# Semantic-guided Cross-Modal Prompt Learning for Skeleton-based Zero-shot Action Recognition

Anqi Zhu[1], Jingmin Zhu[2], James Bailey[1], Mingming Gong[1], Qiuhong Ke[2,*]
[1]The University of Melbourne; [2]Monash University
Parkville VIC 3052, Australia; Wellington Road, Clayton VIC 3800, Australia
azzh1@student.unimelb.edu.au, jzhu0105@student.monash.edu,
{baileyj, mingming.gong}@unimelb.edu.au, qiuhong.ke@monash.edu

## Abstract

*Skeleton-based human action recognition is promising due to its privacy preservation, robustness to visual challenges, and computational efficiency. Especially, the practical necessity to recognize unseen actions has led to increased interest in zero-shot skeleton-based action recognition (ZSSAR). Existing ZSSAR approaches often rely on manually crafted action descriptions or visual assumptions to enhance knowledge transfer, which is limited in flexibility and prone to inaccuracies and noise. To overcome this, we introduce Semantic-guided Cross-Modal Prompt Learning (SCoPLe), a novel framework that replaces manual guidance with data-driven prompt learning for refinement and alignment of skeletal and textual features. Specifically, we introduce a dual-stream language prompting module that preserves the original semantic context from the pre-trained text encoder while still effectively tuning its ouput for ZSSAR task adaptation. We also introduce a joint-shaped prompting module that learns tuning for skeleton features and incorporate an adaptive visual representation sampler that leverages text semantics to strengthen the cross-modal prompting interactions during skeleton-to-text embedding projection. Experimental results on the NTU-RGB+D and PKU-MMD datasets demonstrate the state-of-the-art performance of our method in both ZSSAR and generalized ZSSAR scenarios.*

## 1. Introduction

Human Action Recognition (HAR) is essential in various fields, including surveillance [7, 17, 32], healthcare [1, 46], human-computer interaction [2, 40], and smart environments [4, 24, 41, 45]. Within HAR, skeleton-based action recognition has garnered significant attention due to its unique capability to capture class-determinant body joint movements while ensuring privacy [13, 27, 38]. Moreover, skeleton data is robust to common visual challenges such as background clutter and lighting variations [6], and is computationally efficient, making it suitable for resource-constrained environments [28, 31, 47]. These advantages have led to increased research efforts on the topic in recent years [5, 21, 26, 44].

*Corresponding author

| Model | NTU-RGBD 60 | | NTU-RGBD 120 | |
|---|---|---|---|---|
| | 55/5 split | 48/12 split | 110/10 split | 96/24 split |
| CLIP | 62.58 | 33.16 | 63.92 | 48.15 |
| CLIP-CoOp [49] | 75.88 (+13.3%) | 36.32 (+3.16%) | 64.8 (+0.88%) | 46.02 (-2.13%) |
| CLIP-MaPLe [18] | 65.59 (+3.01%) | 34.93 (+1.77%) | 66.45 (+2.47%) | 47.42 (-0.73%) |
| CLIP-DSLP (ours) | 70.21 (+7.63%) | 40.69 (+7.53%) | 68.88 (+4.96%) | 48.17 (+0.02%) |

Table 1. ZSSAR accuracy for adapting CLIP with different common prompting techniques or our DSLP (see Sec. 3.2) on *NTU-RGB+D 60* and *NTU-RGB+D 120*. For previous approaches, we only preserve their prompting frameworks for the text branch because the shape of the skeleton features is not compatible with their original designs for visual prompting which are applied for image encoding.

Most of the present skeleton-based action recognition approaches [8, 10, 35–37, 43] focus on fully supervised learning, where both training and testing involve the same fixed set of action classes. However, in practical scenarios, the vast range of human actions, combined with the high cost and time required to collect and annotate data for each possible category, demands models that can be effectively adaptable to new recognition tasks involving unseen classes. **Z**ero-**s**hot **s**keleton-based **a**ction **r**ecognition (ZSSAR) is designed to tackle this challenge – enabling models to recognize novel actions without requiring explicit training examples for them.

In addressing ZSSAR, the major challenge is achieving effective knowledge transfer, which requires constructing a shared semantic space to bridge a generalized visual representation distribution for seen and unseen actions. Prior research in ZSSAR [9, 14, 42, 51] has solved this by leveraging label embeddings from pre-trained language models (e.g., BERT) as generalizable intermediate references for distribution learning and developing various strategies to expand the scope of transferable knowledge. While effective, they often rely on careful assumptions and manually crafted techniques to guide the training. For example, some latest solutions involve semantic decomposition that makes simplified assumptions about action labels. [14] assumes that every action label follows a "verb+noun" syntactic structure. [51] and [9] require the manual generation of local movement descriptions for each label by GPT [3] for body-part-based learning. The success of these approaches is highly sensitive to the quality and consistency of the assumptions and descriptions, which might introduce noise or irrelevant information that does not directly aid in skeleton-text alignment.

In this work, we shift away from manually crafted guidance and delve into ZSSAR solutions through purely data-driven prompt tuning for cross-modal knowledge refinement and alignment. In particular, our aim is to investigate both text and skeleton side prompt learning to automatically extract essential contextual signals from action labels and transmit them across the modalities, thereby reducing reliance on artificial assumptions and operations. By automatically capturing task-refined cross-modal semantics from the labels, the model can focus on tuned consistent contexts shared across similar actions, allowing it to establish generalizable visual patterns that define each potential action. This ultimately enhances knowledge transfer, enabling the model to recognize new actions without support from specific analyses or detailed descriptions.

Antecedent works on prompt learning laid the methodology foundation for enhancing pre-trained multi-modal models (PMMs) such as CLIP [30] by concatenating learnable prompt embeddings with class text tokens or image patch embeddings before feeding them into the encoders [49], or by inserting learnable prompts at each encoder layer to fine-tune the attention distribution [18]. Ensuring learning interaction and synchronization between prompts in different data-mode encoders is another important technique to develop tuning efficiency [18, 48]. Showing success in image-based supervised task adaptation, these methods also remain valid in some zero-shot learning (ZSL) scenarios. Yet, according to the 11 dataset experiments in [18], the improvement is unstable and sometimes still fail to surpass the baseline of only using CLIP. In Tab. 1, we conducted a validation experiment by applying common prompting methods (*e.g.* CoOp [49] and MaPLe [18]) to CLIP for ZSSAR, and the results yield similar volatile performance gains. We argue that this instability may arise from the impact of learnable prompts on the generalizability of the original CLIP model to unseen actions. Specifically, a CLIP text encoder, built on a Transformer architecture with a self-attention mechanism, is pre-trained on extensive visual-text pairs to support broad generalization. However, prompts trained exclusively on seen classes can alter the attention outputs of text labels in ways that can significantly impact the model's inherent ability to generalize to the semantic context of new classes. In such cases, the model is overly reliant on these modified attention distributions and feature adjustments introduced by the learnable prompts, which can sometimes enhance performance, but may also compromise the fundamental strengths of CLIP. Additionally, these methods' original prompting strategy for the image branch does not directly accommodate skeleton features due to differences in data shape. The absence of prompting on the visual side will further exacerbate the tuning instability.

