

Zero-shot Skeleton-based Action Recognition via Mutual Information Estimation and Maximization

Yujie Zhou
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing, China
yujiezhou@ruc.edu.cn

Wenwen Qiang
University of Chinese Academy of
Sciences
Institute of Software Chinese
Academy of Sciences
Beijing, China
wenwen2018@iscas.ac.cn

Anyi Rao
Stanford University
Stanford, CA, USA
anyirao@stanford.edu

Ning Lin
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing, China
linning51400@ruc.edu.cn

Bing Su*
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing Key Laboratory of Big Data
Management and Analysis Methods
Beijing, China
subingats@gmail.com

Jiaqi Wang
Shanghai AI Laboratory
Shanghai, China
wjqdev@gmail.com

ABSTRACT

Zero-shot skeleton-based action recognition aims to recognize actions of unseen categories after training on data of seen categories. The key is to build the connection between visual and semantic space from seen to unseen classes. Previous studies have primarily focused on encoding sequences into a singular feature vector, with subsequent mapping the features to an identical anchor point within the embedded space. Their performance is hindered by 1) the ignorance of the global visual/semantic distribution alignment, which results in a limitation to capture the true interdependence between the two spaces. 2) the negligence of temporal information since the frame-wise features with rich action clues are directly pooled into a single feature vector. We propose a new zero-shot skeleton-based action recognition method via mutual information (MI) estimation and maximization. Specifically, 1) we maximize the MI between visual and semantic space for distribution alignment; 2) we leverage the temporal information for estimating the MI by encouraging MI to increase as more frames are observed. Extensive experiments on three large-scale skeleton action datasets confirm the effectiveness of our method.

CCS CONCEPTS

• **Computing methodologies** → *Activity recognition and understanding.*

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3611888>

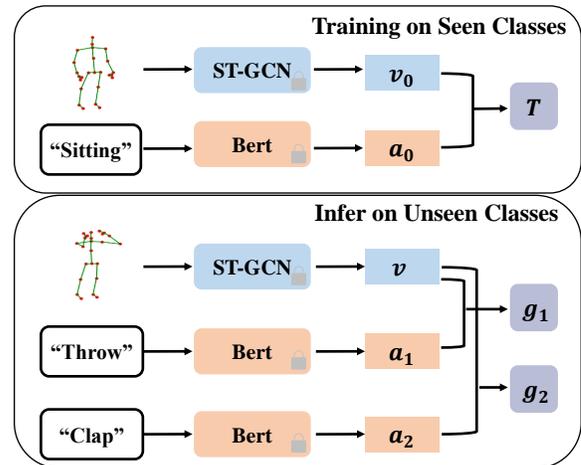


Figure 1: The core of the zero-shot learning lies in constructing a connection model T between visual features v and semantic features a during the training phase. At test time, the learned model T is utilized to predict the most compatible semantic attribute for a given unseen-class visual feature.

KEYWORDS

Zero-shot Learning, Human Skeleton Data, Action Recognition

ACM Reference Format:

Yujie Zhou, Wenwen Qiang, Anyi Rao, Ning Lin, Bing Su, and Jiaqi Wang. 2023. Zero-shot Skeleton-based Action Recognition via Mutual Information Estimation and Maximization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3611888>

1 INTRODUCTION

Human action recognition becomes an essential component in many real-world applications, including but not limited to security

and human-robot interaction. It is not hard to acquire skeleton data due to the development of pose estimation [13] and sensors [45]. To this end, human skeleton data has emerged as a promising alternative to traditional RGB video data due to its robustness to variations in appearance and background, as well as its ability to provide an unbiased representation of individuals.

Many researchers explore fully supervised methods for this task, which requires large amounts of labeled training samples. However, it is not economical to handle numerous action classes in real-world scenarios since the samples of many actions are time-consuming and expensive to collect. Thus, zero-shot learning [18, 29] is used to recognize new classes if there are no training samples but only some semantic information such as the names, attributes, or descriptions of new classes is available. Annotating and labeling 3D skeleton action data presents increased difficulty due to the inclusion of depth information and the complexity of human action semantics. Hence, zero-shot skeleton-based action recognition is highly desirable in practical applications because it can significantly reduce the need for collecting and annotating new actions.

In Fig. 1, Given pre-extracted visual and semantic features, the core of zero-shot learning is to establish a connection model between visual and semantic spaces in the seen classes. During the test phase, the learned model is used to facilitate the knowledge transfer from the seen to the unseen classes. To address the transfer learning problem, zero-shot action recognition relies on the external knowledge base, i.e., the semantic embeddings of each class label from pre-trained large-scale language model such as Sentence-Bert [28] or CLIP [27]. The effective utilization of semantic information is important for bridging the gap between two different modalities.

