii) the volume of data being several orders of magnitude smaller, and (iii) the 3D geometric manipulations of sequences being relatively friendly.

Thus, we propose a FSAR approach that learns on skeleton-based 3D body joints via Joint tEmporal and cAmera viewpoiNt alIgnmEnt (JEANIE). As FSL is based on learning similarity between support-query pairs, to achieve good matching of queries with support sequences representing the same action class, we propose to simultaneously model the optimal (i) temporal and (ii) viewpoint alignments. To this end, we build on soft-DTW [11], a differentiable variant of Dynamic Time Warping (DTW) [10]. Unlike soft-DTW, we exploit the projective camera geometry. We assume that the best smooth path in DTW should simultaneously provide the best temporal and viewpoint alignment, as sequences that are being matched might have been captured under different camera viewpoints or subjects might have followed different trajectories.

To obtain skeletons under several viewpoints, we rotate skeletons (zero-centered by hip) by Euler angles [1] w.r.t. x, y and z axes, or generate skeleton locations given simulated camera positions, according to the algebra of stereo projections [2].

We note that view-adaptive models for action recognition do exist. View Adaptive Recurrent Neural Networks [80,81] is a classification model equipped with a view-adaptive subnetwork that contains the rotation and translation switches within its RNN backbone, and the main LSTM-based network. Temporal Segment Network [62] mod-

els long-range temporal structures with a new segment-based sampling and aggregation module. However, such pipelines require a large number of training samples with varying viewpoints and temporal shifts to learn a robust model. Their limitations become evident when a network trained under a fixed set of action classes has to be adapted to samples of novel classes. Our JEANIE does not suffer from such a limitation.

Our pipeline consists of an MLP which takes neighboring frames to form a temporal block. Firstly, we sample desired Euler rotations or simulated camera viewpoints, generate multiple skeleton views, and pass them to the MLP to get block-wise feature maps, next forwarded to a Graph Neural Network (GNN), *e.g.*, GCN [27], Fisher-Bures GCN [56], SGC [72], APPNP [28] or S^2GC [85,87], followed by an optional transformer [12], and an FC layer to obtain graph-based representations passed to JEANIE.

JEANIE builds on Reproducing Kernel Hilbert Spaces (RKHS) [53] which scale gracefully to FSAR problems which, by their setting, learn to match pairs of sequences rather than predict class labels. JEANIE builds on Optimal Transport [60] by using a transportation plan for temporal and viewpoint alignment in skeletal action recognition.

Below are our contributions:

i. We propose a Few-shot Action Recognition approach for learning on skeleton-based articulated 3D body joints via 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. Warping jointly temporal locations and simulated viewpoint indexes helps meta-learning with limited samples of novel classes.
ii. To simulate different viewpoints of 3D skeleton sequences, we consider rotating them (1) by Euler angles within a specified range along x and y axes, or (2) towards the simulated camera locations based on the algebra of stereo projection.
iii. We investigate several different GNN backbones (including transformer), as well as the optimal temporal size and stride for temporal blocks encoded by a simple 3-layer MLP unit before forwarding them to GNN.
iv. We propose a simple similarity-based loss encouraging the alignment of within-class sequences and preventing the alignment of between-class sequences.

We achieve the state of the art on large-scale NTU-60 [50], NTU-120 [39], Kinetics-skeleton [74] and UWA3D Multiview Activity II [49]. As far as we can tell, the simultaneous alignment in the joint temporal-viewpoint space for FSAR is a novel proposition.

2 Related Works

Below, we describe 3D skeleton-based action recognition, FSAR approaches and GNNs.

Action recognition (3D skeletons). 3D skeleton-based action recognition pipelines often use GCNs [27], *e.g.*, spatio-temporal GCN [74], an a-links inference model [36], shift-graph model [9] and multi-scale aggregation node [40]. However, such models rely on large-scale datasets, and cannot be easily adapted to novel class concepts.

FSAR (videos). Approaches [47,23,73] use a generative model, graph matching on 3D coordinates and dilated networks, respectively. Approach [88] uses a compound memory network. ProtoGAN [14] generates action prototypes. Model [79] uses permutation-invariant attention and second-order aggregation of temporal video blocks, whereas approach [5] proposes a modified temporal alignment for query-support pairs via DTW.

FSAR (3D skeletons). Few FSAR models use 3D skeletons [38,39,45,44]. Global Context-Aware Attention LSTM [38] selectively focuses on informative joints. Action-Part Semantic Relevance-aware (APSR) model [39] uses the semantic relevance between each body part and action class at the distributed word embedding level. Signal Level Deep Metric Learning (DML) [45] and Skeleton-DML [44] one-shot FSL approaches encode signals into images, extract features using CNN and apply multi-similarity miner losses. In contrast, we use temporal blocks of 3D body joints of skeletons encoded by GNNs under multiple viewpoints of skeletons to simultaneously perform temporal and viewpoint-wise alignment of query-support in the meta-learning regime.

Graph Neural Networks. GNNs are popular in the skeleton-based action recognition. We build on GNNs in this paper due to their excellent ability to represent graph-structured data such as interconnected body joints. GCN [27] applies graph convolution in the spectral domain, and enjoys the depth-efficiency when stacking multiple layers due to non-linearities. However, depth-efficiency costs speed due to backpropagation through consecutive layers. In contrast, a very recent family of so-called spectral filters do not require depth-efficiency but apply filters based on heat diffusion to the graph Laplacian. As a result, they are fast linear models as learnable weights act on filtered node representations. SGC [72], APPNP [28] and S^2GC [85] are three methods from this family which we investigate for the backbone.

Multi-view action recognition. Multi-modal sensors enable multi-view action recognition [64,80]. A Generative Multi-View Action Recognition framework [69] integrates complementary information from RGB and depth sensors by View Correlation Discovery Network. Some works exploit multiple views of the subject [50,39,81,69] to overcome the viewpoint variations for action recognition on large training datasets. In contrast, our JEANIE learns to perform jointly the temporal and simulated viewpoint alignment in an end-to-end meta-learning setting. This is a novel paradigm based on similarity learning of support-query pairs rather than learning class concepts.

3 Approach

To learn similarity/dissimilarity between pairs of sequences of 3D body joints representing query and support samples from episodes, our goal is to find a smooth joint viewpoint-temporal alignment of query and support and minimize or maximize the matching distance d_{JEANIE} (end-to-end setting) for same or different support-query labels, respectively. Fig. 2 (top) shows that sometimes matching of query and support may be as easy as rotating one trajectory onto another, in order to achieve viewpoint invariance. A viewpoint invariant distance [25] can be defined as:

$$d_{\text{inv}}(\boldsymbol{\Psi}, \boldsymbol{\Psi}') = \inf_{\gamma,\gamma' \in T} d\big(\gamma(\boldsymbol{\Psi}), \gamma'(\boldsymbol{\Psi}')\big), \tag{1}$$

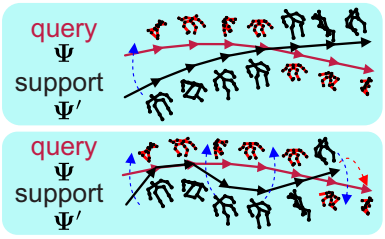


Fig. 2: (*top*) In viewpoint-invariant learning, the distance between query features $\boldsymbol{\Psi}$ and support features $\boldsymbol{\Psi}'$ has to be computed. The blue arrow indicates that trajectories of both actions need alignment. (*bottom*) In real life, subject's 3D body joints deviate from one ideal trajectory, and so advanced viewpoint alignment strategy is needed.

