This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Neuron: Learning Context-Aware Evolving Representations for Zero-Shot Skeleton Action Recognition

Yang Chen[1], Jingcai Guo[1*], Song Guo[2], Dacheng Tao[3]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR
[2]Department of CSE, The Hong Kong University of Science and Technology, Hong Kong SAR
[3]College of Computing and Data Science, Nanyang Technological University, Singapore
jc-jingcai.guo@polyu.edu.hk

Abstract

*Zero-shot skeleton action recognition is a non-trivial task that requires robust unseen generalization with prior knowledge from only seen classes and shared semantics. Existing methods typically build the skeleton-semantics interactions by uncontrollable mappings and conspicuous representations, thereby can hardly capture the intricate and fine-grained relationship for effective cross-modal transferability. To address these issues, we propose a novel dyNamically Evolving dUal skeleton-semantic syneRgistic framework with the guidance of cOntext-aware side informatioN (dubbed **Neuron**), to explore more fine-grained cross-modal correspondence from micro to macro perspectives at both spatial and temporal levels, respectively. Concretely, 1) we first construct the spatial-temporal evolving micro-prototypes and integrate dynamic context-aware side information to capture the intricate and synergistic skeleton-semantic correlations step-by-step, progressively refining cross-model alignment; and 2) we introduce the spatial compression and temporal memory mechanisms to guide the growth of spatial-temporal micro-prototypes, enabling them to absorb structure-related spatial representations and regularity-dependent temporal patterns. Notably, such processes are analogous to the learning and growth of neurons, equipping the framework with the capacity to generalize to novel unseen action categories. Extensive experiments on various benchmark datasets demonstrated the superiority of the proposed method* [1].

1. Introduction

Action recognition has gained considerable attention for years due to its broad applications, *e.g.*, sports analysis [37], rehabilitation assessment [15], anomaly action recognition [29], etc. Benefiting from the excellent privacy-preserving and data-efficiency properties of skeleton modality, a wealth of innovative research has sprung up to advance this field. Among these developments, several challenging and realistic issues have emerged, notably spurring interest in zero-shot skeleton action recognition [6, 32, 41].

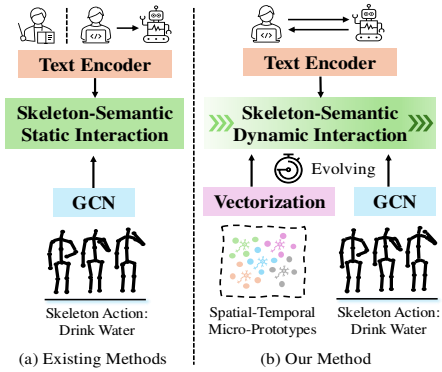


Figure 1. Method sketches: (a) Existing methods employ homogeneous semantics (manual-designed or one-turn LLMs) to align the skeleton space with one-step (static); (b) Our Neuron introduces the context-aware side information (multi-turn LLMs) and evolving micro-prototypes to capture cross-modal correspondence from micro to macro perspectives (dynamic) for controllable alignment.

Conventional supervised and self-supervised skeleton-based action recognition methods [6, 7] are usually hindered by their dependence on expensive annotated data and limited generalization to novel unseen categories. These methods generally confine their classification abilities to actions present in both training and test sets [5, 8, 38, 40]. In contrast, zero-shot skeleton action recognition effectively addresses the above dilemmas by leveraging only limited seen skeleton samples and a predefined semantic corpus, thus enabling the recognition of previously unseen action categories [6, 13, 18, 20, 32, 39, 41]. Furthermore, such a zero-shot learning (ZSL) setting can be extended to generalized zero-shot learning (GZSL), wherein both seen and unseen

*Jingcai Guo is the corresponding author.
[1]Code is available at: https://github.com/cseeyangchen/Neuron.

categories are recognized, making the task more challenging yet more applicable to real-world scenarios. The common solutions involve establishing robust cross-modal relationships by embedding skeleton and semantic data within a joint space, forming a bridge that transfers knowledge from seen to unseen actions [13, 18, 20, 39]. In pursuit of improved generalization, recent studies have further sought ways for fine-grained alignment by decomposing representations either explicitly [6, 41] or implicitly [32].

Despite the progress made, crucial challenges remain unsolved, *i.e.*, 1) existing semantics are limited to manual-designed category names or one-turn LLMs-generated action descriptions (Fig. 1 (a)), leading to high homogeneity semantics among similar actions. Taking "*kicking*" and "*side kick*" as samples, homogeneous semantics have made the joint embedding space indivisible among these highly similar inter-class actions. Additionally, simple action descriptions cannot encapsulate the intra-class diversity of skeletons, where spatial variations (*e.g.*, movement scale) and temporal differences (*e.g.*, speed) vary across individuals; and 2) directly aligning skeleton and semantic features into a joint embedding space lacks fine-grained control over this process, which causes the model to overemphasize irrelevant or shortcut semantics, *i.e.*, they cannot locate the vital skeleton features with intended semantics correctly, resulting in suboptimal cross-modal fine-grained correspondence. Inspired by the growth process of neurons from small to large that can learn and adapt incrementally through repeated stimuli to form complex connections that enable generalization, we consider designing a dynamic and controllable fine-grained aligning framework from micro to macro perspectives that can refine the interactions incrementally by the guidance of context-aware side information, allowing better generalization to unseen actions.

Specifically, we propose a novel dynamically evolving dual skeleton-semantic synergistic framework via the guidance from context-aware side information, termed ***Neuron***, to enable controlled fine-grained cross-modal interactions (Fig. 1 (b)). Firstly, we employ LLMs (*e.g.*, GPT-4o) with multi-turn interactive prompts to generate enriched, contextual side information, providing detailed semantic descriptions of skeletons that span spatial and temporal variations. Spatial variations range from coarse to mid- and fine-grained structural granularity, while temporal dynamics cover the evolution of movements from start to mid and end phases. With this contextual semantic guidance, we construct spatial-temporal micro-prototypes to capture the synergistic skeleton-semantic relationship iteratively. The spatial micro-prototype evolves through the spatial compression mechanism, guided by stepwise semantic granularity upgrades to isolate discriminative structure-related spatial representations. Similarly, the temporal micro-prototype grows dynamically through the temporal memory mechanism, directed by temporal phase-evolution semantics in turn, to search effective regularity-dependent temporal patterns. Both micro-prototypes are updated iteratively from the micro to the macro perspectives, emulating neuron-like incremental growth to refine their understanding of skeleton actions. This process enhances cross-modal transferability, empowering the framework to generalize effectively to unseen action categories.