To address this, we argue that both the attention output of the original and prompted encodings should be considered for optimal generalization. We propose a dual-stream language prompting (DSLP) module. It consists of a pre-trained CLIP text encoder that retains the original encoding stream, along with a second stream that learns layer-level prompts, which tunes the prompted feedforward features while adaptively integrating the original attention distribution from the first stream. This enables the model to balance between the preservation of the original context structure and the tuning of the features. Furthermore, to facilitate the corresponding tuning for skeleton feature refinement, we propose a semantic-guided skeleton prompting module that uses joint-shaped prompts to optimize task-relevant contexts. The design is grounded in the unique data structure of skeleton joint features. The module then incorporates an adaptive visual representation sampler that synchronizes the prompted text semantics from the language branch as queries to gather class-relevant prompted visual clues for projecting skeleton-to-text embeddings. To this end, we establish all the core components of our proposed **S**emantic-guided **Cro**ss-Modal **P**rompt **Le**arning (SCoPLe). In summary, the main contributions of this work include:

- We propose a dual-stream language prompting module that enhances the semantic extraction capabilities of the pre-trained CLIP text encoder, particularly improving its generalization to refine textual features for novel classes.
- We introduce a semantic-guided skeleton prompting module. It learns joint-shaped prompts to refine skeleton features, utilizing an adaptive visual representation sampler to implement cross-modal prompted feature alignment.
- We evaluate SCoPLe in both ZSSAR and Generalized ZSSAR scenarios, demonstrating its ability to achieve state-of-the-art performance on the unified baselines of NTU-RGB+D 60, NTU-RGB+D 120, and PKU-MMD, without requiring extra manual semantic knowledge definition/enrichment or local movement segmentation.

## 2. Related Works

### 2.1. Skeleton-based Zero-shot Action Recognition

Zero-shot skeleton-based action recognition (ZSSAR) addresses the challenge of recognizing skeleton-based actions without access to training samples for target classes. This paradigm represents a specialized knowledge transfer problem in which the source and target domains maintain non-overlapping label distributions while sharing the same input domain of uniformly shaped skeleton sequences. Generalized ZSSAR extends this challenge by incorporating discrimination from the seen class during testing.

Early methodologies primarily focused on two types of strategies: visual-to-semantic projection and shared encoding space construction. In 2019, [42] initially proposed fundamental learning frameworks based on these two theories. They demonstrated that one can either utilize common-space metric learning under a Relation-Net architecture, which transforms both skeleton and text format features, or apply visual-to-semantic embedding projection using a DeViSE [12] model, followed by similarity-based semantic matching. ReViSE [39] aligned latent space outputs from individual data modality autoencoders by minimizing maximum mean discrepancy loss. Moreover, CADA-VAE [33] and SynSE-ZSL [14] employed variational autoencoders for generative latent space with cross/intra-modal reconstruction alignment, introducing part-of-speech decomposition to enhance semantic overlap and improve knowledge transfer. PURLS [51] and STAR [9] regard action recognition as the discrimination of combinations of overlapping sub-motions appearing in both seen and unseen classes. By utilizing GPTs to extract sub-motion semantics, they established a multi-scale cross-modal representation matching to enrich seen-to-unseen knowledge transfer. During semantic alignment, PURLS uses fixed preamble prompts for each sub-motion description, while STAR employs naive prompting concatenation before applying feature projection for each local part, where the main prediction improvements for both solutions arise from the introduction of labor-intensive local pattern learning. SMIE [50] focuses on maximizing mutual information for distributional semantic alignment while incorporating temporal rank loss for more comprehensive feature matching. SA-DVAE [22] tackles information imbalance by disentangling skeleton features into semantically related and irrelevant components.

While preceding baselines incorporate sophisticated artificial designs to boost knowledge transfer in skeleton sequence features, our method fully harnesses the pre-trained CLIP knowledge through purely data-driven task adaptation, eliminating the need for additional manual transferrable knowledge definition, enrichment, or local movement segmentation.

### 2.2. Prompt Learning

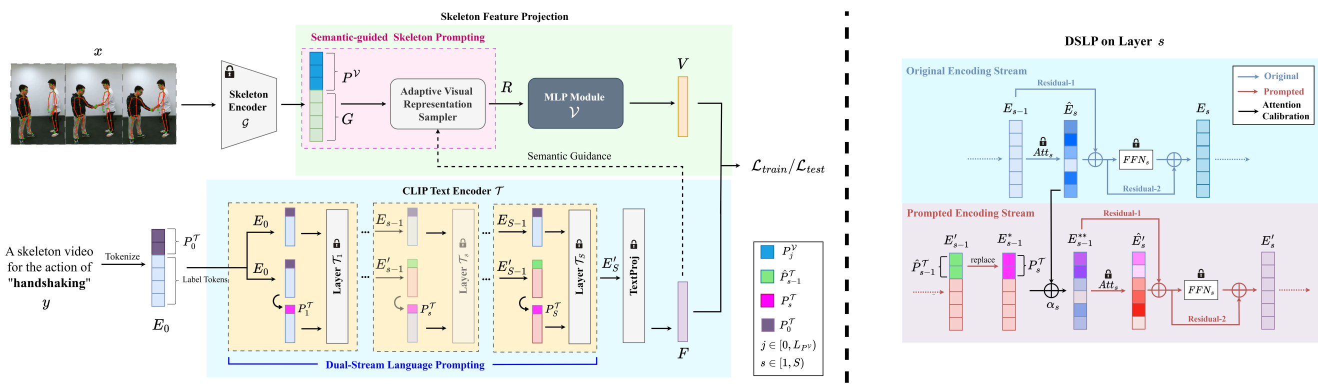
Before the proliferation of prompting techniques, fine-tuning was pivotal for adapting pre-trained large models (PLMs) to specific tasks. Although it efficiently leveraged PLM knowledge, it required labor-intensive supervised training to adjust the model's backbone weights for targeting task-beneficial features. Depending on the domain discrepancies between pre-training and task data, fine-tuning can vary its learning object as either the final layers or the entire model. However, the process eventually becomes cumbersome and costly due to the exponential increase in model size, with parameters ranging from millions to trillions [11, 20, 29].