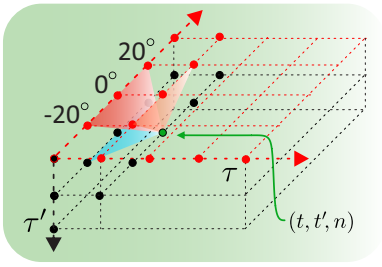


Fig. 3: JEANIE (1-max shift). We loop over all points. At (t, t', n) (green point) we add its base distance to the minimum of accumulated distances at $(t, t'-1, n-1)$, $(t, t'-1, n)$, $(t, t'-1, n+1)$ (orange plane), $(t-1, t'-1, n-1)$, $(t-1, t'-1, n)$, $(t-1, t'-1, n+1)$ (red plane) and $(t-1, t', n-1)$, $(t-1, t', n)$, $(t-1, t', n+1)$ (blue plane).

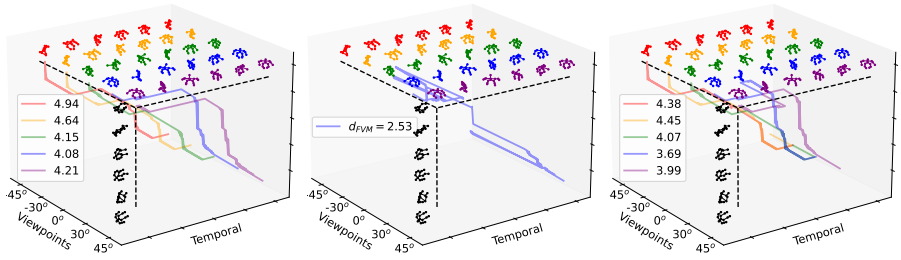
where T is a set of transformations required to achieve a viewpoint invariance, $d(\cdot, \cdot)$ is some base distance, *e.g.*, the Euclidean distance, and $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}'$ are features describing query and support pair of sequences. Typically, T may include 3D rotations to rotate one trajectory onto the other. However, such a global viewpoint alignment of two sequences is suboptimal. Trajectories are unlikely to be straight 2D lines in the 3D space. Fig. 2 (bottom) shows that 3D body joints locally follow complicated non-linear paths.

Thus, we propose JEANIE that aligns and warps query/support sequences based on the feature similarity. One can think of JEANIE as performing Eq. (1) with T containing camera viewpoint rotations, and the base distance $d(\cdot, \cdot)$ being a joint temporal-viewpoint variant of soft-DTW to account for local temporal-viewpoint variations of 3D body joint trajectories. JEANIE unit in Fig. 1 realizes such a strategy (SoftMin operation is equivalent of Eq. (1)). While such an idea sounds simple, it is effective, it has not been done before. Fig. 3 (discussed later in the text) shows one step of the temporal-viewpoint computations of JEANIE.

We present a necessary background on Euler angles and the algebra of stereo projection, GNNs and the formulation of soft-DTW in Appendix Sec. A. Below, we detail our pipeline shown in Figure 1, explain the proposed JEANIE and our loss function.

Notations. $\mathcal{I}_K$ stands for the index set $\{1, 2, ..., K\}$. Concatenation of α_i is denoted by $[\alpha_i]_{i \in \mathcal{I}_I}$, whereas $\mathbf{X}_{:,i}$ means we extract/access column i of matrix $\boldsymbol{D}$. Calligraphic mathcal fonts denote tensors (*e.g.*, $\boldsymbol{\mathcal{D}}$), capitalized bold symbols are matrices (*e.g.*, $\boldsymbol{D}$), lowercase bold symbols are vectors (*e.g.*, $\boldsymbol{\psi}$), and regular fonts denote scalars.

Encoding Network (EN). We start by generating $K \times K'$ Euler rotations or $K \times K'$ simulated camera views (moved gradually from the estimated camera location) of query skeletons. Our EN contains a simple 3-layer MLP unit (FC, ReLU, FC, ReLU, Dropout, FC), GNN, optional Transformer [12] and FC. The MLP unit takes M neighboring frames, each with J 3D skeleton body joints, forming one temporal block. In total,



(a) soft-DTW (view-wise) (b) FVM (c) JEANIE (1-max shift)

Fig. 4: A comparison of paths in 3D for soft-DTW, Free Viewpoint Matching (FVM) and our JEANIE. For a given support skeleton sequence (green color), we choose viewing angles between -45° and 45° for the camera viewpoint simulation. The support skeleton sequence is shown in black color. (a) soft-DTW finds each individual alignment per viewpoint fixed throughout alignment: $d_{\text{shortest}} = 4.08$. (b) FVM is a greedy matching algorithm that in each time step seeks the best alignment pose from all viewpoints which leads to unrealistic zigzag path (person cannot jump from front to back view suddenly): $d_{\text{FVM}} = 2.53$. (c) Our JEANIE (1-max shift) is able to find smooth joint viewpoint-temporal alignment between support and query sequences. We show each optimal path for each possible starting position: $d_{\text{JEANIE}} = 3.69$. While $d_{\text{FVM}} = 2.53$ for FVM is overoptimistic, $d_{\text{shortest}} = 4.08$ for fixed-view matching is too pessimistic, whereas JEANIE strikes the right matching balance with $d_{\text{JEANIE}} = 3.69$.

depending on stride S, we obtain some τ temporal blocks which capture the short temporal dependency, whereas the long temporal dependency is modeled with our JEANIE. Each temporal block is encoded by the MLP into a $d \times J$ dimensional feature map. Subsequently, query feature maps of size $K \times K' \times \tau$ and support feature maps of size τ' are forwarded to a GNN, optional Transformer (similar to ViT [12], instead of using image patches, we feed each body joint encoded by GNN into the transformer), and an FC layer, which returns $\boldsymbol{\Psi} \in \mathbb{R}^{d' \times K \times K' \times \tau}$ query feature maps and $\boldsymbol{\Psi'} \in \mathbb{R}^{d' \times \tau'}$ support feature maps. Feature maps are passed to JEANIE and the similarity classifier.

Let support maps $\boldsymbol{\Psi'}$ be $[f(\boldsymbol{X}'_1; \mathcal{F}), ..., f(\boldsymbol{X}'_{\tau'}; \mathcal{F})] \in \mathbb{R}^{d' \times \tau'}$ and query maps $\boldsymbol{\Psi}$ be $[f(\boldsymbol{X}_1; \mathcal{F}), ..., f(\boldsymbol{X}_{\tau}; \mathcal{F})] \in \mathbb{R}^{d' \times K \times K' \times \tau}$, for query and support frames per block $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{3 \times J \times M}$. Moreover, we define $f(\mathbf{X}; \mathcal{F}) = \text{FC}(\text{Transf}(\text{GNN}(\text{MLP}(\mathbf{X}; \mathcal{F}_{MLP}); \mathcal{F}_{GNN}); \mathcal{F}_{Transf}); \mathcal{F}_{FC})$, $\mathcal{F} \equiv [\mathcal{F}_{MLP}, \mathcal{F}_{GNN}, \mathcal{F}_{Transf}, \mathcal{F}_{FC}]$ is the set of parameters of EN (note optional Transformer [12]). As GNN, we try GCN [27], SGC [72], APPNP [28] or S^{2}GC [85].