Alignment		Adaptability		Semantic Source			Method
GL	FG	ST	DY	Manual	H-LLMs	C-LLMs	
✓	✗	✓	✗	✓	✗	✗	RelationNet [18], SynSE [13]
✓	✗	✓	✗	✓	✓	✗	SMIE [39]
✗	✓	✓	✗	✓	✓	✗	PURLS [41], STAR [6]
✗	✓	✓	✗	✓	✗	✗	SA-DAVE [32]
✗	✓	✗	✓	✗	✗	✓	**Neruon (ours)**

Table 1. Comparison with existing methods. The GL and FG are denoted as global and fine-grained. The ST and DY are represented as static and dynamic. Manual, H-LLMs and C-LLMs represent manual-designed category names, LLMs-generated homogeneous descriptions, and context-aware side information, indicating the relatively Low, Mid and High semantic strength.

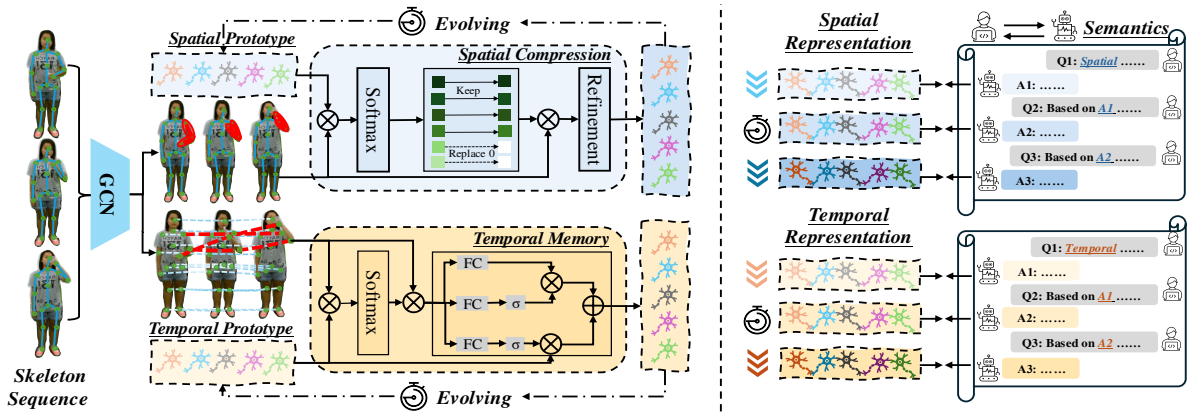
The main contributions can be summarized as follows:

- We introduce dynamic, context-aware side information to guide the spatial-temporal micro-prototypes in evolving from micro to macro perspectives, achieving controlled, fine-grained cross-modal alignment.
- We propose the spatial compression and temporal memory mechanisms, enabling the micro-prototypes to explore structure-related spatial representations and regularity-dependent temporal patterns.
- Extensive experiments show that the proposed method achieves state-of-the-art results in both ZSL and GZSL settings on NTU RGB+D, NTU RGB+D 120, and PKU-MMD datasets.

2. Related Work

2.1. Zero-Shot Skeleton Action Recognition

Zero-shot skeleton action recognition methods aim to establish the relationship between the seen skeletons and their associated semantics to recognize novel action categories. A pioneering approach, RelationNet [18], identified novel categories by learning deep non-linear metrics between global skeleton features and category names in an embedding manner. Differently, SynSE [13] paves the generative way using VAEs to align global skeleton features with verb-noun embeddings of category names. SMIE [39] employs the temporal constraint with maximizing mutual information to achieve global alignment. However, these methods rely on global alignment with manual-designed category names, limiting the granularity and diversity of captured semantics. Recently, PURLS [41] and STAR [6] explicitly decompose skeletons into overlapping parts with LLMs-generated part-specific descriptions, enabling fine-grained alignment. Unfortunately, these generated descriptions are homogenous



(a) Evolving Representation Learning (b) Stepwise Cross-Modal Alignment

Figure 2. The pipeline of the proposed method. (a) represents the evolving spatial-temporal representation learning process. (b) shows the stepwise skeleton-semantic alignment.

across actions, reducing their discriminative capacity. SA-DAVE [32] improves generalization by decoupling skeleton representations into semantic-related and irrelevant components. **[*Our contribution*]**: We design a dynamic, evolving, and synergistic alignment framework to explore better cross-modal correspondence with context-aware evolving representation from LLMs.

2.2. Multi-modal Skeleton Representation Learning

Inspired by the success of CLIP [27], numerous studies have explored multi-modal fusion or cross-model distillation strategies to enhance skeleton representation capacity. These approaches incorporate data from complementary modalities, such as RGB videos [2, 7, 19, 22, 33], depth sequences [10], text descriptions [7, 24, 36], flow information [25, 34], and sensor signals [26]. The core of the multi-modal fusion methods [2, 10, 14, 19, 22, 33, 34] is to compensate for the shortcomings of skeletons by relying on the strengths of other modalities, while the cross-modal distillation methods [7, 24–26, 36] aim to capture modality-agnostic representations for action recognition. Nevertheless, both methods lack explicit control over the representation learning process, resulting in models that capture shortcut features. Recently, some methods have made strides in controlled representation learning [1, 11, 23], yet not in the skeleton action recognition area. **[*Our contribution*]**: We propose spatial compression and temporal memory mechanisms to capture skeleton representations in a controlled manner, progressing from micro to macro perspectives.