To mitigate these challenges, researchers have explored lightweight plug-in modules that can freeze all pre-trained parameters, using small training tasks to find additional learnable embeddings concatenated to the standard inputs to optimize the output. Prompt learning, which originated from manual label prompting operations in CLIP [30], was first systematized by Lester et al. [19], who introduced learnable vectors to Transformer inputs. [49] refines this concept by enabling the training of independent prompt sub-embeddings that are attached to the front of the original label embeddings. However, static prompts lack dynamic and visual-level generalization across various classes [48], leading to the integration of visual cues via Meta-Net for generating conditional prompts based on instances. Concurrently, [23] emphasized that the inclusion of layer-level learnable prompts can further improve task adaptability in feature extraction during inference, and [18] correspondingly designed a prompting framework tailored for visual transformer branches. LoRA [16] advocates for inserting trainable low-rank matrices within the query projection of a Transformer attention module to facilitate nuanced parameter tuning for task adaptation. In 2022, [15] consolidated the theoretical derivation regarding adapters, prefix tuning (*i.e.* prompt learning) and low-rank matrix updates, pointing out that these solutions serve the same adaptation intuition but differ only in design choices. Based on their conclusive formula, they introduced Parallel Adapters for competitive accuracy through combined parallel learning and non-linear low-rank adjustments.

Previous approaches have demonstrated the effectiveness of prompt learning in fully supervised training environments by experimenting with various prompt insertion positions within model architectures and diverse integration methodologies. However, their research focuses on RGB-based and textual inputs, lacking in-depth exploration of adapting skeleton-based features, and present deterioration risks on base-to-novel generalization. Our approach addresses these issues by developing skeleton-feature-friendly prompting and introducing attention normalization to preserve the authentic label contexts for better new class generalization.

## 3. Methodology

**Problem Definition.** Let $\mathcal{D}_{tr} = \{(x^{sc}_{tr}, y^{sc}_{tr})\}$ represents the set of $N_{tr}$ training samples from the available seen classes $\mathcal{Y}^{sc}$. Each skeleton sequence $x^{sc}_{tr} \in \mathbb{R}^{L\times J\times M\times 3}$ records the 3-D locations of $J$ body joints for each actor across $L$ frames. Here, $M$ denotes the maximum number of actors per sequence, and $y^{sc}_{tr} \in \mathcal{Y}^{sc}$ is the corresponding action




(a) (b)

Figure 1. (a) Overview of our proposed SCoPLe to solve (G)ZSSAR within a cross-modal semantic alignment framework. SCoPLe tunes both the skeleton and language branches by learnable prompts, while the original weights of the backbones remain frozen. On the linguistic side, SCoPLe enhances the text branch with a dual-stream language prompting (DSLP) module, effectively preserving the strengths of the original pre-trained CLIP knowledge to normalize textual prompt learning. On the visual side, SCoPLe introduces a semantic-guided skeleton prompting module, which includes joint-shaped prompts and an adaptive visual representation sampler to optimize the skeleton feature projection for synergistic semantic alignment with text-side tuning. (b) Illustration of the dual-stream language prompting module.

label from the seen class label set. Similarly, let $\mathcal{D}_{te} = \{(x_{te}^{uc}, y_{te}^{uc})\}$ denote the set of $N_{te}$ testing samples from the unseen classes $\mathcal{Y}^{uc}$. In a standard ZSSAR setting, we have $\mathcal{Y}^{sc} \cap \mathcal{Y}^{uc} = \phi$. By training solely on seen-class samples, we aim to develop a model capable of robustly projecting skeleton features from unseen classes, enabling accurate predictions $\hat{y} \in \mathcal{Y}^{uc}$ by mapping these features to their corresponding label embeddings. To extend the problem to GZSSAR, we require the model to predict $\hat{y} \in \mathcal{Y}^{sc} \cup \mathcal{Y}^{uc}$, where it should accurately identify both seen and unseen classes. This presents extra challenges for the model, as it must not only generalize to unseen classes but also distinguish them from seen classes.

### 3.1. Model Overview

Fig. 1 (a) presents the overall architecture of the proposed **S**emantic-guided **C**r**o**ss-Modal **P**rompt **Le**arning (SCoPLe) based on a cross-modal embedding alignment framework. Unlike previous approaches that rely on strong assumptions about label semantics or visual information structures to enhance learnable knowledge transfer, our method emphasizes automatic learning of textual and visual prompts to improve feature representation and cross-modal alignment. To achieve this, we propose two key components: 1) a dual-stream language prompting module and 2) a semantic-guided skeleton prompting module. Below we describe the details.

### 3.2. Dual-stream Language Prompting

We propose using learnable textual prompts on a pre-trained CLIP text encoder $\mathcal{T}$ to assist the extraction of key context knowledge signals from action labels, reducing the dependence on using manual descriptions for refining the linguistic semantic targets used for skeleton-side projection alignment. This allows the model to automatically adapt for learning consistent and task-specific features across actions, improving knowledge transfer and facilitating the recognition of new actions. However, due to the learning mechanism in the CLIP text encoder (based on a Transformer architecture), common prompt learning methods that concatenate learnable prompts with label text tokens [18] for integrated tuning can disrupt the entire self-attention distributions that interpret the semantic contexts of the original labels, consequently limiting and degrading the innate generalizability of CLIP, if the prompts are only trained with seen class samples.

To address this, for each layer in $\mathcal{T}$, we attach a dual-stream language prompting module. As shown in Fig. 1 (b), the module processes the layer input in two streams: a regular encoding stream to extract original label token features, which preserves the authentic semantics of the input text label $y$, and a prompted encoding stream that enhances contextual information shared across actions for task adaptation. For the second stream, before conducting prompt-inclusive attention and feedforward calculations, we first decorate its input with a layer-level prompt and apply a learnable weighted aggregation to combine the attention output from the first stream. This enables the model to selectively preserve the original semantic context structure while effectively incorporating the prompted tuning to refine the token feature content. Formally, before sending $y$ to $\mathcal{T}$, we first follow the operation from [49] and concatenate a preamble prompt initialized as the phrase 'A skeleton video for the action of' before the label text. Since each word will have its own token embedding, this allows us to replace some parts of the preamble phrase with learnable prompts without modifying the label texts. We swap the first two tokens in the preamble prompt with a learnable vector $P_0^{\mathcal{T}}$ of the same length to incorporate basic input channel tuning, and these positions are also used as the swapping slots for later layer-wise prompts. The input for $\mathcal{T}$ is now defined as $E_0$. Then, as shown in Fig. 1 (b), for any arbitrary layer $\mathcal{T}_s$ where $s \in [1, S]$, $S \leq Z$ and $Z$ denotes the number of layers in $\mathcal{T}$, it is primitively composed of a multihead attention layer $\mathrm{Att}_s$ and a feedforward layer $\mathrm{FFN}_s$ with pre-trained parameter weights. In our learning process, it handles two streams of input: $E_{s-1}$ and $E'_{s-1}$. $E_{s-1}$ is the output of the previous layer $\mathcal{T}_{s-1}$ that encodes $E_0$ without any additional layer-level tuning modification, while $E'_{s-1}$ is the second stream output from $\mathcal{T}_{s-1}$ that incorporates encoding tuning by the previous layer-wise prompt. For the original encoding stream, the output $E_s$ of $\mathcal{T}_s$ is calculated from $E_{s-1}$ as