There are a few studies on zero-shot skeleton-based action recognition. Existing methods [9, 14] embed action sequences into visual features. To establish the connection model of visual and semantic space, a compatible projection function [14] or a deep metric [6] is learned based on the data of seen classes in the training phase. Then in the testing phase, the similarities between the visual feature of a test action sequence and the sentence embeddings [14] or part-of-speech tagged words [41] of the unseen classes are measured either in the projected common space or by the learned metrics. However, the projection operation merely maps the visual or semantic features to a common anchor point in the embedding space, overlooking the global alignment between the distributions of visual and semantic features. Furthermore, the learned projections or metrics inadequately utilize semantic information to capture the associations between the two modalities. Their attempt to perform cross-modal reconstruction without aligning the distributions is challenging due to the significant difference between the visual and semantic spaces, ultimately resulting in the difficulty of generalizing to novel classes with diverse distributions.

Secondly, The information loss becomes severe in the zero-shot action recognition task scenario since some semantic classes require dynamics information to differentiate from each other. For example, "walking" and "skipping" differ only in local parts since the initial frames for these two action sequences are similar; "Skipping" cannot be identified until a human-rising procedure is observed. Thus, for a human action, utilizing the inherent temporal dynamics information also plays a role in the generalization ability of the zero-shot connection model.

In this paper, we propose a **Skeleton-based Mutual Information Estimation and maximization framework for zero-shot action recognition (SMIE)**. To better capture the dependencies between visual and semantic spaces, our approach avoids direct mapping and instead aligns the distributions of these two spaces using a global alignment module. This module utilizes mutual information as a measure of similarity and applies an estimator based on Jensen-Shannon divergence (JSD) to maximize the mutual information between paired visual and semantic features while minimizing mutual information between unpaired visual and semantic features. A neural network is employed as the connecting model to estimate the similarity score in the JSD estimator, which is used during the test phase on unseen classes. Then considering the inherent temporal information of actions, SMIE proposes a temporal constraint module to encourages the mutual information between visual and semantic features to increase when more parts of the action are executed. Specifically, the JSD estimator applies contrastive learning to estimate the global mutual information. The paired visual and semantic features form positive samples, while unpaired ones form negative samples. To perceive keyframes that contain more discriminative information in the action sequence, the temporal constraint module computes the motion attention of each sequence and masks the keyframes with higher attention to generate extra positive samples, which contain partial temporal information loss. During training, the temporal-constrained mutual information is computed with the same negative samples and is kept smaller than the global mutual information.

The major contributions of this paper are three-fold:

- We propose a skeleton-based mutual information estimation and maximization framework (SMIE), a new zero-shot approach to skeleton-based action recognition based on mutual information maximization, which can capture the complex statistical correlations between the distributions of the visual space and the text semantic space.
- A novel temporal constraint module is proposed to compute the temporal-constrained mutual information and a temporal rank loss is applied to help the connection model capture the inherent temporal information of actions.
- Extensive experiments and analyses demonstrate the effectiveness of the proposed method, which outperforms the baseline methods by a large margin.

2 RELATED WORK

Zero-shot Action Recognition. Most of the existing zero-shot video classification methods aim to build the connection between the visual and semantic spaces using feature projections, which mainly focuses on the visual space [10, 17, 40], the semantic space [1, 47], and the intermediate space [7, 42]. Specifically, The visual features are first extracted from videos using a pre-trained network such as Convolutional 3D Network (C3D) [36], ResNet [11], and Inflated 3D Network(I3D) [3]. And then they map the visual or semantic features to the fixed anchor points in the embedding space. Different from these works, we focus on zero-shot skeleton-based action recognition, where the visual features for skeleton-based action sequences greatly differ from those for RGB videos. We use mutual information instead of projections to associate the skeleton-based visual and semantic features.

Skeleton-based Action Recognition. With the development of highly accurate depth sensors such as Kinect cameras and pose estimation algorithms [2, 32], skeleton-based action recognition has attracted increasing attention recently. Human skeleton-based representation [8, 19] is robust to variations of appearance and background environment, where each skeleton contains different types of joints, and each joint records its 3D position.