3. Method

3.1. Problem Formulation

For the zero-shot skeleton action recognition problem, the skeleton dataset $\mathcal{D} = \{x_i, y_i, a_i\}_{i=1}^{N}$ with $|\mathcal{Y}|$ categories can be divided into three overlapping parts: the train-set $\mathcal{D}_{tr}^{s}$ with seen categories $\mathcal{Y}^s$, the test-set $\mathcal{D}_{te}^{u}$ with unseen categories $\mathcal{Y}^u$, and another test-set $\mathcal{D}_{te}^{s}$ with seen categories $\mathcal{Y}^s$. Here, $x_i \in \mathbb{R}^{3\times T\times V\times M}$ represents the skeleton sequence, where 3 corresponds to the 3D coordinates of human joints, T denotes the sequence frames, V is the number of human joints, and M indicates the human numbers. The symbol $y_i \in \mathcal{Y}$ is the corresponding action label, where $a_i \in \mathcal{A}$ signifies the associated semantics. Notably, $\mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u$ and $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$. For each action category, a manual-designed category name and one-turn LLMs-generated action description are adopted as the semantics a_i in prior works [6, 13, 18, 32, 39, 41]. During the training phase, we establish the relationship between the x_i^s and a_i^s within the train-set $\mathcal{D}_{tr}^{s}$. In the inference phase, we need to predict categories in both $\mathcal{D}_{te}^{u}$ and $\mathcal{D}_{te}^{s} \cup \mathcal{D}_{te}^{u}$ under the ZSL and GZSL settings through the above-constructed relationship, respectively. For simplicity, we omit the superscript for seen (s) and unseen (u) categories in the following sections.

3.2. Contextual Side Information Generation

We generate more heterogeneous and multifaceted contextual semantic descriptions of skeletons from the perspective of spatial structure granularity and temporal resolution variation than previous methods [6, 13, 18, 32, 39, 41]. Specifically, we engage the LLM (*e.g.*, GPT-4o) to produce semantic descriptions of skeleton movements in multi-turn inquiries, including the hierarchical spatial structure and evolving temporal resolutions, as shown in Fig. 2 (b). The multi-turn spatial prompts vary at the coarse-, mid-, and fine-grained levels. Differently, the temporal prompts vary across the first, second, and third phases. We present the temporal prompts and obtained semantics of "nod head/bow" action as samples: [Q1]: You should split the "nod head/bow" action into three phases along the time axis. Please describe the body movements of temporal variations in the first phase. [A1]: Begin by lowering your head slightly. [Q2]: Based on [A1], describe them in the second phase further. [A2]: Continue by maintain-

ing a slight head. [Q3]: Based on [A2], describe them in the third phase further. [A3]: Finish by returning the head to a neutral position. In the e-th phase, we can obtain N_a state descriptions $\mathcal{A}_y^e$ for the same skeleton category y, represented by different individuals, thereby enriching the intra-class diversity of semantics. With these step-by-step queries, the semantics among the categories become more discriminative, prompting inter-class separability of semantics. Upon this iterative process, we enhance the coherent and contextual representation capabilities of the semantics $\mathcal{A} = \{\mathcal{A}_y^1, ..., \mathcal{A}_y^e, ..., \mathcal{A}_y^{N_e}\}_{y=1}^{|\mathcal{Y}|}$, where N_e is the number of phase. Meanwhile, the evolving property of semantics from LLMs can guide the learning process of skeleton representation for better cross-modal correspondences. Afterward, we utilize pre-trained text encoder $\Psi(\cdot)$ to extract the contextual semantic features $\mathcal{Z} = \{\mathcal{Z}_y\}_{y=1}^{|\mathcal{Y}|}$, where $\mathcal{Z}_y = \{\mathcal{Z}_y^{s_e}, \mathcal{Z}_y^{t_e} \in \mathbb{R}^{N_a \times d}\}_{e=1}^{N_e}$, where $\mathcal{Z}_y^{s_e}$ and $\mathcal{Z}_y^{t_e}$ are the spatial and temporal semantics of y-th action category at e-th phase, d is the dimension of semantic feature.

3.3. Spatial-Temporal Synergistic Growing

Spatial-Temporal Micro-prototypes Construction. Decoupling spatial structures and temporal variations is vital in distinguishing similar actions. As described in [38], temporal variations help differentiate "put things into a bag" and "take things out of a bag", while spatial structures are significant for distinguishing them from "reach into pockets". Thus, we separate spatial and temporal representations for alignment independently. First, we extract the skeleton representations $\mathcal{F} = \Phi(x) \in \mathbb{R}^{\widehat{T} \times \widehat{V} \times C}$ by skeleton encoder $\Phi(\cdot)$, where $\widehat{T} \times \widehat{V}$ represents the spatial-temporal dimension after transformations, C is the feature dimension. Then, we pool the temporal and spatial dimensions respectively to obtain the spatial feature $\mathcal{F}_s \in \mathbb{R}^{\widehat{V} \times C}$ and the temporal feature $\mathcal{F}_t \in \mathbb{R}^{\widehat{T} \times C}$. To effectively align the skeleton and semantic spaces, we design the learnable spatial micro-prototype $\mathcal{P}_s = \{p_k^s \in \mathbb{R}^{C \times 1}\}_{k=1}^{N_s}$ to capture the patterns of skeleton topology structure and construct the learnable temporal micro-prototype $\mathcal{P}_t = \{p_k^t \in \mathbb{R}^{C \times 1}\}_{k=1}^{N_t}$ to record the temporal motion variations of the skeleton, where p_k^s and p_k^t are randomly initialized attributes of skeleton representation (*e.g.*, amplitude and velocity [6]), N_s and N_t are the number of primitives. Guided by semantics $\mathcal{Z}$, $\mathcal{P}_s$ and $\mathcal{P}_t$ progressively learn useful skeleton representations, similar to the growth process of neurons.

Spatial Compression Mechanism. For the given spatial micro-prototype $\mathcal{P}_s$, we aim to enable it to progressively and controllably absorb structure-related spatial patterns from skeletons across three phases, enhancing the cross-modal correspondence at the spatial level from coarse-grained to fine-grained. Specifically, in the first phase, we treat the initialized $\mathcal{P}_s^0$ as the query matrix to calculate similarity scores with spatial skeleton representations $\mathcal{F}^s$:

$$\mathcal{H}^{s_0} = softmax(\mathcal{F}_s \mathcal{P}_s^0), \tag{1}$$

where $h_{j,k}^{s_0} \in \mathcal{H}^{s_0}$ denotes the similarity between the j-th joint feature and the k-th spatial micro-prototype attribute at the coarse-grained level. Normally, a straightforward approach is to aggregate the joint features to obtain the updated micro-prototype for alignment with semantics:

$$\mathcal{P}_s^1 = \varphi_s(\mathcal{F}_s^T \mathcal{H}^{s_0}), \tag{2}$$

where $\varphi_s(\cdot)$ is the refinement function composed of MLPs. However, not all the micro-prototype attributes are relevant for each joint across different samples; aggregating all attributes may inadvertently overemphasize those unrelated to the action itself (*e.g.*, camera views). To mitigate this issue, we selectively drop joint-unrelated attributes, thus avoiding shortcut feature learning. We sort the matrix $\mathcal{H}^{s_0}$ row-wise and partition it into two subsets with the hyper-parameter $\alpha \in [0, 1]$. Then, we maintain the top-α scores and replace others with zeros. Using the modified matrix $\hat{\mathcal{H}}^{s_0}$ in Eq 2, we can get the updated micro-prototype $\mathcal{P}_s^1$. Considering coarse-grained exploration alone is insufficient, we repeat the above operations at the mid-grained and fine-grained levels further. Meanwhile, α is incrementally increased in each phase to allow for more precise attribute selection. Through these controllable compression processes, the spatial micro-prototype is growing to $\mathcal{P}_s^3$, capturing more general and concise spatial pattern representations.

Temporal Memory Mechanism. Compared to the spatial micro-prototype $\mathcal{P}_s$ growing through the refinement of hierarchical structure, the temporal micro-prototype $\mathcal{P}_t$ develops alongside the evolution of resolution. Unlike the complete spatial feature $\mathcal{F}_s$, we split the temporal feature into three phases to obtain segment features $\mathcal{F}_t = \{\mathcal{F}_t^e \in \mathbb{R}^{\frac{\widehat{T}}{3} \times C}\}_{e=1}^{N_e}$ for progressive alignment. Initially, we calculate the similarity scores between the initialized P_t^0 and 1-st temporal feature F_t^1, and then aggregate all frame features to update the temporal micro-prototype as follows:

$$\mathcal{H}^{t_0} = softmax(\mathcal{F}_t^0 \mathcal{P}_t^0), \tag{3}$$

$$\hat{\mathcal{P}}_t^0 = (\mathcal{F}_t^0)^T \mathcal{H}^{t_0}, \tag{4}$$

We do not adopt the compression mechanism here because the sequential temporal features in each phase can constrain the learning process not stuck into the short features. However, the updated temporal micro-prototype in the e-th phase faces the challenge of knowledge oblivion that is learned in the $(e-1)$-th phase, as shown in the first row of Fig. 5 (a) - 5 (c). To address this, in each phase, we introduce the memory mechanism to remember new information

while selectively recalling the knowledge from the previous phase:

$$\mathcal{P}_t^1 = \underbrace{\sigma(\rho_r(\hat{\mathcal{P}}_t^0)) \cdot \mathcal{P}_t^0}_{recall} + \underbrace{\sigma(\rho_m(\hat{\mathcal{P}}_t^0)) \cdot \varphi_t(\hat{\mathcal{P}}_t^0)}_{remember}, \quad (5)$$

where $\sigma(\cdot)$ is the sigmoid function to quantize the remember and recall probability, $\rho_r(\cdot)$ and $\rho_m(\cdot)$ are MLPs for projection, $\varphi_t(\cdot)$ is the refine functions. Through the evolution from the first to the third phase, the micro-prototype $\mathcal{P}_t^0$ progressively captures the regularity-dependent temporal clues, controllable forming to the $\mathcal{P}_t^3$ at the macro level.

3.4. Overall Objective

To align the skeleton and semantic spaces at the fine-grained level, we start from the coarse-grained level, step-by-step, instructing micro-prototypes to update iteratively and controllably from micro to macro perspectives. For each e-th phase, the spatial-temporal micro-prototypes $\mathcal{P}_s^e \in \mathbb{R}^{C \times N_s}$ and $\mathcal{P}_t^e \in \mathbb{R}^{C \times N_t}$ contain the specific skeleton patterns for different attributes, so we pool all the channels to produce the skeleton representation $\mathcal{X}_s^e, \mathcal{X}_t^e \in \mathbb{R}^C$. Meanwhile, we pool all the contextual semantic features $\mathcal{Z}_y^{s_e}, \mathcal{Z}_y^{t_e} \in \mathbb{R}^{Na \times d}$ into $\hat{\mathcal{Z}}_y^{s_e}, \hat{\mathcal{Z}}_y^{t_e} \in \mathbb{R}^d$. Then, we define the alignment operation as:

$$\mathcal{L}_s^e = -log(\frac{exp(\phi_s(\mathcal{X}_s^e)^T \psi(\hat{\mathcal{Z}}_y^{s_e}))}{\sum_{k \in \mathcal{Y}^s} exp(\phi_s(\mathcal{X}_s^e)^T \psi(\hat{\mathcal{Z}}_k^{s_e}))}), \quad (6)$$

$$\mathcal{L}_t^e = -log(\frac{exp(\phi_t(\mathcal{X}_t^e)^T \psi(\hat{\mathcal{Z}}_y^{t_e}))}{\sum_{k \in \mathcal{Y}^s} exp(\phi_t(\mathcal{X}_t^e)^T \psi(\hat{\mathcal{Z}}_k^{t_e}))}), \quad (7)$$

where $\phi_s(\cdot)$, $\phi_t(\cdot)$, and $\psi(\cdot)$ are project functions. Last, the overall optimization objective is defined as:

$$\mathcal{L} = \sum_{e=1}^{N_e} (\mathcal{L}_s^e + \mathcal{L}_t^e). \quad (8)$$

3.5. ZSL/GZSL Prediction

In the inference stage, we can contain the spatial-temporal features $\mathcal{X}_s^m, \mathcal{X}_t^m$ that are updated by micro-prototypes. Then, we can predict the unseen skeleton category by calculating the similarity with unseen semantics. Furthermore, the calibrated stacking method [4] is employed to mitigate the domain shift for GZSL prediction. The whole process is formulated as follows:

$$\hat{y}^s = \underset{y \in \mathcal{Y}^u/\mathcal{Y}}{arg\max} \rho_s(\mathcal{X}_s^3)^T \phi(\hat{\mathcal{Z}}_y^{s_3}) - \gamma_s \mathbb{I}[y \in \mathcal{Y}^s], \quad (9)$$

$$\hat{y}^t = \underset{y \in \mathcal{Y}^u/\mathcal{Y}}{arg\max} \rho_t(\mathcal{X}_t^3)^T \phi(\hat{\mathcal{Z}}_y^{t_3}) - \gamma_t \mathbb{I}[y \in \mathcal{Y}^s], \quad (10)$$

where $\mathcal{Y}^u/\mathcal{Y}$ corresponds to the ZSL/GZSL setting, and γ_s/γ_t are calibration factors. As mentioned above, different actions have different propensities for spatial and temporal, so we define the final prediction as $\hat{y}^* = \{\hat{y}^s, \hat{y}^t\}$.