JEANIE. Matching query-support pairs requires temporal alignment due to potential offset in locations of discriminative parts of actions, and due to potentially different dynamics/speed of actions taking place. The same concerns the direction of the dominant action trajectory w.r.t. the camera. Thus, JEANIE, our advanced soft-DTW, has the transportation plan $\mathcal{A}' \equiv \mathcal{A}_{\tau,\tau',K,K'}$, where apart from temporal block counts τ and τ', for query sequences, we have possible η_{az} left and η_{az} right steps from the initial camera azimuth, and η_{alt} up and η_{alt} down steps from the initial camera altitude.

Thus, $K=2\eta_{az}+1$, $K'=2\eta_{alt}+1$. For the variant with Euler angles, we simply have $\mathcal{A}''\equiv\mathcal{A}_{\tau,\tau',K,K'}$ where $K=2\eta_x+1$, $K'=2\eta_y+1$ instead. Then, JEANIE is given as:

$$d_{\text{JEANIE}}(\boldsymbol{\Psi},\boldsymbol{\Psi}')=\underset{\mathbf{A}\in\mathcal{A}'}{\text{SoftMin}_\gamma}\left\langle\mathbf{A},\boldsymbol{\mathcal{D}}(\boldsymbol{\Psi},\boldsymbol{\Psi}')\right\rangle, \tag{2}$$

where $\boldsymbol{\mathcal{D}}\in\mathbb{R}_+^{K\times K'\times\tau\times\tau'}\equiv[d_{\text{base}}(\boldsymbol{\psi}_{m,k,k'},\boldsymbol{\psi}'_n)]_{\substack{(m,n)\in\mathcal{I}_\tau\times\mathcal{I}_{\tau'}\\(k,k')\in\mathcal{I}_K\times\mathcal{I}_{K'}}}$ and $\boldsymbol{\mathcal{D}}$ contains distances.

Figure 3 shows one step of JEANIE (1-max shift). Suppose the given viewing angle set is $\{-40°,-20°,0°,20°,40°\}$. For 1-max shift, we loop over (t,t',n). At location (t,t',n), we extract the base distance and add it together with the minimum of aggregated distances at the shown 9 predecessor points. We store that total distance at (t,t',n), and we move to the next point. Note that for viewpoint index n, we look up $(n-1,n,n+1)$. Extension to the ι-max shift is straightforward.

Algorithm 1 illustrates JEANIE. For brevity, let us tackle the camera viewpoint alignment in a single space, *e.g.*, for some shifting steps $-\eta,...,\eta$, each with size $\Delta\theta_{az}$. The maximum viewpoint change from block to block is ι-max shift (smoothness). As we have no way to know the initial optimal camera shift, we initialize all possible origins of shifts in accumulator $r_{n,1,1}=d_{\text{base}}(\boldsymbol{\psi}_{n,1},\boldsymbol{\psi}'_1)$ for all $n\in\{-\eta,...,\eta\}$. Subsequently, a phase related to soft-DTW (temporal-viewpoint alignment) takes place. Finally, we choose the path with the smallest distance over all possible viewpoint ends by selecting a soft-minimum over $[r_{n,\tau,\tau'}]_{n\in\{-\eta,...,\eta\}}$. Notice that accumulator $\boldsymbol{\mathcal{R}}\in\mathbb{R}^{(2\iota+1)\times\tau\times\tau'}$. Moreover, whenever either index $n-i$, $t-j$ or $t'-k$ in $r_{n-i,t-j,t'-k}$ (see algorithm) is out of bounds, we define $r_{n-i,t-j,t'-k}=\infty$.

FVM. To ascertain whether JEANIE is better than performing separately the temporal and simulated viewpoint alignments, we introduce a baseline called the Free Viewpoint Matching (FVM). FVM, for every step of DTW, seeks the best local viewpoint alignment, thus realizing non-smooth temporal-viewpoint path in contrast to JEANIE. To this end, we apply DTW in Eq. (2) with the base distance replaced by:

$$d_{\text{FVM}(\boldsymbol{\psi}_t,\boldsymbol{\psi}'_{t'})}=\underset{m,n,m',n'\in\{-\eta,...,\eta\}}{\text{SoftMin}_{\bar{\gamma}}}\ d_{\text{base}}(\boldsymbol{\psi}_{m,n,t},\boldsymbol{\psi}'_{m',n',t'}), \tag{3}$$

where $\boldsymbol{\Psi}\in\mathbb{R}^{d'\times K\times K'\times\tau}$ and $\boldsymbol{\Psi}'\in\mathbb{R}^{d'\times K\times K'\times\tau'}$ are query and support feature maps. We abuse the notation by writing $d_{\text{FVM}(\boldsymbol{\psi}_t,\boldsymbol{\psi}'_{t'})}$ as we minimize over viewpoint indexes in Eq. (3). We compute the distance matrix $\boldsymbol{D}\in\mathbb{R}_+^{\tau\times\tau'}\equiv[d_{\text{FVM}}(\boldsymbol{\psi}_t,\boldsymbol{\psi}'_{t'})]_{(t,t')\in\mathcal{I}_\tau\times\mathcal{I}_{\tau'}}$.

Fig. 4 shows the comparison between soft-DTW (view-wise), FVM and our JEANIE. FVM is a greedy matching method which leads to complex zigzag path in 3D space (assuming the camera viewpoint single space in $\boldsymbol{\psi}_{n,t}$ and no viewpoint in $\boldsymbol{\psi}'_{t'}$). Although FVM is able to find the smallest distance path compared to soft-DTW and JEANIE, it suffers from several issues (i) It is unreasonable for poses in a given sequence to match under sudden jumps in viewpoints. (ii) Suppose the two sequences are from two different classes, FVM still yields the smallest distance (decreased inter-class variance).

Loss Function. For the N-way Z-shot problem, we have one query feature map and $N\times Z$ support feature maps per episode. We form a mini-batch containing B

Algorithm 1 Joint tEmporal and cAmera viewpoiNt alIgnmEnt (JEANIE).

Input (forward pass): $\boldsymbol{\Psi}, \boldsymbol{\Psi}'$, $\gamma>0$, $d_{\text{base}}(\cdot,\cdot)$, ι-max shift.