Specifically, skeletons are organized as pseudo-images [15, 16] or a sequence of long-term contextual information [5] and feed it into CNNs/RNNs. Later, ST-GCN [43] constructs a predefined spatial graph based on the natural connections of joints in the human body and utilizes GCN to integrate the skeleton joint information. In terms of consecutive frames, ST-GCN constructs the temporal edges between corresponding joints. After that, many variants of ST-GCN are proposed [4, 31, 39, 44], which contain more data streams or add attention mechanisms. In this paper, we employ ST-GCN [43] and Shift-GCN [4] as the backbone to extract visual features. To explore the temporal relationship of skeleton data, we retain the pre-trained GCN model and acquire the partial visual feature by inputting frames with different indices in the skeleton sequences.

Zero-shot Skeleton-Based Action Recognition. Fewer works have been devoted to zero-shot skeleton-based action recognition though it is of great importance. DeVISE [6] and RelationNet [14] are extended to tackle this problem by extracting visual features from skeleton sequences with ST-GCN and semantic embeddings with Word2Vec [23] or Sentence-Bert [28]. DeVISE uses a simple learnable linear projection between the visual and semantic feature spaces. Based on it, RelationNet utilizes an attribute network and a relation network to achieve the same goal. In the above works, a projection operation is needed to build the visual-semantic embedding, they map the visual or semantic features to the fixed anchor points in the embedding space, which does not consider the global distribution of the semantic features. Recently, SynSE [9] uses a generative multi-modal alignment module to align the visual features with parts of speech-tagged words. It works but needs extra PoS syntactic information to divide labels into verbs and nouns. Our method differs from these works in two aspects: We maximize the mutual information between the two modalities, and the distribution between visual and semantic features can be aligned. The global information of the semantic distribution can be utilized, which can help to guide the knowledge transfer from seen domain to the unseen domain. Second, to exploit the temporal information of the skeleton sequence, a temporal constraint module is used to encourage the mutual information between visual and semantic features to increase with the number of observed frames.

Mutual Information for Zero-shot Learning. Recently, the mutual information between the visual space and the semantic space has been explored in zero-shot learning. In [33], local patches from other images with the same label are drawn to estimate the mutual information for local interpretability. In [34], mutual information is utilized to learn latent visual and semantic representations so that multi-modalities can be aligned for generalized zero-shot learning. Different from the above, we use mutual information to bridge the skeleton-based visual space and the text label semantic space and

impose a novel constraint on the mutual information to capture the temporal semantic of the skeleton sequences.

3 METHOD

3.1 Problem Definition

The zero-shot learning setting addressed in this paper is the same as [14], where the model is trained on seen classes and tested on disjoint unseen classes. Specifically, the training dataset consists of the skeleton sequence and the corresponding class name from **seen** classes. Each training sample is denoted as (x^s, e^s) , $x^s \in \mathbb{R}^{K \times J \times C}$ represents the training skeleton sequence with K frames and J recorded 3D-joints for each frame, and e^s is the corresponding class name. The test dataset comes from **unseen** classes, a sample of the test dataset is denoted as (x^u, e^u) .

Generally, we use a visual feature extractor F_v , which has been pre-trained on seen classes, and a semantic feature extractor F_e , which has been pre-trained on large-scale language models. These two extractors are employed to acquire the visual and semantic features v and a by taking in the skeleton sequence and class name as inputs. Then a layer-norm layer (without learnable parameters) N is used to normalize the visual features,

$$v^i = N(F_v(x^i)), a^i = F_e(e^i), i \in \{s, u\}. \quad (1)$$

Let V and A represent the random variables of v and a respectively. The zero-shot learning aims to classify the sample of the unseen classes by the model learned based on the training data from seen classes. The utilization of semantic features is important because the semantic space is shared between seen and unseen classes, which can help the learned model transfer knowledge from different domains. For simplicity, we omit the subscript for seen (s) and unseen (u) classes during the following model training introduction.

3.2 Method Overview

In Fig. 2, we propose a skeleton-based mutual information estimation and maximization framework (SMIE). Our SMIE consists of two modules: The global alignment module uses mutual information estimation and maximization to capture the statistical correlations between visual and semantic distributions. The temporal constraint module is utilized to enable the connection model T to perceive the keyframes of sequences, which allows for the exploration of action dynamics and the capture of inherent temporal information.

In the inference phase, the trained mutual information estimation network T calculates the similarity score g between tested visual sequences and all semantic features of unseen classes. The unseen class with the highest similarity score is chosen as the prediction.