4. Experiments

4.1. Datasets

NTU RGB+D 60 [31]. It contains 56880 skeleton sequences with 60 action categories performed by 40 subjects captured from 3 distinct views. Generally, two benchmarks are commonly utilized for evaluation: cross-subject (Xsub) and cross-view (Xview). The Xsub splits all sequences based on subject indices, with 20 subjects allocated for training and the remaining 20 subjects for testing. The Xview divides data based on view variations, using views 2 and 3 for training while reserving view 1 for testing.

NTU RGB+D 120 [21]. As an extended version of the NTU RGB+D 60 dataset, it encompasses 114,480 skeleton sequences with 120 action categories. Similarly, it also provides the two official benchmarks: cross-subject (Xsub) and cross-setup (Xset). The Xsub task requires 53 subjects for training, with the remaining subjects designated for testing. Meanwhile, the Xset task utilizes data captured from even camera IDs for training and others for testing.

PKU-MMD [9]. Almost 20000 skeleton sequences with 51 categories are collected in this dataset, which can be split into two phases with increasing challenges. It offers two benchmarks: cross-subject (Xsub) and cross-view (Xview). For the Xsub task, 57 subjects are designated for training, while 9 subjects are reserved for testing. For Xview, the middle and right views data is included in the training set, with the left view used for testing. In this work, we utilize the first phase for experiments following the [6, 39].

4.2. Evaluation Settings

Protocols. We follow the previous studies [6] to conduct the experiments using predefined seen/unseen protocols and tasks in the ZSL and GZSL settings. For NTU 60, we adopt the 55/5 and 48/12 protocols. For NTU 120, they are extended to 110/10 and 96/24 protocols. For PKU-MMD I, we take the 46/5 and 39/12 split strategies.

Metrics. We evaluate the performance of our method using Top-1 classification accuracy $\mathcal{A}cc = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[y_i \in \hat{y}^*]$. In the ZSL setting, we calculate accuracy ($\mathcal{A}cc$) on the $\mathcal{D}_{te}^u$. For the GZSL setting, we first compute the seen accuracy ($\mathcal{S}$) on the $\mathcal{D}_{te}^s$ and unseen accuracy ($\mathcal{U}$) on the $\mathcal{D}_{te}^u$, respectively. Then, the harmonic mean accuracy ($\mathcal{H}$) can be calculated by $\mathcal{H} = (2 \times \mathcal{S} \times \mathcal{U})/(\mathcal{S} + \mathcal{U})$.

4.3. Implementation Details

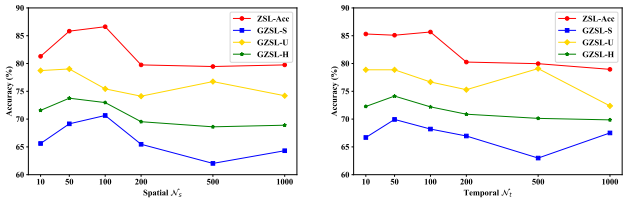
We sample the skeleton sequences to 64 frames using the data processing procedure adopted in [6]. For the skeleton encoder, we employ the Shift-GCN [8], which is pre-trained on the seen categories. Concurrently, we take the text encoder from the pre-trained **ViT-L/14@336px** model

Table 2. State-of-the-art comparisons on NTU RGB+D 60 dataset under the ZSL and GZSL setting. The best and the second-best results are marked in **Red** and **Blue**, respectively.

Method	Xsub								Xview							
	55/5 Split				48/12 Split				55/5 Split				48/12 Split			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$
ReViSE [17]	69.5	40.8	50.2	45.0	24.0	21.8	14.8	17.6	54.4	25.8	29.3	27.4	17.2	34.2	16.4	22.1
JPoSE [35]	73.7	66.5	53.5	59.3	27.5	28.6	18.7	22.6	72.0	61.1	59.5	60.3	28.9	29.0	14.7	19.5
CADA-VAE [30]	76.9	56.1	56.0	56.0	32.1	50.4	25.0	33.4	75.1	65.7	56.1	60.5	32.9	49.7	25.9	34.0
SynSE [13]	71.9	51.3	47.4	49.2	31.3	44.1	22.9	30.1	68.0	65.5	45.6	53.8	29.9	61.3	24.6	35.1
SMIE [39]	77.9	-	-	-	41.5	-	-	-	79.0	-	-	-	41.0	-	-	-
PURLS [41]	79.2	-	-	-	41.0	-	-	-	-	-	-	-	-	-	-	-
SA-DAVE [32]	**82.4**	62.8	**70.8**	66.3	41.4	50.2	36.9	42.6	-	-	-	-	-	-	-	-
STAR [6]	81.4	**69.0**	69.9	**69.4**	**45.1**	**62.7**	**37.0**	**46.6**	**81.6**	**71.9**	**70.3**	**71.1**	**42.5**	**66.2**	**37.5**	**47.9**
Neuron (Ours)	**86.9**	**69.1**	**73.8**	**71.4**	**62.7**	**61.6**	**56.8**	**59.1**	**87.8**	**70.6**	**75.9**	**73.2**	**63.3**	**65.3**	**58.1**	**61.5**

Table 3. State-of-the-art comparisons on NTU RGB+D 120 dataset under the ZSL and GZSL setting. The best and the second-best results are marked in **Red** and **Blue**, respectively.