$$\hat{E}_s = \mathrm{Att}_s(\mathrm{LayerNorm}(E_{s-1})), \tag{1}$$

$$\hat{E}_s^{res} = E_{s-1} + \hat{E}_s, \tag{2}$$

$$\begin{aligned} E_s &= \mathcal{T}_s(E_{s-1}) \\ &= \hat{E}_s^{res} + \mathrm{FFN}_s(\mathrm{LayerNorm}(\hat{E}_s^{res})), \end{aligned} \tag{3}$$

where $\hat{E}_s$ is the output of $\mathrm{Att}_s$ and $\hat{E}_s^{res}$ is the result of adding residual connections. For the prompted encoding stream, $E'_{s-1}$ first replaces its decorated positions prompted by the previous layer with the current layer prompt $P_s^{\mathcal{T}}$. We denote this output by $E^*_{s-1}$. Then, it is enhanced with the attention output of $\hat{E}_s$ to emphasize the information of the original label semantics. This is achieved through a weighted aggregation controlled by a learning layer parameter $\alpha_s$ that balances between the tuning effect and the original encoding for a best generalized adaptation. The resulting $E^{**}_{s-1}$ then goes through the ordinary attention learning and feedforward network in $\mathcal{T}_s$ using the pre-trained parameters and produces $E'_s$ as the prompted stream input for the next layer. Note that the location decorated with layer prompts is now encoded as $\hat{P}_s^{\mathcal{T}}$ rather than the native value of $P_s^{\mathcal{T}}$. The actual procedures are implemented as

$$E^*_{s-1} = [P_s^{\mathcal{T}}, E'_{s-1} \setminus \hat{P}_{s-1}^{\mathcal{T}}], \tag{4}$$

$$E^{**}_{s-1} = (1 - \alpha_s) E^*_{s-1} + \alpha_s \hat{E}_s, \tag{5}$$

$$\hat{E}'_s = \mathrm{Att}_s(\mathrm{LayerNorm}(E^{**}_{s-1})), \tag{6}$$

$$\hat{E'}_s^{res} = E^{**}_{s-1} + \hat{E}'_s, \tag{7}$$

$$\begin{aligned} E'_s &= \mathcal{T}_s(E'_{s-1}) \\ &= \hat{E'}_s^{res} + \mathrm{FFN}_s(\mathrm{LayerNorm}(\hat{E'}_s^{res})), \end{aligned} \tag{8}$$

where $[\cdot, \cdot]$ refers to the concatenation operation and $\setminus$ refers to the removal operation. We implement layer-wise prompting on the first $S$ layers while keeping the remaining blocks in $\mathcal{T}$ unchanged as regular layers. This ensures that the subtle contextual tuning in each layer eventually accumulates and is comprehended in a non-linear manner at the end to influence the final output. The final output $F \in \mathbb{R}^{d_{text}}$ of $\mathcal{T}$ is computed from the subsequent encoding of the rest $Z - S$ layers according to $E'_S$ from the $S$th prompted layer as

$$(E_S, E'_S) = (\mathcal{T}_s(E_{S-1}), \mathcal{T}_s(E'_{S-1})), \tag{9}$$

$$F = \mathrm{TextProj}(E'_S). \tag{10}$$

$d_{text}$ is the final text feature dimension and $\mathrm{TextProj}(\cdot)$ is the combination of the remaining $Z - S$ layers. $S$ is a pre-definable hyperparameter.

### 3.3. Semantic-guided Skeleton Prompting

While each skeleton joint feature captures spatial-temporal visual information, they lack direct semantic context unless learned in conjunction with the entire human body. The primary objective of visual prompting is to strengthen semantic tuning interactions between the skeleton joint features and the corresponding label embeddings. To achieve this, we propose a semantic-guided skeleton prompting module that learns joint-shaped prompts for the visual side based on the prompted textual output. This refines the skeleton features to enable a more accurate representation and understanding of the visual data in relation to textual cues.

**Joint-shaped Prompts**. Previous work on learnable prompts [18, 48, 49] has established a basic paradigm for effectively integrating prompts by integrating them into the original data structure of the sample during the feature feedforward process. Therefore, we propose a joint-shaped prompt that matches the unique encoding format of skeleton features. Given a 3D skeleton sequence input $x \in \mathbb{R}^{L \times J \times M \times 3}$, we first use a universal pre-trained Shift-GCN [10] backbone to extract its joint visual features $G = \mathcal{G}(x) \in \mathbb{R}^{J \times d_{skel}}$, where $d_{skel}$ is the skeleton feature dimension. To simplify the calculations, we average the features across $M$ performers and the temporal dimensions. The visual prompt is then established in the format of $P^{\mathcal{V}} \in \mathbb{R}^{L_{P^{\mathcal{V}}} \times d_{skel}}$, where $L_{P^{\mathcal{V}}}$ is the prompt length and the feature shape aligns with the real joint feature. The prompt is initialized with random values $z \sim N(0, 1)$. The design intuition is to simulate $L_{P^{\mathcal{V}}}$ fake joint features that introduce contextual information facilitating text semantic alignment.

**Adaptive Visual Representation Sampler**. To extract textual-relevant alignment information from the skeleton features, we propose an Adaptive Visual Representation Sampler that leverages the textual branch output to perform cross-attention sampling on the original joint features and visual prompts, with weights determined by their semantic relevance to the targeted labels (see Fig. 2). This obtains a semantic-relevant visual feature for better skeleton-text alignment. Given the skeleton feature output $G$, the joint-level prompt $P^{\mathcal{V}}$ and the textual embedding $F$, we establish three attention matrices $\mathbf{W}_Q \in \mathbb{R}^{d_{text} \times h}$, $\mathbf{W}_K \in \mathbb{R}^{d_{skel} \times h}$, and $\mathbf{W}_V \in \mathbb{R}^{d_{skel} \times d_R}$. $h$ refers to the dimension size of the projected features and $d_R$ is the dimension size of the output feature. We project $F$ to get the query vector $F_Q \in$

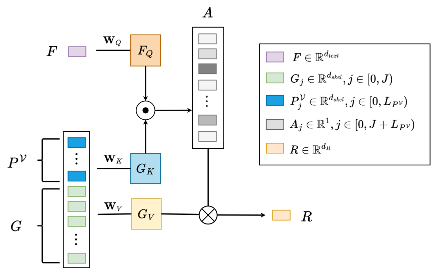



Figure 2. Illustration of the internal architecture of Adaptive Visual Representation Sampler. Each colored block represents either an text encoding output, a fake joint feature (*i.e*. skeleton-shaped prompt), or a true joint feature.