3.3 Global Alignment

Previous zero-shot skeleton-based action recognition approaches [6, 14] learn projections that pull the visual features and the corresponding semantic features closer in seen classes without considering the whole distribution of features. Due to the huge gap between visual and semantic spaces, it is difficult to bridge the cross-domain gap and generalize the projections to unseen classes.

To tackle this issue, we design a global alignment module to learn an estimation network T through maximizing the mutual information $I(V; A)$ [25, 38] between the random variables of visual

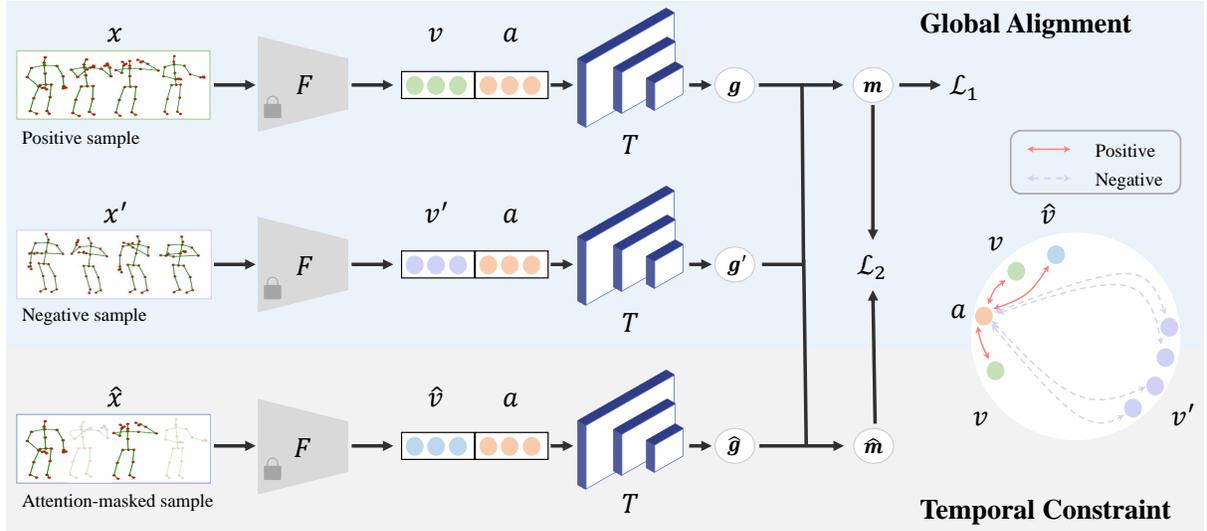


Figure 2: The framework of our proposed SMIE includes a global alignment module to estimate the mutual information between visual and semantic features, and a temporal constraint module to incorporate temporal dynamics information into the estimation network.

and semantic features:

$$\begin{aligned} I(V; A) &= D_{KL}(p(v, a) || p(x)p(a)) \\ &= \mathbb{E}_{p(v, a)} \left[\log \frac{p(v|a)}{p(v)} \right]. \end{aligned} \quad (2)$$

D_{KL} is the KL-divergence between $p(v, a)$ and $p(v)p(a)$, which represents the joint distribution and product of the marginal distributions of v and a , respectively. The information on joint distribution can be utilized, and help the model use the global semantic attributes. Meanwhile, the mutual information can be written with the joint entropy among V , A and their conditional entropy,

$$I(V; A) = H(V, A) - (H(V | A) + H(A | V)). \quad (3)$$

As pointed out above, maximizing the mutual information between V and A is equivalent to maximizing the common information, i.e., the difference between joint entropy and conditional entropy.

Following Eq. (2), instead of directly modeling $p(v | a)$, mutual information is utilized to encode V and A into compact distributed vector representations via a connection network learned from data. Benefiting from it, the captured shared information between visual and semantic features maintains a better global structure, while low-level information and noise will be discarded.

However, directly calculating the mutual information of two random variables in high-dimensional spaces is extremely hard. Inspired by the Jensen-Shannon divergence (JSD) [24], we propose to learn an estimation network T that takes the global visual feature v and the semantic feature a as inputs. The output of the network can be served as a similarity metric score g between the visual and semantic features, which is applied to match a skeleton sequence to the unseen classes in the testing phase.