Method	Xusb								Xset							
	110/10 Split				96/24 Split				110/10 Split				96/24 Split			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$
ReViSE [17]	19.8	0.6	14.5	1.1	8.5	3.4	1.5	2.1	30.2	4.0	23.7	6.8	13.5	2.6	3.4	2.9
JPoSE [35]	57.3	53.6	11.6	19.1	38.1	41.0	3.8	6.9	52.8	23.6	4.4	7.4	38.5	79.3	2.6	4.9
CADA-VAE [30]	52.5	50.2	43.9	46.8	38.7	48.3	27.5	35.1	52.5	46.0	44.5	45.2	38.7	47.6	26.8	34.3
SynSE [13]	52.4	57.3	43.2	49.5	41.9	48.1	32.9	39.1	59.3	58.9	49.2	53.6	41.4	46.8	31.8	37.9
SMIE [39]	61.3	-	-	-	42.3	-	-	-	57.0	-	-	-	42.3	-	-	-
PURLS [41]	**72.0**	-	-	-	52.0	-	-	-	-	-	-	-	-	-	-	-
SA-DAVE [32]	68.8	**61.1**	**59.8**	**60.4**	**46.1**	**58.8**	35.8	**44.5**	-	-	-	-	-	-	-	-
STAR [6]	63.3	59.9	52.7	56.1	44.3	51.2	**36.9**	42.9	**65.3**	**59.3**	**59.5**	**59.4**	**44.1**	**53.7**	**34.1**	**41.7**
Neuron (Ours)	**71.5**	**67.6**	**59.5**	**63.3**	**57.1**	**67.5**	**44.4**	**53.6**	**71.1**	**67.5**	**58.9**	**62.9**	**54.0**	**67.0**	**44.9**	**53.8**



(a) Spatial $\mathcal{N}_s$ (b) Temporal $\mathcal{N}_t$

Figure 3. The influence of hyper-parameters on the NTU 60.

of CLIP [27] to extract semantic embeddings. The optimizer is SGD with a weight decay of 0.0005. The batch size is 64. The base learning rate is 0.1 and reduced with 0.1 multiplied at epochs 10 and 20. The number of N_s and N_t are set to 80. The hyperparameter λ_s and λ_t are set to 0.0003 and 0.0002. All the experiments are finished using the PyTorch platform on a GeForce RTX 4090 Ti GPU.

4.4. Comparison with State-of-the-Art

Zero-Shot Learning. We evaluate the performance of our method, Neuron, against state-of-the-art approaches in the ZSL setting, as shown in Table 2 - 4. Across three distinct datasets representing varied scenarios, our proposed method achieves superior performance. Furthermore, the performance of our method is robust across all tasks, demonstrating its ability to learn core spatial-temporal patterns from skeletons that enhance cross-modal correspondences, thereby facilitating better cross-modal transferability.

Generalized Zero-Shot Learning. As shown in Table 2 - 4, our method outperforms previous studies in both seen and unseen categories. Additionally, our method achieves an excellent balance between the seen and unseen categories, *i.e.*, mitigating the tendency of the model towards the seen categories (domain bias), improving a considerable margin over others on the value of harmonic mean metrics. Thus, our synergistic framework actually helps to avoid shortcut representation learning, emphasizing spatial structure-related and temporal regularity-dependent feature learning.

4.5. Ablation Study

Influence of Seen-Unseen Category Settings. We evaluate the robustness of our method under the different seen-unseen categories settings. To this end, we re-partition all skeleton categories into three different seen-unseen combinations that are non-overlap, keeping the same as the previous studies [6, 39], and then compute the average results

Table 4. State-of-the-art comparisons on PKU-MMD I dataset under the ZSL and GZSL setting. The best and the second-best results are marked in **Red** and **Blue**, respectively.

Method	Xsub								Xview							
	46/5 Split				39/12 Split				46/5 Split				39/12 Split			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$	$\mathcal{A}cc$	$\mathcal{S}$	$\mathcal{U}$	$\mathcal{H}$
ReViSE [17]	54.2	44.9	34.5	39.1	19.3	35.7	13.0	19.0	54.1	50.7	39.9	44.6	12.7	34.5	9.43	14.8
JPoSE [35]	57.4	67.0	43.0	52.4	27.0	64.8	26.5	37.6	53.1	72.9	42.5	53.7	22.8	57.6	20.2	29.9
CADA-VAE [30]	73.9	76.2	51.8	61.7	33.7	69.0	29.3	41.1	74.5	**79.9**	61.5	69.5	29.5	62.4	28.3	39.0
SynSE [13]	69.5	**77.8**	40.2	53.0	36.5	71.9	30.0	42.3	71.7	69.9	51.1	59.0	25.4	61.9	22.6	33.1
SMIE [39]	72.9	-	-	-	44.2	-	-	-	71.6	-	-	-	40.7	-	-	-
STAR [6]	**76.3**	59.1	**72.3**	**65.0**	**50.2**	**72.7**	**44.7**	**55.4**	**75.4**	73.5	**72.2**	**72.8**	**50.5**	**69.8**	**47.5**	**56.5**
Neuron (Ours)	**89.2**	**78.9**	**75.9**	**77.4**	**61.4**	**78.3**	**56.8**	**65.9**	**88.2**	**78.7**	**77.8**	**78.3**	**62.2**	**75.4**	**57.1**	**65.0**

Table 5. Influence of seen-unseen categories settings on NTU RGB+D 60 and PKU-MMD I datasets. The columns of the ZSL and GZSL represent the Top1 accuracy and harmonic mean. The best and the second-best results are marked in **Red** and **Blue**.

Method	NTU RGB+D 60 55/5 (Xsub)		PKU-MMD I 46/5 (Xsub)	
	ZSL	GZSL	ZSL	GZSL
ReViSE [17]	54.7	27.4	48.7	32.8
JPoSE [35]	56.6	44.7	39.2	31.7
CADA-VAE [30]	58.0	47.1	49.0	52.7
SynSE [13]	59.9	49.9	43.5	40.4
SMIE [39]	64.2	-	66.4	-
STAR [6]	**77.5**	**62.8**	**70.6**	**67.1**
Our Method	**84.5**	**71.2**	**74.4**	**69.2**

Table 6. Analysis of different components on NTU RGB+D 60 Xsub dataset under the 55/5 split strategies.

Spatial Stream		Temporal Stream		$\mathcal{A}cc$ (%)	
Granularity Update	Spatial Compression	Phase Evolving	Temporal Memory	ZSL	GZSL
✗	✗	✗	✗	79.9	65.9
✓	✗	✗	✗	81.0	66.3
✓	✗	✓	✗	81.2	67.1
✓	✓	✓	✗	81.7	69.9
✓	✓	✓	✓	86.9	71.4

of three settings to minimize variance. We report the ZSL and GZSL results in Table 5. It can be seen that our method consistently achieves state-of-the-art results in different settings, showing excellent stability and robustness.