$r_{:,:,:}=\infty$, $r_{n,1,1}=d_{\text{base}}(\boldsymbol{\psi}_{n,1}, \boldsymbol{\psi}'_1),\ \forall n\in\{-\eta,...,\eta\}$
$\Pi \equiv \{-\iota,...,0,...,\iota\}\times\{(0,1),(1,0),(1,1)\}$
for $t\in\mathcal{I}_\tau$:
  for $t'\in\mathcal{I}_{\tau'}$:
    if $t\neq 1$ or $t'\neq 1$:
      for $n\in\{-\eta,...,\eta\}$:
        $r_{n,t,t'} = d_{\text{base}}(\boldsymbol{\psi}_{n,t}, \boldsymbol{\psi}'_{t'}) + \text{SoftMin}_\gamma\left([r_{n-i,t-j,t'-k}]_{(i,j,k)\in\Pi}\right)$

Output: $\text{SoftMin}_\gamma\left([r_{n,\tau,\tau'}]_{n\in\{-\eta,...,\eta\}}\right)$

episodes. Thus, we have query feature maps $\{\boldsymbol{\Psi}_b\}_{b\in\mathcal{I}_B}$ and support feature maps $\{\boldsymbol{\Psi}'_{b,n,z}\}_{b\in\mathcal{I}_B,n\in\mathcal{I}_N,z\in\mathcal{I}_Z}$. Moreover, $\boldsymbol{\Psi}_b$ and $\boldsymbol{\Psi}'_{b,1,:}$ share the same class, one of N classes drawn per episode, forming the subset $C^\ddagger\equiv\{c_1,...,c_N\}\subset\mathcal{I}_C\equiv\mathcal{C}$. To be precise, labels $y(\boldsymbol{\Psi}_b)=y(\boldsymbol{\Psi}'_{b,1,z}),\forall b\in\mathcal{I}_B,z\in\mathcal{I}_Z$ while $y(\boldsymbol{\Psi}_b)\neq y(\boldsymbol{\Psi}'_{b,n,z}),\forall b\in\mathcal{I}_B,n\in\mathcal{I}_N\setminus\{1\},z\in\mathcal{I}_Z$. In most cases, $y(\boldsymbol{\Psi}_b)\neq y(\boldsymbol{\Psi}_{b'})$ if $b\neq b'$ and $b,b'\in\mathcal{I}_B$. Selection of $C^\ddagger$ per episode is random. For the N-way Z-shot protocol, we minimize:

$$l(\boldsymbol{d}^+,\boldsymbol{d}^-)=\left(\mu(\boldsymbol{d}^+)-\{\mu(\text{TopMin}_\beta(\boldsymbol{d}^+))\}\right)^2 \tag{4}$$

$$+\left(\mu(\boldsymbol{d}^-)-\{\mu(\text{TopMax}_{\text{NZ}\beta}(\boldsymbol{d}^-))\}\right)^2, \tag{5}$$

where $\boldsymbol{d}^+=[d_{\text{JEANIE}}(\boldsymbol{\Psi}_b,\boldsymbol{\Psi}'_{b,1,z})]_{\substack{b\in\mathcal{I}_B\\ z\in\mathcal{I}_Z}}$ and $\boldsymbol{d}^-=[d_{\text{JEANIE}}(\boldsymbol{\Psi}_b,\boldsymbol{\Psi}'_{b,n,z})]_{\substack{b\in\mathcal{I}_B,\\ n\in\mathcal{I}_N\setminus\{1\},z\in\mathcal{I}_Z}}$,

where $\boldsymbol{d}^+$ is a set of within-class distances for the mini-batch of size B given N-way Z-shot learning protocol. By analogy, $\boldsymbol{d}^-$ is a set of between-class distances. Function $\mu(\cdot)$ is simply the mean over coefficients of the input vector, $\{\cdot\}$ detaches the graph during the backpropagation step, whereas $\text{TopMin}_\beta(\cdot)$ and $\text{TopMax}_{\text{NZ}\beta}(\cdot)$ return β smallest and $NZ\beta$ largest coefficients from the input vectors, respectively. Thus, Eq. (4) promotes the within-class similarity while Eq. (5) reduces the between-class similarity. Integer $\beta\geq 0$ controls the focus on difficult examples, *e.g.*, $\beta=1$ encourages all within-class distances in Eq. (4) to be close to the positive target $\mu(\text{TopMin}_\beta(\cdot))$, the smallest observed within-class distance in the mini-batch. If $\beta>1$, this means we relax our positive target. By analogy, if $\beta=1$, we encourage all between-class distances in Eq. (5) to approach the negative target $\mu(\text{TopMax}_{\text{NZ}\beta}(\cdot))$, the average over the largest NZ between-class distances. If $\beta>1$, the negative target is relaxed.

4 Experiments

We provide network configurations and training details in Appendix Sec. H. Below, we describe the datasets and evaluation protocols on which we validate our JEANIE.

Datasets. Appendix Sec. B. and Table 9. contain details of datasets described below.

i. *UWA3D Multiview Activity II* [49] contains 30 actions performed by 9 people in a cluttered environment. In this dataset, the Kinect camera was moved to different

positions to capture the actions from 4 different views: front view (V_1), left view (V_2), right view (V_3), and top view (V_4).

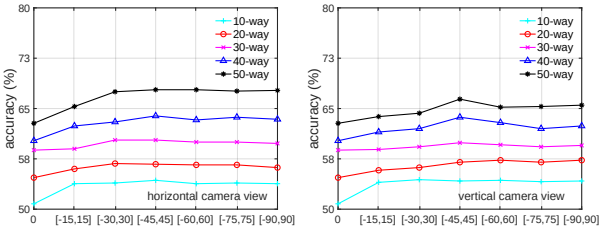
ii. *NTU RGB+D (NTU-60)* [50] contains 56,880 video sequences and over 4 million frames. This dataset has variable sequence lengths and high intra-class variations.

iii. *NTU RGB+D 120 (NTU-120)* [39], an extension of NTU-60, contains 120 action classes (daily/health-related), and 114,480 RGB+D video samples captured with 106 distinct human subjects from 155 different camera viewpoints.

iv. *Kinetics* [26] is a large-scale collection of 650,000 video clips that cover 400/600/700 human action classes. It includes human-object interactions such as *playing instruments*, as well as human-human interactions such as *shaking hands* and *hugging*. As the Kinetics-400 dataset provides only the raw videos, we follow approach [74] and use the estimated joint locations in the pixel coordinate system as the input to our pipeline. To obtain the joint locations, we first resize all videos to the resolution of 340 $\times$ 256, and convert the frame rate to 30 FPS. Then we use the publicly available *OpenPose* [6] toolbox to estimate the location of 18 joints on every frame of the clips. As OpenPose produces the 2D body joint coordinates and Kinetics-400 does not offer multiview or depth data, we use a network of Martinez *et al*. [43] pre-trained on Human3.6M [8], combined with the 2D OpenPose output to estimate 3D coordinates from 2D coordinates. The 2D OpenPose and the latter network give us (x, y) and z coordinates, respectively.

Evaluation protocols. For the UWA3D Multiview Activity II, we use standard multi-view classification protocol [49,63,64], but we apply it to one-shot learning as the view combinations for training and testing sets are disjoint. For NTU-120, we follow the standard one-shot protocol [39]. Based on this protocol, we create a similar one-shot protocol for NTU-60, with 50/10 action classes used for training/testing respectively. To evaluate the effectiveness of the proposed method on viewpoint alignment, we also create two new protocols on NTU-120, for which we group the whole dataset based on (i) horizontal camera views into left, center and right views, (ii) vertical camera views into top, center and bottom views. We conduct two sets of experiments on such disjoint view-wise splits: (i) using 100 action classes for training, and testing on the same 100 action classes (ii) training on 100 action classes but testing on the rest unseen 20 classes. Appendix Sec. G details new/additional eval. protocols on NTU-60/NTU-120.