The estimation network T can be trained by maximizing the following JSD estimator.

$$\begin{aligned} I(V; A) \approx m &= \mathbb{E}_{p(v, a)} [-f_{sp}(-T(v, a))] \\ &\quad - \mathbb{E}_{p(v)p(a)} [f_{sp}(T(v', a))], \end{aligned} \quad (4)$$

where m is the estimated mutual information shown in Fig. 2. Note that (v, a) are paired visual/semantic features and (v', a) are negative pairs. Specifically, v is a visual feature extracted from skeleton sequences x , which is related to a . And v' is a visual feature extracted from negative sample x' related to other classes. f_{sp} is the soft-plus function $f_{sp}(z) = \log(1 + e^z)$. Then, the visual feature v is concatenated with the paired semantic feature a to form the positive pair, while the negative pair is generated by concatenating this semantic feature with the visual feature v' of another sequence. The two pairs are sent to the estimation network T to obtain the scores g and g' , respectively, for contrastive learning.

In this way, the estimation network encourages the semantic feature to have a larger similarity to the corresponding visual feature than those unpaired visual features. Thus, to maximize the estimated mutual information m for the global alignment, we have the following loss:

$$\mathcal{L}_1 = -m. \quad (5)$$

The parameters of the estimation network T are updated by gradient descent during training.

3.4 Temporal Constraint Module

Compared with the image data, 3D human skeleton data is more complicated because of the additional temporal dimension. Utilizing the temporal dynamics information within a skeleton sequence can help the model capture subtle differences among various classes. A temporal constraint module is proposed to incorporate such temporal information.

Generally, for human action sequences, the more frames observed the more dynamics information model can capture, which encourages the visual feature to get a stronger correlation with its corresponding semantic feature. Furthermore, the keyframes in the action sequences are often richer in discriminative information, so the sequence loss of such frames has lower semantic relevance to their labels. Inspired by PSTL [46], which utilizes the motion of

skeleton data to find the key frames of each sequence. We further adopt bidirectional motion attention to enhance the effectiveness. As shown in the bottom half of Fig. 2, to acquire the attention-masked sample, we first calculate the bidirectional action attention for each action sequence. Specifically, the motion $p \in \mathbb{R}^{K \times J \times C}$ of the sequence is computed by the temporal displacement between frames: $p_{k,j,c}^{nex} = x_{k+1,j,c} - x_{k,j,c}$, which represents the subsequent variation of action in each frame. Then we further incorporate the displacement between the current frame and its preceding frame into the motion information: $p_{k,j,c}^{pre} = x_{k-1,j,c} - x_{k,j,c}$. The bidirectional motion of the sequence can be defined as:

$$p_{k,j,c} = (p_{k,j,c}^{nex})^2 + (p_{k,j,c}^{pre})^2. \quad (6)$$

Then, we calculate the average motion value for each frame to acquire p_k :

$$p_k = \frac{1}{J \times C} \sum_{j=1}^J \sum_{c=1}^C p_{k,j,c}. \quad (7)$$

With the bidirectional motion p_k , we can acquire the overall motion rate of a frame which serves as the bidirectional attention weight:

$$q_k = \frac{p_k}{\sum_{i=1}^K p_i}. \quad (8)$$

Then the top P frames with the highest attention scores q_{k_1}, \dots, q_{k_P} are selected and the frame list x_{k_1}, \dots, x_{k_P} serves as the keyframes which contain more discriminative information about the action. We mask such key frames on the original skeleton sequence x to construct the attention-masked sample sequence \hat{x} , which suffers some information loss and has lower semantic relevance to their semantic feature. With the help of the visual feature extractor F and layer-norm layer N , the temporal-constrained visual feature \hat{v} is extracted and concatenated with the corresponding semantic feature a to construct the temporal-constrained positive pair.

Similar to the usage of the JSD estimator in the global alignment module, the temporal-constrained mutual information \hat{m} is formulated as follows,

$$\hat{m} = \mathbb{E}_{p(\hat{v}, a)} [-f_{sp}(-T(\hat{v}, a))] - \mathbb{E}_{p(\hat{v})p(a)} [f_{sp}(T(v', a))], \quad (9)$$

Here, the mutual information between the temporal-constrained visual features and corresponding semantic features is maximized. Note that the negative pair still consist of the original negative sample x' and the unpaired semantic feature a to ensure the consistency of the negative sample space. Our temporal constraint module aims to encourage the connection module to perceive the importance of the keyframes during mutual information estimation. So a hinge loss is utilized to force the global mutual information m greater than the partial one during training,

$$\mathcal{L}_2 = \max(0, \beta - (m - \hat{m})), \quad (10)$$

where β is a hyper-parameter to control the distance between two types of mutual information. By adjusting β , the model can adapt to different datasets. In brief, the temporal constraint module serves as a regularization on the JSD estimator, which helps the model incorporate the dynamics information and be more robust.