Influence of Components. We evaluate the efficiency of different components of our method in Table 6. It is noted that the performance of Neuron drops significantly without the temporal memory mechanism, around 5.2%/1.5%, showing this technique can mitigate the situation of knowledge oblivion effectively. Additionally, the dual guidance on spatial granularity updating and temporal phase evolving is crucial in progressively controlling the process of cross-modal alignment. Meanwhile, the spatial compression mechanism can improve generalization, increasing the GZSL from 67.1% to 69.9%.

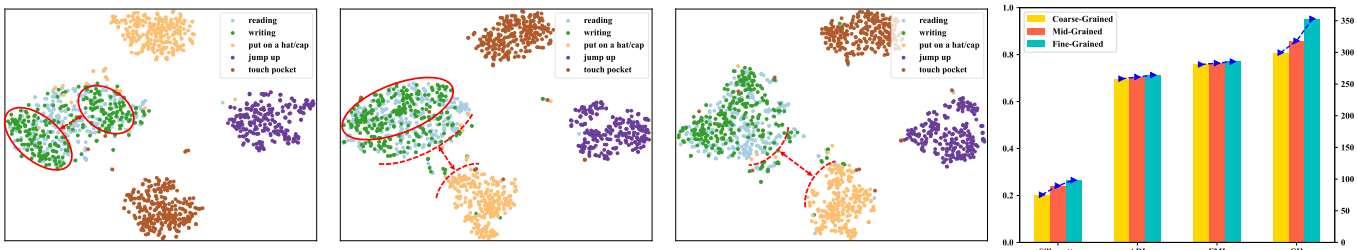
Influence of $\mathcal{N}_s$ and $\mathcal{N}_t$ in Micro-prototypes. As illustrated in Fig. 3 (a) - 3 (b), we observe that the accuracy initially improves and then decreases with the attribute number increases. The best performance of our method can be obtained when the $\mathcal{N}_s$ is around 100 and the $\mathcal{N}_t$ is approximately 50. With further increases in attribute values, accuracy stabilizes and remains consistently high, highlighting the effectiveness of our progressive learning approach.

4.6. Qualitative Analysis

Visualization of Spatial Compression. Here, we plot the t-SNE visualization of unseen skeleton representation distribution at each phase during the spatial compression process, as shown in Fig. 4 (a) - 4 (c). Moving from the coarse-grained to the mid-grained phase, we can see that the compactness of features in "writing" becomes better. Furthermore, the inter-class separability between the "reading" and "put on a hat/cup" actions can also be improved in transitioning from the mid-grained to fine-grained steps. To quantify intra-class compactness, we calculate the compactness of features based on mainstream clustering evaluation metrics, including the silhouette score [28], Adjusted Rand Index (ARI) [16], Fowlkes-Mallows Index (FMI) [12], and Calinski-Harabasz Index (CH) [3], as shown in Fig. 4 (d). The intra-class compactness continually improves, underscoring Neuron's effective core feature learning.

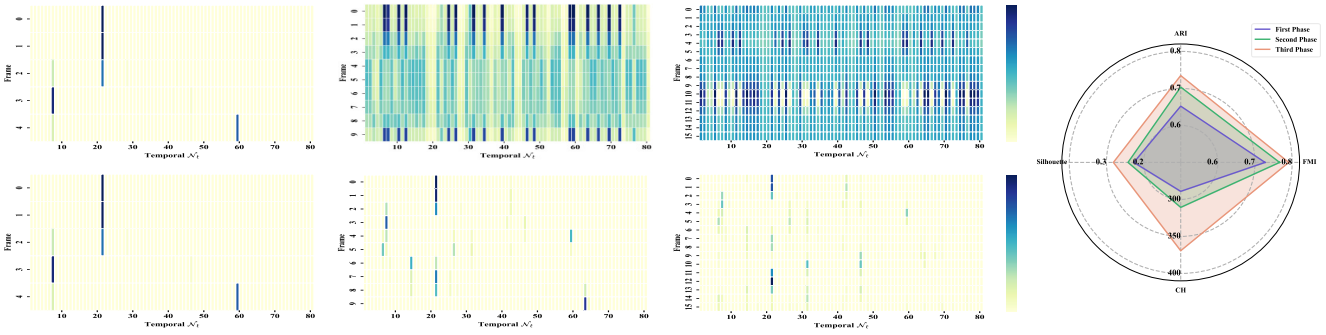
Visualization of Temporal Memory. As shown in Fig. 5 (a) - 5 (c), we draw the updating process of temporal micro-prototypes over multiple steps, comparing results with and without the temporal memory mechanism. In the first row, the updated temporal micro-prototype without the memory mechanism guidance will be chaotic as frames increase, leading to knowledge oblivion. In contrast, the second row exhibits a more effective learning process, where past knowledge is retained while new information is integrated. We also compute the intra-class compactness with the memory mechanism based on the above metrics [3, 12, 16, 28], all of them achieve better with the learning process ongoing.

Fine-Grained Action Recognition. To compare the perfor-



(a) Coarse-Grained Phase (b) Mid-Grained Phase (c) Fine-Grained Phase (d) Intra-Class Compactness

Figure 4. (a) - (c) represent the t-SNE visualization of the skeleton spatial spaces for unseen categories on different phases. The color denotes different unseen categories from the cross-subject task of the NTU 60 dataset under the 55/5 split settings. (d) represents the spatial intra-class compactness metrics of corresponding phases (zoom in for a better view).



(a) First Phase (b) Second Phase (c) Third Phase (d) Intra-Class Compactness

Figure 5. (a) - (c) represent the visualization of updated temporal micro-prototype in each phase for a randomly selected skeleton sequence. The first row denotes the updated process without the proposed temporal memory mechanism, while the second row is equipped. (d) represents the temporal intra-class compactness metrics of corresponding phases (zoom in for a better view).