Stereo projections. For simulating different camera viewpoints, we estimate the fundamental matrix F (Eq. 7 in Appendix), which relies on camera parameters. Thus, we use the Camera Calibrator from MATLAB to estimate intrinsic, extrinsic and lens distortion parameters. For a given skeleton dataset, we compute the range of spatial coordinates x and y, respectively. We then split them into 3 equally-sized groups to form roughly left, center, right views and other 3 groups for bottom, center, top views. We choose $\sim$15 frame images from each corresponding group, upload them to the Camera Calibrator, and export camera parameters. We then compute the average distance/depth and height per group to estimate the camera position. On NTU-60 and NTU-120, we simply group the whole dataset into 3 cameras, which are left, center and right views, as provided in [39], and then we compute the average distance per camera view based on the height and distance settings given in the table in [39].



(a) horizontal camera view (b) vertical camera view

Fig. 5: The impact of viewing angles on NTU-60.

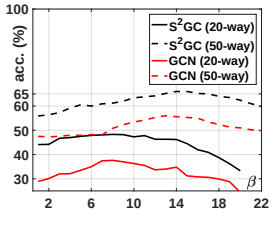


Fig. 6: The impact of β in loss function on NTU-60 with S^{2}GC and GCN.

Table 1: Experimental results on NTU-60 (left) and NTU-120 (right) for different camera viewpoint simulations. Below the dashed line are ablated few variants of JEANIE.

	NTU-60					NTU-120				
# Training Classes	10	20	30	40	50	20	40	60	80	100
Euler simple ($K+K'$)	54.3	56.2	60.4	64.0	68.1	30.7	36.8	39.5	44.3	46.9
Euler ($K\times K'$)	**60.8**	67.4	67.5	70.3	**75.0**	32.9	39.2	43.5	48.4	50.2
CamVPC ($K\times K'$)	59.7	**68.7**	**68.4**	**70.4**	73.2	**33.1**	**40.8**	**43.7**	**48.4**	**51.4**
V(Euler)	54.0	56.0	60.2	63.8	67.8	30.6	36.7	39.2	44.0	47.0
2V(Euler simple)	54.3	56.2	60.4	64.0	68.1	30.7	36.8	39.5	44.3	46.9
2V(Euler)	60.8	67.4	67.5	70.3	75.0	32.9	39.2	43.5	48.4	50.2
2V(CamVPC)	59.7	68.7	68.4	70.4	73.2	33.1	40.8	43.7	48.4	51.4
2V(CamVPC+crossval.)	63.4	72.4	73.5	73.2	78.1	37.2	43.0	49.2	50.0	55.2
2V(CamVPC+crossval.)+Transf.	**65.0**	**75.2**	**76.7**	**78.9**	**80.0**	**38.5**	**44.1**	**50.3**	**51.2**	**57.0**

4.1 Ablation Study

We start our experiments by investigating the GNN backbones (Appendix Sec. C.1), camera viewpoint simulation and their hyper-parameters (Appendix Sec. C.3, C.4, C.5).

Camera viewpoint simulations . We choose 15 degrees as the step size for the viewpoints simulation. The ranges of camera azimuth/altitude are in $[-90°, 90°]$. Where stated, we perform a grid search on camera azimuth/altitude with Hyperopt. Below, we explore the choice of the angle ranges for both horizontal and vertical views. Fig. 5a and 5b (evaluations on the NTU-60 dataset) show that the angle range $[-45°, 45°]$ performs the best, and widening the range in both views does not increase the performance any further. Table 1 (top) shows results for the chosen range $[-45°, 45°]$ of camera viewpoint simulations. Euler simple ($K+K'$) denotes a simple concatenation of features from both horizontal and vertical views, whereas Euler/CamVPC($K\times K'$) represents the grid search of all possible views. It shows that Euler angles for the viewpoint augmentation outperform Euler simple, and CamVPC (viewpoints of query sequences are generated by the stereo projection geometry) outperforms Euler angles in almost all the experiments on NTU-60 and NTU-120. This proves the effectiveness of using the stereo projection geometry for the viewpoint augmentation. More baseline experiments with/without viewpoint alignment are in Appendix Sec. C.2.

Table 2: Experimental results on NTU-60 (left) and NTU-120 (right) for ι-max shift. ι-max shift is the max. viewpoint shift from block to block in JEANIE.

	NTU-60					NTU-120				
	10	20	30	40	50	20	40	60	80	100
$\iota=1$	60.8	70.7	72.5	72.9	75.2	36.3	42.5	48.7	**50.0**	54.8
$\iota=2$	**63.8**	**72.9**	**74.0**	**73.4**	**78.1**	**37.2**	**43.0**	**49.2**	**50.0**	**55.2**
$\iota=3$	55.2	58.9	65.7	67.1	72.5	36.7	**43.0**	48.5	49.0	54.9
$\iota=4$	54.5	57.8	63.5	65.2	70.4	36.5	42.9	48.3	48.9	54.3

Table 3: The impact of the number of frames M in temporal block under stride step S on results (NTU-60). $S = pM$, where $1-p$ describes the temporal block overlap percentage. Higher p means fewer overlap frames between temporal blocks.

	$S=M$		$S=0.8M$		$S=0.6M$		$S=0.4M$		$S=0.2M$	
M	50-class	20-class	50-class	20-class	50-class	20-class	50-class	20-class	50-class	20-class
5	69.0	55.7	71.8	57.2	69.2	59.6	73.0	60.8	71.2	61.2
6	69.4	54.0	65.4	54.1	67.8	58.0	72.0	57.8	**73.0**	**63.0**
8	67.0	52.7	67.0	52.5	**73.8**	**61.8**	67.8	60.3	68.4	59.4
10	62.2	44.5	63.6	50.9	65.2	48.4	62.4	57.0	70.4	56.7
15	62.0	43.5	62.6	48.9	64.7	47.9	62.4	57.2	68.3	56.7
30	55.6	42.8	57.2	44.8	59.2	43.9	58.8	55.3	60.2	53.8
45	50.0	39.8	50.5	40.6	52.3	39.9	53.0	42.1	54.0	45.2

Evaluation of β . Figure 6 shows that if $\beta=8$ and 14, our loss function performs the best on 20- and 50-class protocol, respectively, on NTU-60 for the S^{2}GC and GCN backbone. Moreover, β is not affected by backbone.

The ι-max shift . Table 2 shows the evaluations of ι for the maximum shift. We notice that $\iota=2$ yields the best results for all the experimental settings on both NTU-60 and NTU-120. Increasing ι does not help improve the performance.

Block size and strides . Table 3 shows evaluations of block size M and stride S, and indicates that the best performance (both 50- and 20-class) is achieved for smaller block size (frame count in the block) and smaller stride. Longer temporal blocks decrease the performance due to the temporal information not reaching the temporal alignment step. Our block encoder encodes each temporal block for learning the local temporal motions, and aggregate these block features finally to form the global temporal motion cues. Smaller stride helps capture more local motion patterns. Considering the computational cost and the performance, we choose $M=8$ and $S=0.6M$.

Euler *vs*. CamVPC . Table 1 (bottom) shows that using the viewpoint alignment simultaneously in two dimensions, x and y for Euler angles, or azimuth and altitude the stereo projection geometry (*CamVPC*), improves the performance by 5-8% compared to (*Euler simple*), a variant where the best viewpoint alignment path was chosen from the best alignment path along x and the best alignment path along y. Euler simple is better than Euler with y rotations only ((*V*) includes rotations along y while (*2V*) includes rotations along two axes). Using HyperOpt [4] to search for the best angle range in

Table 4: Results on NTU-60 (S^2GC backbone). Models use temporal alignment by soft-DTW or JEANIE (joint temporal-viewpoint alignment) except if indicated otherwise.