Table 1: The hyper-parameters on NTU-60, NTU-120, and PKU-MMD datasets.

Parameter	NTU-60	NTU-120	PKU-MMD
β	0.1	0.5	0.01
P	15	15	15

The overall loss function combines the global mutual information maximization term and the temporal constraint term together, as shown in the following,

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2. \quad (11)$$

λ is the trade-off parameter and is set to 0.5 for all experiments.

4 EXPERIMENTS

4.1 Datasets

NTU-RGB+D 60 [30] contains 56,578 skeleton sequences of 60 action categories, performed by 40 volunteers. The skeleton sequences are collected by Microsoft Kinect sensors and each subject is represented by 25 joints. Two official dataset splits are applied: 1) Cross-Subject (xsub): the training set contains half of the subjects, and the rest make up the testing sets; 2) Cross-View (xview): The data from different views constitute the training and test set.

NTU-RGB+D 120 [21] is the extended version of the NTU-60. It is performed by 106 volunteers and contains 113,945 skeleton sequences of 120 action categories. NTU-120 also has two official dataset splits: 1) Cross-Subject (xsub): 53 subjects belong to the training set and the testing data is performed by the rest volunteers; 2) Cross-Setup (xset): the training set is captured by cameras with even IDs and the test set is captured with odd IDs.

PKU-MMD [22] has almost 20000 action samples in 51 categories collected by 66 subjects. It is captured via the Kinect v2 sensors from multiple viewpoints. The dataset has two parts: 1) Part I contains 21539 samples; 2) Part II contains 6904 samples.

4.2 Implementation Details and Baselines

Detailed Implementation of SMIE. We follow the same data processing procedure in Cross-CLR [19], which removes the invalid frames and resizes the skeleton sequences to 50 frames by linear interpolation. ST-GCN [43] with 16 hidden channels is used as the visual feature extractor and the extracted feature dimension is 256. For the semantic feature, we use Sentence-Bert [28] to obtain the 768-dimensional word embeddings, and then all semantic features are processed by L2 normalization to improve the stability of the training phase. For all experiments, we adopt the Adam optimizer and the CosineAnnealing scheduler with 100 epochs. The mini-batch size is 128. The learning rate is $1e-5$ for NTU-60 and PKU-MMD datasets, while for the NTU-120 dataset with larger data size, the learning rate is $1e-4$. Tab. 1 shows the choices of the hyper-parameters for all datasets. P refers to the number of masked keyframes, which remains 15 for all three datasets. The hyper-parameter margin β controls the distance between global and temporal-constrained mutual information. By adjusting β , the model can adapt to different datasets. Note that, with decreasing margins, the impact of temporal constraints on the overall loss

Table 2: Comparison of SMIE with the State-of-the-Art methods on NTU-60 and NTU-120 datasets.

Method Split	NTU-60(%)		NTU-120(%)	
	55/5	48/12	110/10	96/24
DeViSE	60.72	24.51	47.49	25.74
RelationNet	40.12	30.06	52.59	29.06
ReViSE	53.91	17.49	55.04	32.38
JPoSE	64.82	28.75	51.93	32.44
CADA-VAE	76.84	28.96	59.53	35.77
SynSE	75.81	33.30	62.69	38.70
SMIE	77.98	40.18	65.74	45.30

increases. Table 1 indicates a positive correlation between dataset size and margin parameter β . Smaller datasets necessitate a smaller β for optimal performance. The reason is that the temporal constraint module serves as a regularization on the JSD estimator. By providing necessary constraints on the model, the module prevents overfitting to the limited dataset.

For the details of our SMIE model, the network T in Eq. (4) is composed of three MLP layers with ReLU activation functions. For the negative pairs, we shift the skeleton visual samples in a batch to make the visual and semantic features do not correspond.

Baseline Methods. The core of skeleton-based zero-shot learning lies in evaluating the efficacy of the connection model, which functions as the intermediary between the visual and semantic spaces. The recent method SynSE [9] follows this main idea and provides various zero-shot learning comparative methods, such as Devise [6], ReViSE [37], RelationNet [14], JPoSE [41] and CADA-VAE [29]. Specifically, DeViSE and RelationNet both use linear projections to map visual features and semantic features into the same space. After the projection, DeViSE calculates the dot-product similarities between projected visual and semantic features. RelationNet utilizes a relation module to acquire the similarity between projected features. ReViSE uses a maximum mean discrepancy loss as cross-domain learning criteria to align the latent embeddings. JPoSE performs cross-modal fine-grained action retrieval between text and skeleton data. It learns PoS-aware embeddings and builds a separate multi-modal space for each PoS tag. CADA-VAE learns a latent space for both visual features and semantic embeddings via aligned variational autoencoders. Based on the above, the state-of-art method SynSE infuses latent skeleton visual representations with PoS syntactic information. We conducted an apple-to-apple comparison between our SMIE and such baseline methods.