$\mathbb{R}^{1\times h}$ and find its semantic relevance with the projected key vector $G_K \in \mathbb{R}^{(J+L_{P^{\mathcal{V}}})\times h}$ of the concatenated vector of $P^{\mathcal{V}}$ and $G$. After obtaining the relevance score $A \in \mathbb{R}^{1\times(J+L_{P^{\mathcal{V}}})}$ through the external product between $F_Q$ and $G_K$, we can apply a dot production between $A$ and the projected value vector $G_V \in \mathbb{R}^{(J+L_{P^{\mathcal{V}}})\times d_R}$ to produce the visual representation $R \in \mathbb{R}^{d_R}$ that adaptively highlights the most class-relevant prompted visual contexts according to the given label.

$$F_Q = F\mathbf{W}_Q, \tag{11}$$

$$G_K = [P^{\mathcal{V}}, G]\mathbf{W}_K, G_V = [P^{\mathcal{V}}, G]\mathbf{W}_V, \tag{12}$$

$$A = \mathit{Softmax}((F_Q \times G_K^T)/\sqrt{h}), \tag{13}$$

$$R = AG_V. \tag{14}$$

As indicated in Fig. 1 (a), $R$ then goes through an MLP $\mathcal{V}$ to generate $V \in \mathbb{R}^{d_{text}}$, which expresses the projected visual embedding of $x$ in the text encoding space. $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are randomly initialized and learn along with $\mathcal{V}$.

### 3.4. Learning Losses

At this point, to achieve a complete learning optimization for generalization towards dual-modal alignment in ZSSAR, we have proposed respective prompting solutions for each branch that appropriately enhance the visual and textual representations of $V$ and $F$ from the pre-trained skeleton and CLIP backbones. During the training stage, we employ the standard contrastive learning loss [30] to minimize the embedding similarity between the projected visual and textual representations. This can be written as

$$\mathcal{L}_{train}(V, F) = -\frac{1}{2}\log\frac{\exp(\frac{VF}{\tau})}{\sum_o^{\mathcal{Y}^{batch}}\exp(\frac{VF_o}{\tau})} \tag{15}$$

$$-\frac{1}{2}\log\frac{\exp(\frac{VF}{\tau})}{\sum_w \exp(\frac{V_w F}{\tau})}, \tag{16}$$

where $F_o$ refers to the class textual embedding of all the other $o$ (negative) samples in the same batch, and $V_w$ represents the projected visual embedding from all the other $w$ (negative) skeleton samples. The temperature parameter $\tau$ regulates the training gradient.

During the inference stage, a skeleton sequence input $x_{te}$ goes through the frozen skeleton backbone to get $G_{x_{te}}$. Using $G_{x_{te}}$ and the learned prompted text encoder $\mathcal{T}$ to obtain the textual embedding $F_y$ and the corresponding representation $V_y^{x_{te}}$ for each candidate $y \in \mathcal{Y}^{uc}$ based on cross-attention query, the classification is based on the search for the nearest class neighbors of $V_y^{x_{te}}$. The following equations interpret the mentioned prediction process.

$$\mathcal{L}_{test}(V_y^{x_{te}}, F_y) = -\frac{1}{2}\log\frac{\exp(\frac{V_y^{x_{te}}F_y}{\tau})}{\sum_o^{\mathcal{Y}^{uc}}\exp(\frac{V_o^{x_{te}}F_o}{\tau})}, \tag{17}$$

$$\hat{y}_{te} = \mathop{\arg\min}_{y\in\mathcal{Y}^{uc}} \mathcal{L}_{test}(V_y^{x_{te}}, F_y). \tag{18}$$

## 4. Results

### 4.1. Datasets

We follow [22] to evaluate the ZSSAR performance of our model on three datasets according to the metrics of a) general performance and b) random three-split prediction from [22]. The splits are based on cross-subjects, ensuring that the training and testing subjects do not overlap. In a test split formatted as 'A/B', A refers to the number of seen classes, while B refers to the number of unseen classes.

***NTU-RGB+D 60*** [34] is a prominent benchmark dataset in skeleton-based HAR, offering a vast and diverse collection of data for 60 different common actions. It comprises $56,880$ video samples from $40$ subjects, captured from $80$ distinct camera viewpoints. In our experiment, we used its skeleton sequences collected from $25$ body joints per person (setting $2$ performers in each sample). Following the zero-shot evaluation protocols in [14], we tested our model under a $55/5$ split and a $48/12$ split.

***NTU-RGB+D 120*** [25] is an extension of NTU-60, which stands as one of the most extensive datasets for HAR at present. This dataset includes $114,480$ video samples across $120$ action classes, significantly expanding the diversity and complexity of actions compared to its predecessor. The actions are performed by $106$ subjects and captured from $155$ different camera viewpoints. Analogous to the previous dataset, we follow [14] and use the splits of $110/10$ and $96/24$ for the evaluation setups.

***PKU-MMD*** is a mid-scale dataset that encompasses 51 action categories and provides more than 20,000 action instances. It is notable for its inclusion of complex and interactive actions, which makes it suitable for evaluating algorithms in both individual and group activity recognition. We calculated the average accuracy from three random split settings in the format of 46/5. The selected splits are referenced from the same experiments in [22].