# Training Classes	10	20	30	40	50
Each frame to frontal view	52.9	53.3	54.6	54.2	58.3
Each block to frontal view	53.9	56.1	60.1	63.8	68.0
Traj. aligned baseline (video-level)	36.1	40.3	44.5	48.0	50.2
Traj. aligned baseline (block-level)	52.9	55.8	59.4	63.6	66.7
Matching Nets [61]	46.1	48.6	53.3	56.2	58.8
Matching Nets [61]+2V	47.2	50.7	55.4	57.7	60.2
Prototypical Net [54]	47.2	51.1	54.3	58.9	63.0
Prototypical Net [54]+2V	49.8	53.1	56.7	60.9	64.3
TAP [55]	54.2	57.3	61.7	64.7	68.3
S^2GC (no soft-DTW)	50.8	54.7	58.8	60.2	62.8
soft-DTW	53.7	56.2	60.0	63.9	67.8
(no soft-DTW)+Transf.	56.0	64.2	67.3	70.2	72.9
soft-DTW+Transf.	57.3	66.1	68.8	72.3	74.0
JEANIE+Transf.	**65.0**	**75.2**	**76.7**	**78.9**	**80.0**

which we perform the viewpoint alignment (*CamVPC+crossval.*) improves results. Enabling the viewpoint alignment for support sequences yields extra improvement. With Transformer (*2V+Transf.*), JEANIE boosts results by $\sim 2\%$.

4.2 Comparisons With the State-of-the-Art Methods

One-shot action recognition (NTU-60). Table 4 shows that aligning query and support trajectories by the angle of torso 3D joint, denoted (*Traj. aligned baseline*) is not very powerful, as alluded to in Figure 2 (top). Aligning piece-wise parts (blocks) is better than aligning entire trajectories. In fact, aligning individual frames by torso to the frontal view (*Each frame to frontal view*) and aligning block average of torso direction to the frontal view (*Each block to frontal view*)) were marginally better. We note these baselines use soft-DTW. We show more comparisons in Appendix Sec. E. Our JEANIE with Transformer (*JEANIE+Transf.*) outperforms soft-DTW with Transformer (*soft-DTW+Transf.*) by 7.46% on average.

One-shot action recognition (NTU-120) . Table 5 shows that JEANIE outperforms recent SL-DML and Skeleton-DML by 6.1% and 2.8% respectively (100 training classes). For comparisons, we extended the view adaptive neural networks [81] by combining them with Prototypical Net [54]. VA-RNN+VA-CNN [81] uses 0.47M+24M parameters with random rotation augmentations while JEANIE uses 0.25–0.5M params. Their *rotation+translation* keys are not proven to perform smooth optimal alignment as JEANIE. In contrast, d_{JEANIE} performs jointly a smooth viewpoint-temporal alignment via a principled transportation plan ($\geq$3 dim. space) by design. Their use Euler angles which are a worse option than the camera projection of JEANIE. We notice that ProtoNet+VA backbones is 12% worse than our JEANIE. Even if we split skeletons into blocks to let soft-DTW perform temporal alignment of prototypes and query, JEANIE is still 4–6% better. JEANIE outperforms FVM by 2-4%. This shows that seeking jointly the best temporal-viewpoint alignment is more valuable than considering viewpoint alignment as a local task (free range alignment per each step of soft-DTW).

Table 5: Experimental results on NTU-120 (S^2GC backbone). Methods use temporal alignment by soft-DTW or JEANIE (joint temporal-viewpoint alignment) except VA [80,81] and other cited works. For VA*, we used soft-DTW on temporal blocks while VA generated temporal blocks.

# Training Classes	20	40	60	80	100
APSR [39]	29.1	34.8	39.2	42.8	45.3
SL-DML [45]	36.7	42.4	49.0	46.4	50.9
Skeleton-DML [44]	28.6	37.5	48.6	48.0	54.2
Prototypical Net+VA-RNN(aug.) [80]	25.3	28.6	32.5	35.2	38.0
Prototypical Net+VA-CNN(aug.) [81]	29.7	33.0	39.3	41.5	42.8
Prototypical Net+VA-fusion(aug.) [81]	29.8	33.2	39.5	41.7	43.0
Prototypical Net+VA*-fusion(aug.) [81]	33.3	38.7	45.2	46.3	49.8
TAP [55]	31.2	37.7	40.9	44.5	47.3
S^2GC(no soft-DTW)	30.0	35.9	39.2	43.6	46.4
soft-DTW	30.3	37.2	39.7	44.0	46.8
(no soft-DTW)+Transf.	31.2	37.5	42.3	47.0	50.1
soft-DTW+Transf.	31.6	38.0	43.2	47.8	51.3
FVM+Transf.	34.5	41.9	44.2	48.7	52.0
JEANIE+Transf.	**38.5**	**44.1**	**50.3**	**51.2**	**57.0**

Table 6: Experiments on 2D and 3D Kinetics-skeleton. Note that we have no results on JEANIE or FVM for 2D coordinates (aligning viewpoints is an operation in 3D).

	S^2GC (no soft-DTW)	soft-DTW	*FVM*	JEANIE	JEANIE +Transf.
2D skel.	32.8	34.7	-	-	-
3D skel.	35.9	39.6	44.1	**50.3**	**52.5**

JEANIE on the Kinetics-skeleton . We evaluate our proposed model on both 2D and 3D Kinetics-skeleton. We split the whole dataset into 200 actions for training, and the rest half for testing. As we are unable to estimate the camera location, we simply use Euler angles for the camera viewpoint simulation. Table 6 shows that using 3D skeletons outperforms the use of 2D skeletons by 3-4%, and JEANIE outperforms the baseline (temporal alignment only) and Free Viewpoint Matching (FVM, for every step of DTW, seeks the best local viewpoint alignment, thus realizing non-smooth temporal-viewpoint path in contrast to JEANIE) by around 5% and 6%, respectively. With the transformer, JEANIE further boosts results by 2%.

Few-shot multiview classification . Table 7 (UWA3D Multiview Activity II) shows that adding temporal alignment to SGC, APPNP and S^2GC improves the performance, and the big performance gain is obtained by further adding the viewpoint alignment. As this dataset is challenging in recognizing the actions from a novel view point, our proposed method performs consistently well on all different combinations of training/testing viewpoint variants. This is predictable as our method aligns both temporal and camera viewpoints which allows a robust classification. JEANIE outperforms FVM by 4.2%, and outperforms the baseline (with temporal alignment only) by 7% on average.

Table 7: Experiments on the UWA3D Multiview Activity II.