4.3 Comparison with State-of-the-Art

Evaluation Settings. In zero-shot learning, the selection of distinct class splits gives rise to varied sets of seen and unseen classes, exerting a significant impact on the empirical outcomes. Moreover, the accuracy is also affected by the selection of the feature extractor. To facilitate a direct comparison with the state-of-the-art approach and fully demonstrate the effectiveness of our connection model, we employ identical experimental settings as SynSE. That means we use the same class splits, and pre-extracted visual and semantic features as supplied in their codebase. Specifically, SynSE maintains

Table 3: Ablation studies under optimized experimental setting on NTU-60, NTU-120, and PKU-MMD datasets.

Method Split	NTU-60(%)	NTU-120(%)	PKU-MMD(%)
	55/5	110/10	46/5
DeViSE	49.80	44.59	47.94
RelationNet	48.16	40.55	51.97
ReViSE	56.97	49.32	65.65
SMIE w/o \mathcal{L}_2	62.17	55.34	66.14
SMIE	63.57	56.37	67.15

two types of fixed class splits on NTU-60 and NTU-120 datasets. For the NTU-60 dataset, SynSE provides 55/5 and 48/12 splits, which include 5 and 12 unseen classes, respectively. Meanwhile, for the larger NTU-120 dataset with more categories, SynSE offers 110/10 and 96/24 splits. The visual feature extractor employed in the study is Shift-GCN [4], while the semantic feature extractor utilized is Sentence-Bert [28].

Results and Analysis. Tab. 2 presents the comparison results with baseline methods on NTU-60 and NTU-120. For the 55/5 and 48/12 split on NTU-60 datasets, SMIE outperforms SynSE 2.17% and 6.88% (relatively 2.86% and 20.66%), respectively. For the 110/10 split and 96/24 split on NTU-120, SMIE achieves 3.05% and 6.60% (relatively 4.87% and 17.05%), respectively. As the number of unseen classes increases, the difficulty of generalizing the learned knowledge from seen to unseen classes also increases. Notably, Our proposed SMIE method utilizes mutual information to capture global semantic information and effectively bridges the gap between visual and semantic space, which shows promising potential in enhancing the performance of unseen classes.

4.4 Ablation Study

Optimized Experimental Setting. From the experimental setting of SynSE, we find the objective of zero-shot learning experiments is to verify the effectiveness of the learned connection model. However, due to the significant impact of different class splits on the results, there can be a considerable deviation in accuracy even if the number of unseen classes is the same. Meanwhile, it is advisable to minimize the impact of the feature extractors with complex structures on the results, in order to focus on the effectiveness of the connection model itself. Thus, we provide an optimized experimental setting for zero-shot skeleton-based action recognition. First, we expand the dataset from two to three large-scale skeleton datasets, i.e., NTU-60, NTU-120, and PKU-MMD datasets, which increases the credibility of the results. Second, for each dataset, a three-fold test is applied to eliminate variance. Each fold has different groups of seen and unseen classes and the average results are reported. At last, we follow most skeleton-based self-supervised methods [19, 20, 35, 46] and apply the classical ST-GCN [43] as the visual feature extractor to minimize the impact of the feature extractors. The semantic feature extractor utilized in this study is Sentence-Bert, which is consistent with SynSE.

Overall Analysis on Optimized Experimental Setting. Under the optimized experimental setting, we aim to conduct an ablation study on the temporal constraint module of the SMIE. To provide

Table 4: Comparisons of different margin β in SMIE on NTU-60, NTU-120, and PKU-MMD datasets.