| Model | Backbone | NTU-RGBD 60 55/5 split | | | | NTU-RGBD 60 48/12 split | | | | NTU-RGBD 120 110/10 split | | | | NTU-RGBD 120 96/24 split | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ZSL | S | U | H | ZSL | S | U | H | ZSL | S | U | H | ZSL | S | U | H |
| ReViSE [39] | BERT | 53.91 | 74.22 | 34.73 | 47.32 | 17.49 | 62.36 | 20.77 | 31.16 | 55.04 | 48.69 | 44.84 | 46.68 | 32.38 | 49.66 | 25.06 | 33.31 |
| JPoSE [42] | BERT | 64.82 | 64.44 | 50.29 | 56.49 | 28.75 | 60.49 | 20.62 | 30.75 | 51.93 | 47.66 | 46.40 | 47.05 | 32.44 | 38.62 | 22.79 | 28.67 |
| CADA-VAE [33] | BERT | 76.84 | 69.38 | 61.79 | 65.37 | 28.96 | 51.32 | 27.03 | 35.41 | 59.53 | 47.16 | 49.78 | 48.44 | 35.77 | 41.11 | 34.14 | 37.31 |
| SynSE [14] | BERT | 75.81 | 61.27 | 56.93 | 59.02 | 33.30 | 52.21 | 27.85 | 36.33 | 62.69 | 52.51 | 57.60 | 54.94 | 38.70 | 56.39 | 32.25 | 41.04 |
| SMIE [50] | BERT | 77.98 | / | / | / | 40.18 | / | / | / | 65.74 | / | / | / | 45.30 | / | / | / |
| STAR [9] | CLIP | 81.40 | 69.00 | 69.90 | 69.40 | 45.10 | 62.70 | 37.00 | 46.60 | 63.30 | 59.90 | 52.70 | 56.10 | 44.30 | 51.20 | 36.90 | 42.90 |
| PURLS [51] | CLIP | 79.23 | / | / | / | 40.99 | / | / | / | 71.95 | / | / | / | 52.01 | / | / | / |
| SA-DVAE [22] | BERT | 82.37 | 62.28 | 70.80 | 66.27 | 41.38 | 50.20 | 36.94 | 42.56 | 68.77 | 61.10 | 59.75 | 60.42 | 46.12 | 58.82 | 35.79 | 44.50 |
| SCoPLe (ours) | CLIP | **84.10** | 69.60 | 71.94 | **70.75** | **52.96** | 54.49 | 61.83 | **57.93** | **74.53** | 63.51 | 61.08 | **62.27** | **52.17** | 53.33 | 51.18 | **52.23** |

Table 2. ZSL and GZSL accuracy (%) on the NTU RGB+D datasets. Each baseline chooses either a pretrained Sentence-BERT or CLIP-ViT-B/32 as its text backbone. S is for seen classes, U is for unseen classes, H is for the harmonic mean of seen/unseen classes.

| Model | NTU-RGBD 60 55/5 split | | | | NTU-RGBD 120 48/12 split | | | | PKU-MMD 46/5 split | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ZSL | S | U | H | ZSL | S | U | H | ZSL | S | U | H |
| ReViSE [39] | 60.94 | 71.75 | 52.06 | 60.34 | 44.90 | 48.29 | 34.64 | 40.34 | 59.34 | 60.89 | 42.16 | 49.82 |
| JPoSE [42] | 59.44 | 66.25 | 54.92 | 60.05 | 46.69 | 49.43 | 39.14 | 43.69 | 57.17 | 60.26 | 45.18 | 51.64 |
| CADA-VAE [33] | 61.84 | 77.35 | 58.14 | 66.38 | 45.15 | 51.09 | 41.24 | 45.64 | 60.74 | 63.17 | 35.86 | 45.75 |
| SynSE [14] | 64.19 | 75.84 | 60.77 | 67.47 | 47.28 | 41.73 | 45.36 | 43.47 | 53.85 | 63.09 | 40.69 | 49.47 |
| SMIE [50] | 65.08 | / | / | / | 46.40 | / | / | / | 60.83 | / | / | / |
| SA-DVAE [22] | **84.20** | 78.16 | 72.60 | 75.27 | 50.67 | 58.09 | 40.23 | 47.54 | 66.54 | 58.49 | 51.40 | 54.72 |
| SCoPLe (ours) | 83.72 | 75.32 | 80.17 | **77.67** | **53.34** | 70.47 | 44.29 | **54.08** | **71.41** | 62.17 | 49.69 | **54.85** |

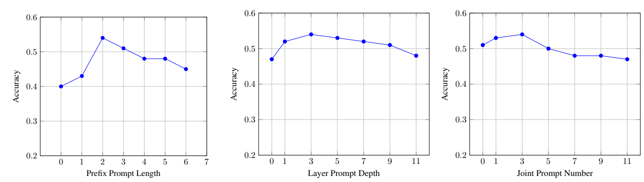
Table 3. Average ZSL and GZSL accuracy (%) under the random split settings on the NTU-60, NTU120, and PKU-MMD datasets.

## 4.2. Implementation Details

To fairly highlight the superiority of our method, we use Shift-GCN as the skeleton backbone, consistent with the previous baseline (see Sec. 3.1 in our **Supplementary Materials** for ablation study with various backbones). Our text prompt is embedded within a pre-trained CLIP-ViT-B/32 text encoder. Adhering to the training protocols in [14], we only use samples from the seen classes to pre-train the skeleton backbone for each split. The visual features are then obtained from the $d_{skel}$-dimensional penultimate layer of the backbone, preserving the standard feature format of $G \in \mathbb{R}^{J \times d_{skel}}$ where $d_{skel}$ is 256. The textual features $F$ are represented by the 512-dimensional encoding outputs.

We implement the projection layer $\mathcal{V}$ as a two-layer MLP with SiLU activations and a Tanh output layer. The size $d_{proj}$ of each hidden layer in $\mathcal{V}$ is 512. We set the length of prefix prompts $P_0^{\mathcal{T}}$ to 2 and the depth of layer prompt tuning $S$ to 3. The number of joint-shaped prompts $L_{P\mathcal{V}}$ is 3. Following the operations in [18], we randomly initialize the layer-level prompts $P_s^{\mathcal{T}}$ for each Transformer block of $s \in [0, S)$ according to a normal distribution $\sim N(0, 1)$. The Adaptive Visual Representation Sampler holds $\mathbf{W}_Q \in \mathbb{R}^{512\times150}$, $\mathbf{W}_K \in \mathbb{R}^{256\times150}$ and $\mathbf{W}_V \in \mathbb{R}^{256\times256}$ where $h = 150$ and $d_R = 256$, using single-headed attention learning. $\alpha_s$ in each prompted layer in $\mathcal{T}$ is initialized to 0.5 to balance contributions from the original text encoding and the prompted tuning streams at the beginning. We conducted our experiments using PyTorch on a machine equipped with a single A100 GPU. Each training process takes 100 epochs and a learning rate of $1e-4$. The batch size is set to 256. We use an Adam optimizer with $\beta1 = 0.9$ and $\beta2 = 0.999$.