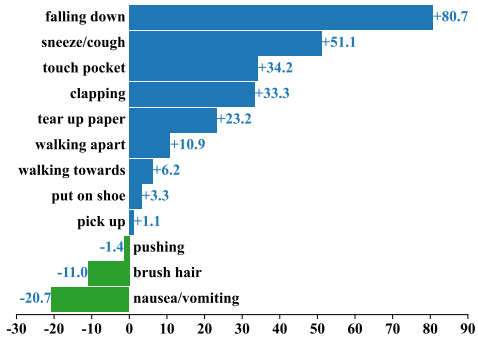


Figure 6. Unseen action classification accuracy compared with STAR [6] on the Xsub task of NTU 60 under the 48/12 split setting.

mance of our method with the fine-grained method STAR [6] more intuitive, we calculate the absolute difference value of recognition accuracy between Neuron and STAR in each unseen category. As shown in Fig. 6, we find the accuracy of 9 action categories recognized by our method over the STAR a lot. Specifically, the performance improved for "walking apart" and "walking towards" actions as we paid isolated attention to temporal cues during the training process. Meanwhile, some spatial-oriented actions, *e.g.*, "clapping" and "falling down" actions, are improving significantly. However, some failure cases, such as the "nausea/vomiting" action, are incorrectly classified into the "sneeze/cough" action with highly similar skeletons. Meanwhile, the homogeneous situation still exists between their contextual side information. So, further exploration is still needed for those extremely similar and abstract actions.

5. Conclusion

In this paper, we propose **Neuron**, a novel dy**N**amically **E**volving d**U**al skeleton-semantic syne**R**gistic framework guided by c**O**ntext-aware side informatio**N**, to explore the desirable fine-grained cross-modal correspondence. By introducing the spatial-temporal evolving micro-prototypes to imitate the learning and growth of neurons, the synergistic skeleton-semantic correlations can be established and refined step-by-step under the guidance of contextual side information. In addition, the spatial compression and temporal memory mechanisms are appropriately designed to mitigate the shortcut feature learning and the situation of temporal knowledge oblivion, absorbing the structure-related and regularity-dependent core spatial-temporal representations. Experimental results indicate excellent cross-modal transferability and effective generalizability from seen to unseen categories.

Acknowledgments

This research was supported by funding from the Hong Kong RGC General Research Fund (No. 15221123 and 15216424) and the PolyU Internal Fund (No. P0043932).

References

[1] Maurits Bleeker. Multi-modal learning algorithms and network architectures for information extraction and retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6925–6929, 2022. 3

[2] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022. 3

[3] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 7

[4] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 52–68. Springer, 2016. 5

[5] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13359–13368, 2021. 1

[6] Yang Chen, Jingcai Guo, Tian He, Xiaocheng Lu, and Ling Wang. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 778–786, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[7] Yang Chen, Tian He, Junfeng Fu, Ling Wang, Jingcai Guo, Ting Hu, and Hong Cheng. Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning. *IEEE Transactions on Multimedia*, 2024. 1, 3

[8] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 183–192, 2020. 1, 5

[9] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 5

[10] Yufeng Cui and Yimei Kang. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17957, 2023. 3

[11] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583, 2021. 3

[12] Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569, 1983. 7

[13] Pranay Gupta, Divyanshu Sharma, and Ravi Kiran Sarvadevabhatla. Syntactically guided generative embeddings for zero-shot skeleton action recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 439–443. IEEE, 2021. 1, 2, 3, 6, 7

[14] Tian He, Yang Chen, Xu Gao, Ling Wang, Ting Hu, and Hong Cheng. Enhancing skeleton-based action recognition with language descriptions from pre-trained large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[15] Tian He, Yang Chen, Ling Wang, and Hong Cheng. An expert-knowledge-based graph convolutional network for skeleton-based physical rehabilitation exercises assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024. 1

[16] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985. 7

[17] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International conference on Computer Vision*, pages 3571–3580, 2017. 6, 7

[18] Bhavan Jasani and Afshaan Mazagonwalla. Skeleton based zero shot action recognition in joint pose-language semantic space. *arXiv preprint arXiv:1911.11344*, 2019. 1, 2, 3

[19] Sangwon Kim, Dasom Ahn, and Byoung Chul Ko. Cross-modal learning with 3d deformable attention for action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10265–10275, 2023. 3

[20] Ming-Zhe Li, Zhen Jia, Zhang Zhang, Zhanyu Ma, and Liang Wang. Multi-semantic fusion model for generalized zero-shot skeleton-based action recognition. In *International Conference on Image and Graphics*, pages 68–80. Springer, 2023. 1, 2

[21] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 5

[22] Jinfu Liu, Chen Chen, and Mengyuan Liu. Multi-modality co-learning for efficient skeleton-based action recognition. *arXiv preprint arXiv:2407.15706*, 2024. 3

[23] Chenyu Lu, Jun Yin, Hao Yang, and Shiliang Sun. Enhancing multi-modal fusion in visual dialog via sample debiasing and feature interaction. *Information Fusion*, 107:102302, 2024. 3

[24] Mingqi Lu, Siyuan Yang, Xiaobo Lu, and Jun Liu. Cross-modal contrastive pre-training for few-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[25] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. Cmd: Self-supervised 3d action representation learning with cross-modal mutual distillation. In *European*

Conference on Computer Vision, pages 734–752. Springer, 2022. 3

[26] Jianyuan Ni, Anne HH Ngu, and Yan Yan. Progressive cross-modal knowledge distillation for human action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5903–5912, 2022. 3

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[28] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 7

[29] Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6471–6480, 2023. 1

[30] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–57, 2019. 6, 7

[31] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 5

[32] Wei-Jie Chen Yi-Hsin Yu Chih-Yuan Yang Jane Yung-jen Hsu Sheng-Wei Li, Zi-Xiang Wei. Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 6

[33] Sijie Song, Jiaying Liu, Yanghao Li, and Zongming Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020. 3

[34] Shengkai Sun, Daizong Liu, Jianfeng Dong, Xiaoye Qu, Junyu Gao, Xun Yang, Xun Wang, and Meng Wang. Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2973–2984, 2023. 3

[35] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 450–459, 2019. 6, 7

[36] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10276–10285, 2023. 3

[37] Jinglin Xu, Guohao Zhao, Sibo Yin, Wenhao Zhou, and Yuxin Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21773–21782, 2024. 1

[38] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10608–10617, 2023. 1, 4

[39] Yujie Zhou, Wenwen Qiang, Anyi Rao, Ning Lin, Bing Su, and Jiaqi Wang. Zero-shot skeleton-based action recognition via mutual information estimation and maximization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5302–5310, 2023. 1, 2, 3, 5, 6, 7

[40] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024. 1

[41] Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. Part-aware unified representation of language and skeleton for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18761–18770, 2024. 1, 2, 3, 6