Training view	V_1 & V_2		V_1 & V_3		V_1 & V_4		V_2 & V_3		V_2 & V_4		V_3 & V_4		Mean
Testing view	V_3	V_4	V_2	V_4	V_2	V_3	V_1	V_4	V_1	V_3	V_1	V_2	
GCN	36.4	26.2	20.6	30.2	33.7	22.4	43.1	26.6	16.9	12.8	26.3	36.5	27.6
SGC	40.9	60.3	44.1	52.6	48.5	38.7	50.6	52.8	52.8	37.2	57.8	49.6	48.8
+soft-DTW	43.9	60.8	48.1	54.6	52.6	45.7	54.0	58.2	56.7	40.2	60.2	51.1	52.2
+JEANIE	47.0	62.8	50.4	57.8	53.6	47.0	57.9	62.3	57.0	44.8	61.7	52.3	54.6
APPNP	42.9	61.9	47.8	58.7	53.8	44.0	52.3	60.3	55.1	38.2	58.3	47.9	51.8
+soft-DTW	44.3	63.2	50.7	62.3	53.9	45.0	56.9	62.8	56.4	39.3	60.1	51.9	53.9
+JEANIE	46.8	64.6	51.3	65.1	54.7	46.4	58.2	65.1	58.8	43.9	60.3	52.5	55.6
S^2GC	45.5	64.4	46.8	61.6	49.5	43.2	57.3	61.2	51.0	42.9	57.0	49.2	52.5
+soft-DTW	48.2	67.2	51.2	67.0	53.2	46.8	62.4	66.2	57.8	45.0	62.2	53.0	56.7
+FVM	50.7	68.8	56.3	69.2	55.8	47.1	63.7	68.8	62.5	51.4	63.8	55.7	59.5
+JEANIE	**55.3**	**70.2**	**61.4**	**72.5**	**60.9**	**50.8**	**66.4**	**73.9**	**68.8**	**57.2**	**66.7**	**60.2**	**63.7**

Table 8: Results on NTU-120 (multiview classification). Baseline is soft-DTW + S^2GC.

Training view	bott.	bott.	bott.& cent.	left	left	left & cent.
Testing view	cent.	top	top	cent.	right	right
100/same 100 (baseline)	74.2	73.8	75.0	58.3	57.2	68.9
100/same 100 (*FVM*)	79.9	78.2	80.0	65.9	63.9	75.0
100/same 100 (*JEANIE*)	**81.5**	**79.2**	**83.9**	**67.7**	**66.9**	**79.2**
100/novel 20 (baseline)	58.2	58.2	61.3	51.3	47.2	53.7
100/novel 20 (*FVM*)	66.0	65.3	68.2	58.8	53.9	60.1
100/novel 20 (*JEANIE*)	**67.8**	**65.8**	**70.8**	**59.5**	**55.0**	**62.7**

Table 8 (NTU-120) shows that adding more camera viewpoints to the training process helps the multiview classification. Using bottom and center views for training and top view for testing, or using left and center views for training and the right view for testing yields 4% gain ('*same 100*' means the same train/test classes but different views). Testing on 20 novel classes ('*novel 20*' never used in training) yields 62.7% and 70.8% for multiview classification in horizontal and vertical camera viewpoints, respectively.

5 Conclusions

We have proposed a Few-shot Action Recognition (FSAR) approach for learning on 3D skeletons via JEANIE. We have demonstrated that the joint alignment of temporal blocks and simulated viewpoints of skeletons between support-query sequences is efficient in the meta-learning setting where the alignment has to be performed on new action classes under the low number of samples. Our experiments have shown that using the stereo camera geometry is more efficient than simply generating multiple views by Euler angles in the meta-learning regime. Most importantly, we have introduced a novel FSAR approach that learns on articulated 3D body joints.

Acknowledgements We thank Dr. Jun Liu (SUTD) for discussions on FSAR for 3D skeletons, and CSIRO's Machine Learning and Artificial Intelligence Future Science Platform (MLAI FSP).

References

1. Euler angles. Wikipedia, https://en.wikipedia.org/wiki/Euler_angles, accessed: 08-03-2022 2
2. Lecture 12: Camera projection. On-line, http://www.cse.psu.edu/~rtc12/CSE486/lecture12.pdf, accessed: 08-03-2022 2
3. Bart, E., Ullman, S.: Cross-generalization: Learning novel classes from a single example by feature replacement. CVPR pp. 672–679 (2005) 1
4. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., Cox, D.D.: Hyperopt: a python library for model selection and hyperparameter optimization. CSD **8**(1), 014008 (2015) 11
5. Cao, K., Ji, J., Cao, Z., Chang, C.Y., Niebles, J.C.: Few-shot video classification via temporal alignment. In: CVPR (2020) 1, 4
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017) 9
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017) 1
8. Catalin, Ionescu, Dragos, Papava, Vlad, Olaru, Cristian, Sminchisescu: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI (2014) 9
9. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: CVPR (2020) 3
10. Cuturi, M.: Fast global alignment kernels. In: ICML (2011) 2
11. Cuturi, M., Blondel, M.: Soft-dtw: a differentiable loss function for time-series. In: ICML (2017) 2
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020) 3, 5, 6
13. Dvornik, N., Schmid, C., Mairal1, J.: Selecting relevant features from a multi-domain representation for few-shot classification. In: ECCV (2020) 1
14. Dwivedi, S.K., Gupta, V., Mitra, R., Ahmed, S., Jain, A.: Protogan: Towards few shot learning for action recognition. arXiv (2019) 1, 4
15. Elsken, T., Staffler, B., Metzen, J.H., Hutter, F.: Meta-learning of neural architectures for few-shot learning. In: CVPR (2020) 1
16. Fei, N., Guan, J., Lu, Z., Gao, Y.: Few-shot zero-shot learning: Knowledge transfer with less supervision. In: ACCV (2020) 1
17. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE TPAMI **28**(4), 594–611 (2006) 1
18. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: CVPR (2017) 1
19. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016) 1
20. Fink, M.: Object classification from a single example utilizing class relevance metrics. NeurIPS pp. 449–456 (2005) 1
21. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (eds.) ICML. vol. 70, pp. 1126–1135. PMLR (2017) 1
22. Guan, J., Zhang, M., Lu, Z.: Large-scale cross-domain few-shot learning. In: ACCV (2020) 1
23. Guo, M., Chou, E., Huang, D.A., Song, S., Yeung, S., Fei-Fei, L.: Neural graph matching networks for fewshot 3d action recognition. In: ECCV. pp. 653–669 (2018) 1, 4

24. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: ECCV (2020) 1
25. Haasdonk, B., Burkhardt, H.: Invariant kernel functions for pattern analysis and machine learning. Mach. Learn. **68**(1), 35–61 (2007) 4
26. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. arXiv (2017) 9
27. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017) 3, 4, 6
28. Klicpera, J., Bojchevski, A., Gunnemann, S.: Predict then propagate: Graph neural networks meet personalized pagerank. In: ICLR (2019) 3, 4, 6
29. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2 (2015) 1
30. Koniusz, P., Wang, L., Cherian, A.: Tensor representations for action recognition. IEEE TPAMI (2020) 1
31. Koniusz, P., Wang, L., Sun, K.: High-order tensor pooling with attention for action recognition. arXiv (2021) 1
32. Koniusz, P., Zhang, H.: Power normalizations in fine-grained image, few-shot image and graph classification. IEEE Trans. Pattern Anal. Mach. Intell. **44**(2), 591–609 (2022) 1
33. Lake, B.M., Salakhutdinov, R., Gross, J., Tenenbaum, J.B.: One shot learning of simple visual concepts. CogSci (2011) 1
34. Li, F.F., VanRullen, R., Koch, C., Perona, P.: Rapid natural scene categorization in the near absence of attention. PNAS **99**(14), 9596–9601 (2002) 1
35. Li, K., Zhang, Y., Li, K., Fu, Y.: Adversarial feature hallucination networks for few-shot learning. In: CVPR (2020) 1
36. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: CVPR (2019) 3
37. Lichtenstein, M., Sattigeri, P., Feris, R., Giryes, R., Karlinsky, L.: Tafssl: Task-adaptive feature sub-space learning for few-shot classification. In: ECCV (2020) 1
38. Liu, J., Wang, G., Hu, P., Duan, L., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: CVPR. pp. 3671–3680 (2017) 2, 4
39. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE TPAMI (2019) 2, 3, 4, 9, 13
40. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: CVPR (2020) 3
41. Lu, C., Koniusz, P.: Few-shot keypoint detection with uncertainty learning for unseen species. CVPR (2022) 1
42. Luo, Q., Wang, L., Lv, J., Xiang, S., Pan, C.: Few-shot learning via feature hallucination with variational inference. In: WACV (2021) 1
43. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV. pp. 2659–2668 (2017) 9
44. Memmesheimer, R., Häring, S., Theisen, N., Paulus, D.: Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition. arXiv (2021) 2, 4, 13
45. Memmesheimer, R., Theisen, N., Paulus, D.: Signal level deep metric learning for multimodal one-shot action recognition. arXiv (2020) 2, 4, 13
46. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. CVPR **1**, 464–471 (2000) 1
47. Mishra, A., Verma, V.K., Reddy, M.S.K., Arulkumar, S., Rai, P., Mittal, A.: A generative approach to zero-shot and few-shot action recognition. In: WACV. pp. 372–380 (2018) 1, 4