**Adaptation to Generalized ZSSAR:** We design an extra step to mitigate the classification confusion between seen and unseen classes. Drawing inspiration from OOD domains, we implement a gating mechanism based on the prediction entropy confidence. In detail, we first let SCoPLe make similarity predictions $p_y = \mathcal{L}_{test}(V_y^{x^{te}}, F_y)$ for a testing sample $x_{te}$ with only seen class candidates $y \in \mathcal{Y}^{sc}$. Using this as an inference logit, we obtain the entropy $H_y$ by calculating $H_y = -p_y \log p_y$ for all seen class candidates. Given a pre-defined entropy threshold $\delta$, we assume that the model will more confidently predict $H_y$ to be below $\delta$ if $x_{te}$ has a ground truth label $y$ from seen classes, since its feature should exhibit more representative characteristics when being predicted as a seen class. If all $H_y$ exceed $\delta$, we then perform the standard zero-shot classification, in which the model classifies $x_{test}$ as one of the unseen classes in $\mathcal{Y}^{uc}$.




(a) (b) (c)

Figure 3. Ablation study on Prefix Prompt Length (a), Layer Prompt Depth (b), and Joint-shaped Prompt Number (c) in SCoPLe with the testing split of 12/48 in NTU-60.

## 4.3. Overall Performance

As shown in Tab. 2, we compare our method with all previous ZSSAR solutions following the same testing protocols originally established by [14]. For zero-shot learning, SCoPLe outperforms other baselines and achieves state-of-the-art prediction accuracy in each split case. Note that, due to varying design requirements, the textual backbones used among methods can differ and may significantly impact performance improvement. Nevertheless, during the random 3-split testing experiment, all solutions are equipped with the same CLIP backbone, and Tab. 3 demonstrates that our method still has an advantage over the previous SOTA baseline of SA-DVAE [22] for the 48/12 split and the 46/5 split, with accuracy increases of +2.67%, and +4.87%. Similarly, for the more challenging task of Generalized ZSSAR in both the general and 3-split protocols, SCoPLe improves

| Model | NTU-RGBD 60 | | | | | | | | NTU-RGBD 120 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 55/5 split | | | | 48/12 split | | | | 110/10 split | | | | 96/24 split | | | |
| | ZSL | S | U | H | ZSL | S | U | H | ZSL | S | U | H | ZSL | S | U | H |
| SA-DVAE [22] (last SOTA) | 82.37 | 62.28 | 70.80 | 66.27 | 41.38 | 50.20 | 36.94 | 42.56 | 68.77 | 61.10 | 59.75 | 60.42 | 46.12 | 58.82 | 35.79 | 44.50 |
| SCoPLe w/o all prompts | 62.58 | 41.16 | 80.63 | 54.50 | 33.16 | 53.78 | 58.35 | 55.97 | 63.92 | 39.13 | 62.35 | 48.08 | 48.15 | 32.02 | 64.33 | 42.76 |
| SCoPLe w/o text prompts | 63.39 | 37.21 | 82.87 | 51.36 | 32.41 | 53.67 | 57.48 | 55.51 | 64.69 | 49.69 | 61.00 | 54.77 | 46.14 | 32.22 | 64.24 | 42.92 |
| SCoPLe w/ text prompts (no dual) | 75.32 | 57.85 | 70.61 | 63.59 | 44.73 | 62.74 | 51.14 | 56.35 | 65.93 | 58.66 | 56.90 | 57.76 | 46.02 | 54.18 | 41.74 | 47.15 |
| SCoPLe w/o skel. prompts | 70.21 | 67.92 | 61.03 | 64.29 | 40.69 | 55.35 | 55.94 | 55.64 | 68.88 | 67.82 | 43.75 | 53.19 | 48.17 | 51.84 | 49.46 | 50.62 |
| SCoPLe w/ skel. prompts (no sampler) | 70.72 | 65.60 | 66.04 | 65.82 | 40.78 | 48.90 | 61.93 | 54.65 | 67.92 | 67.47 | 45.17 | 54.11 | 40.41 | 48.66 | 52.93 | 50.71 |
| SCoPLe (full) | **84.10** | 69.60 | 71.94 | **70.75** | **52.96** | 54.49 | 61.83 | **57.93** | **74.53** | 63.51 | 61.08 | **62.27** | **52.17** | 53.33 | 51.18 | **52.23** |

Table 4. Ablation study on each proposed module with all testing splits for ZSL and GZSL on NTU-RGB+D datasets.

the best hybrid accuracy of previous baselines with +1.35%, +11.33%, +1.85%, +7.73% in Tab. 2, and +2.40%, +6.54%, +0.13% in Tab. 3. In Sec. 2.4 of our **Supplementary Materials**, we include additional experiments to examine the limit when training data becomes significantly insufficient to produce a valid effect on ZSSAR.

Furthermore, we made a training efficiency comparison between PURLS and SCoPLe using the 110/10 split on NTU 120. The latter simplifies the operations of preparing eight labor-intensive text description embedding (7 local&1 global) and uses the original single label embedding with cross-attention query, achieving better performance with a lower training time of 53 mins while PURLS took 72 mins.

## 4.4. Ablation Study

**Tuning impacts of different prompting modules.** Tab. 4 records our ablation experiment results on NTU datasets by removing each proposed prompting module or mechanism along the linguistic or visual branches. Specifically, we thoroughly examine the gradual ablation by testing the removal of all prompt learning (**w/o all prompts**), all layer-level text prompting (**w/o text prompts**), the attention normalization introduced by the dual-stream mechanism (**w/ text prompts (no dual)**), all skeleton prompting (**w/o skeleton prompts**), and the Adaptive Visual Representation Sampler (**w/ skeleton prompts (no sampler)**). Their results are horizontally compared with the performance of the full model (**full**). Based on the outcomes, we can arrive at three obvious conclusions. First, by comparing the ablation study among **w/o all prompts**, **w/o text prompts**, **w/o skeleton prompts**, and **full**, we can validate that each prompting brings orthogonal prediction improvements over directly using CLIP. Especially, using only skeleton prompting without any language guidance and textual prompting (*i.e.* **w/o text prompts**) results in unstable advancement. This again corroborates the fact that ZSL is highly dependent on the quality of language embedding to enhance the efficiency of knowledge transfer. In addition, comparing the study between **w/ text prompts (no dual)** and **full**, we find that adding dual-stream normalization not only improves prediction in ZSL protocols, but also better balances prediction accuracy between seen and unseen classes for GZSL protocols, thus improving overall performance in this scenario. Finally, comparing the study between **w/ skeleton prompts (no sampler)** and **full**, we reveal that the introduction of textual guidance through the adaptive visual representation sampler significantly improves the auxiliary effect of visual side prompting.

**Hyperparameter Selection.** In Fig. 3, we conduct a hyperparameter grid search and illustrate the effect of changing the length of $P_0^{\mathcal{T}}$ (left), the layer prompting depth of $S$ (middle), and $L_{P\mathcal{V}}$ (right). We find that optimal design choices occur when the length of $P_0^{\mathcal{T}}$ is 2, $S$ is 3, and $L_{P\mathcal{V}}$ is 3. The performance for values larger than these begins to decrease, which can be interpreted as an overfitting situation where the model cannot effectively allocate new parameters to enhance prediction generalization.

A key characteristic of SCoPLe is that each module is entirely data-driven and functions at full feature level. This theoretically enhances its versatility beyond skeleton-based recognition, especially for the text-based prompting brought by DSLP. In Sec. 4 of our **Supplementary Materials**, we have also conducted a preliminary exploration of image-based base-to-novel generalization tasks.

# 5. Conclusion

In conclusion, this paper presents SCoPLe, a novel semantic-guided cross-modal prompt learning framework. By introducing dual-stream language prompting learning and semantic-guided adaptive skeleton prompting learning modules, our approach shifts from manually crafted assumptions to a purely data-driven solution for adapting PLMs to skeleton-based zero-shot learning tasks. The experimental results demonstrate that SCoPLe achieves state-of-the-art performance on the NTU-RGB+D 60, NTU-RGB+D 120, and PKU-MMD datasets. A limitation of our method compared to manually refined solutions like PURLS [51] is the deficiency of interpretability, considering that SCoPLe trades it for improved generalization and efficiency through automatic learning on prompt tuning. In addition, future research can further explore the extension of the concept of our prompting normalization to other data modalities.

# 6. Acknowledgement

This research was partially supported by the Australian Government through the Australian Research Council's DECRA funding scheme (Grant No: DE250100030) and Discovery Projects scheme (Grant No: DP230101534).

## References


[1] Automatic video annotation of human health care action via clustering. *J. Medical Imaging Health Informatics*, 10(10): 2512–2521, 2020.

[2] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Cesar Cadena, Anthony Dick, and Ian Reid. Multi-modal auto-encoders as joint estimators for robotics scene understanding. 2016.

[5] Yi Cao, Chen Liu, Zilong Huang, Yongjian Sheng, and Yongjian Ju. Skeleton-based action recognition with temporal action graph and temporal adaptive graph convolution structure. 80(19):29139–29162, 2021.

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[7] Ronald Juárez Chambi, Juan Carlos Gutiérrez Cáceres, and César Armando Beltrán Castañón. Densenet 3d for violent action recognition in surveillance video sequences. In *41st International Conference of the Chilean Computer Science Society, SCCC 2022, Santiago, Chile, November 21-25, 2022*, pages 1–8. IEEE, 2022.

[8] 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.

[9] Yang Chen, Jingcai Guo, Tian He, and Ling Wang. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition, 2024.

[10] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 180–189, 2020.

[11] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022.

[12] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013.

[13] Thomas Golda, Johanna Thiemich, Mickael Cormier, and Jürgen Beyerer. For the sake of privacy: Skeleton-based salient behavior recognition. In *2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16-19 October 2022*, pages 3983–3987. IEEE, 2022.

[14] 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, 2021.

[15] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning, 2022.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[17] Mustaqeem Khan, Jamil Ahmad, Abdulmotaleb El-Saddik, Wail Gueaieb, Giulia De Masi, and Fakhri Karray. Drone-hat: Hybrid attention transformer for complex action recognition in drone surveillance videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 4713–4722. IEEE, 2024.

[18] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning, 2023.

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

[20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

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

[22] Sheng-Wei Li, Zi-Xiang Wei, Wei-Jie Chen, Yi-Hsin Yu, Chih-Yuan Yang, and Jane Yung jen Hsu. Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders, 2024.

[23] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, 2021. Association for Computational Linguistics.

[24] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Bo Wu, Yifeng Lu,

Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection, 2022.

[25] 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.

[26] Qiang Nie and Yunhui Liu. View transfer on human skeleton pose: Automatically disentangle the view-variant and view-invariant information for pose representation learning. *IJCV*, 129(1):1–22, 2021.

[27] Yoshihisa Nitta and Yuko Murayama. Privacy-aware remote monitoring system by skeleton recognition. In *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, pages 1–8. ScholarSpace, 2019.

[28] Quang-Tien Pham, Duc-Anh Nguyen, Tien-Thanh Nguyen, Thanh Nam Nguyen, Duy-Tung Nguyen, Dinh-Tan Pham, Thanh-Hai Tran, Thi-Lan Le, and Hai Vu. A study on skeleton-based action recognition and its application to physical exercise recognition. In *The 11th International Symposium on Information and Communication Technology, SoICT 2022, Hanoi, Vietnam, December 1-3, 2022*, pages 239–246. ACM, 2022.

[29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[31] M Rashmi and Ram Mohana Reddy Guddeti. Skeleton based human action recognition for smart city application using deep learning. In *2020 international conference on communication systems & networks (COMSNETS)*, pages 756–761. IEEE, 2020.

[32] Justin Sanchez, Christopher Neff, and Hamed Tabkhi. Real-world graph convolution networks (rw-gcns) for action recognition in smart video surveillance. *CoRR*, abs/2201.05739, 2022.

[33] E. Schonfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata. Generalized zero- and few-shot learning via aligned variational autoencoders. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8239–8247, Los Alamitos, CA, USA, 2019. IEEE Computer Society.

[34] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016.

[35] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035. Computer Vision Foundation / IEEE, 2019.

[36] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *CVPR*, pages 7912–7921. Computer Vision Foundation / IEEE, 2019.

[37] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020.

[38] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022.

[39] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings, 2017.

[40] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds, 2022.

[41] Christian Wojek, Stefan Walk, Stefan Roth, and Bernt Schiele. Monocular 3d scene understanding with explicit occlusion reasoning. pages 1993–2000, 2011.

[42] Michael Wray, Gabriela Csurka, Diane Larlus, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 450–459, 2019.

[43] Wentian Xin, Ruyi Liu, Yi Liu, Yu Chen, Wenxin Yu, and Qiguang Miao. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*, 537: 164–186, 2023.

[44] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.

[45] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021.

[46] Yan Zhang. *Human action parsing in untrimmed videos and its applications for elderly people healthcare*. PhD thesis, University of Ulm, Germany, 2020.

[47] Yunkai Zhang, Yinghong Tian, Pingyi Wu, and Dongfan Chen. Application of skeleton data and long short-term memory in action recognition of children with autism spectrum disorder. *Sensors*, 21(2):411, 2021.

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022.

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[50] 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, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5302–5310. ACM, 2023.

[51] Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. Part-aware unified representation of language and skeleton for zero-shot action recognition. *CoRR*, abs/2406.13327, 2024.