48. Qin, Z., Liu, Y., Ji, P., Kim, D., Wang, L., McKay, B., Anwar, S., Gedeon, T.: Fusing higher-order features in graph neural networks for skeleton-based action recognition. IEEE TNNLS (2022) 2
49. Rahmani, H., Mahmood, A., Huynh, D.Q., Mian, A.: Histogram of Oriented Principal Components for Cross-View Action Recognition. IEEE TPAMI pp. 2430–2443 (2016) 3, 8, 9
50. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: CVPR (2016) 3, 4, 9
51. Simon, C., Koniusz, P., Harandi, M.: On learning the geodesic path for incremental learning. In: CVPR. pp. 1591–1600 (2021) 1
52. Simon, C., Koniusz, P., Nock, R., Harandi, M.: On modulating the gradient for meta-learning. In: ECCV (2020) 1
53. Smola, A.J., Kondor, R.: Kernels and regularization on graphs. COLT (2003) 3
54. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) NeurIPS. pp. 4077–4087 (2017) 1, 12
55. Su, B., Wen, J.R.: Temporal alignment prediction for supervised representation learning and few-shot sequence classification. In: ICLR (2022) 12, 13
56. Sun, K., Koniusz, P., Wang, Z.: Fisher-Bures adversary graph convolutional networks. UAI **115**, 465–475 (2019) 3
57. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: CVPR. pp. 1199–1208 (2018) 1
58. Tang, L., Wertheimer, D., Hariharan, B.: Revisiting pose-normalization for fine-grained few-shot recognition. In: CVPR (2020) 1
59. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015) 1
60. Villani, C.: Optimal Transport, Old and New. Springer (2009) 3
61. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) NeurIPS. pp. 3630–3638 (2016) 1, 12
62. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE TPAMI **41**(11), 2740–2755 (2019) 2
63. Wang, L.: Analysis and Evaluation of Kinect-based Action Recognition Algorithms. Master's thesis, School of the Computer Science and Software Engineering, The University of Western Australia (2017) 1, 9
64. Wang, L., Huynh, D.Q., Koniusz, P.: A comparative review of recent kinect-based action recognition algorithms. IEEE TIP **29**, 15–28 (2020) 1, 4, 9
65. Wang, L., Huynh, D.Q., Mansour, M.R.: Loss switching fusion with similarity search for video classification. ICIP (2019) 1
66. Wang, L., Koniusz, P.: Self-supervising action recognition by statistical moment and subspace descriptors. In: ACM-MM. p. 4324–4333 (2021) 1
67. Wang, L., Koniusz, P.: Uncertainty-DTW for time series and sequences. ECCV (2022) 1, 2
68. Wang, L., Koniusz, P., Huynh, D.Q.: Hallucinating IDT descriptors and I3D optical flow features for action recognition with cnns. In: ICCV (2019) 1
69. Wang, L., Ding, Z., Tao, Z., Liu, Y., Fu, Y.: Generative multi-view human action recognition. In: ICCV (2019) 4
70. Wang, S., Yue, J., Liu, J., Tian, Q., Wang, M.: Large-scale few-shot learning via multi-modal knowledge discovery. In: ECCV (2020) 1
71. Wang, Y., Long, M., Wang, J., Yu, P.S.: Spatiotemporal pyramid network for video action recognition. In: CVPR (2017) 1
72. Wu, F., Zhang, T., de Souza Jr., A.H., Fifty, C., Yu, T., Weinberger, K.Q.: Simplifying graph convolutional networks. In: ICML (2019) 3, 4, 6

73. Xu, B., Ye, H., Zheng, Y., Wang, H., Luwang, T., Jiang, Y.G.: Dense dilated network for few shot action recognition. In: ACM ICMR. pp. 379–387 (2018) 1, 4
74. Yan, S., Xiong, Y., Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In: AAAI (2018) 3, 9
75. Yu, X., Zhuang, Z., Koniusz, P., Li, H.: 6DoF object pose estimation via differentiable proxy voting regularizer. In: BMVC. BMVA Press (2020) 1
76. Zhang, H., Koniusz, P.: Power normalizing second-order similarity network for few-shot learning. In: WACV. pp. 1185–1193 (2019) 1
77. Zhang, H., Koniusz, P., Jian, S., Li, H., Torr, P.H.S.: Rethinking class relations: Absolute-relative supervised and unsupervised few-shot learning. In: CVPR. pp. 9432–9441 (June 2021) 1
78. Zhang, H., Li, H., Koniusz, P.: Multi-level second-order few-shot learning. IEEE Transactions on Multimedia (2022) 1
79. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: ECCV (2020) 1, 4
80. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: ICCV (2017) 2, 4, 13
81. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive neural networks for high performance skeleton-based human action recognition. IEEE TPAMI **41**(8), 1963–1978 (2019) 2, 4, 12, 13
82. Zhang, S., Luo, D., Wang, L., Koniusz, P.: Few-shot object detection by second-order pooling. In: ACCV. Lecture Notes in Computer Science, vol. 12625, pp. 369–387. Springer (2020) 1
83. Zhang, S., Murray, N., Wang, L., Koniusz, P.: Time-rEversed diffusioN tEnsor Transformer: A new TENET of Few-Shot Object Detection. In: ECCV (2022) 1
84. Zhang, S., Wang, L., Murray, N., Koniusz, P.: Kernelized few-shot object detection with efficient integral aggregation. In: CVPR. pp. 19207–19216 (June 2022) 1
85. Zhu, H., Koniusz, P.: Simple spectral graph convolution. In: ICLR (2021) 3, 4, 6
86. Zhu, H., Koniusz, P.: EASE: Unsupervised discriminant subspace learning for transductive few-shot learning. CVPR (2022) 1
87. Zhu, H., Sun, K., Koniusz, P.: Contrastive laplacian eigenmaps. In: NeurIPS. pp. 5682–5695 (2021) 3
88. Zhu, L., Yang, Y.: Compound memory networks for few-shot video classification. In: ECCV (2018) 4