β	NTU-60 (%)	NTU-120 (%)	PKU-MMD (%)
0	62.17	55.34	66.14
0.01	62.98	55.93	67.15
0.1	63.57	55.78	66.77
0.5	63.29	56.37	66.29
1	63.12	54.98	65.99

more baseline results for subsequent research work, we reproduced two mapping methods (DeViSE and RelationNet) and one distribution alignment method (ReViSE) under this setting. Specifically, for the NTU-60 and PKU-MMD datasets, 5 unseen classes are selected randomly and the rest serves as the seen classes. The visual extractor only pre-trains on the seen classes. For the NTU-120 datasets, the number of unseen classes is 10. All the datasets get 3 groups of random splits and the average results are reported on Tab. 3. We found out that the projection-based methods obtain relatively lower accuracy. By aligning the distributions of the latent embeddings in the two domains, ReViSE achieves some improvements on the three datasets, which further confirms the importance of global information to the semantic features. However, ReViSE is much more complicated for the utilization of extra auto-decoders. Our SMIE achieves significant improvements on all datasets. Specifically, SMIE outperforms other projection methods by a margin of about 13.77%, 11.78%, and 15.18% on the three datasets. For ReViSE, our SMIE still achieves 6.60%, 7.05%, and 1.50% increments. By utilizing mutual information as the similarity metric, SMIE aligns the distributions of the two modalities and incorporates more discriminative information between different features in Eq. (3). For the ablation study of the temporal constraint module, "SMIE (w/o \mathcal{L}_2)" in the table indicates that the model with the global alignment module only, and "SMIE" is the full model. It is observed the temporal constraint module brings about 1.40%, 1.03%, and 1.01% performance on the three datasets, respectively. The results show the temporal constraint can help to integrate useful temporal information.

Influence of Hyper-parameters. To determine the best choice of the margin β in SMIE, we differ it based on our full model and conduct a test with all three splits on NTU-60, NTU-120, and PKU-MMD datasets. Tab. 4 shows the overall results. Note that, with decreasing margins, the impact of temporal constraints on the overall loss increases. It is found that as β increases, the performance first increases and then drops. The NTU-60, NTU-120, and PKU-MMD datasets achieve the best results at β values of 0.1, 0.5, and 0.01, respectively. The choice of margin keeps a balance between the global alignment module and the temporal constraint module. When the $\beta = 1$, the temporal constraint can not be fully used, which results in a performance drop.

Ablation Studies on Different Semantic Features. Under the optimized experimental setting, we also explored the influence of different semantic feature extractors on experimental results. As shown in Table 5, we use CLIP [27] as the semantic feature instead of Sentence-BERT [28]. Similarly, with different semantic features, the global align methods still achieve better performances than the direct mapping methods. Our proposed SMIE can also

Table 5: Results of CLIP semantic feature extractor on NTU-60, NTU-120, and PKU-MMD datasets.

Method	NTU-60(%)	NTU-120(%)	PKU-MMD(%)
Split	55/5	110/10	46/5
DeViSE	56.61	41.55	61.72
RelationNet	56.12	32.68	56.96
ReViSE	55.70	46.72	66.61
SMIE	61.11	45.74	71.50

Table 6: Results with the expanding category descriptions by ChatGPT on NTU-60, NTU-120, and PKU-MMD datasets.

Method	NTU-60 (%)	NTU-120 (%)	PKU-MMD (%)
SMIE	63.57	56.37	67.15
SMIE_Chat	70.21	58.85	69.26

outperform the baseline methods by a large margin on NTU-60 and PKU-MMD datasets. On the NTU-120 dataset with more data, our SMIE method achieves results comparable to the more complex ReViSE, which utilizes extra visual and textual auto-decoders. These results demonstrate that our SMIE method can achieve good performance on different semantic feature extractors, with a simple and efficient structure.

Expanding Category Descriptions using ChatGPT. Conventional approaches typically rely on taking the category label as input to a semantic feature extractor, to obtain the corresponding semantic feature. However, these labels contain only a few words and can not fully and accurately describe the corresponding action semantics. Based on it, we expand each action label name into a complete action description using ChatGPT and then extract its semantic feature. For example, "Wear jacket" can be expanded to "the act of putting on a garment designed to cover the upper body and arms". Following our optimized experimental setting, the results are shown in Tab. 6. Significant improvement in SMIE_Chat indicates that a more comprehensive description of action semantics leads to the improved representational capacity of semantic features, facilitating the connection model to capture the relationship between visual and semantic spaces.

Qualitative Results and Analysis. We visualize some predictive scores of the four unseen classes given by our approach in Fig. 4. As for the one-subject scenario shown in Fig. 4 (a, b, c), the actions of "clapping" and "cough" are very similar to each other. Our model not only makes the right predictions but also predicts plausible predictions for similar classes. For example, the class with the second highest score is "cough" when the ground truth is "clapping". A similar conclusion could be made in the two subjects' scenarios as shown in Fig. 4 (d, e), which proves the rationality of our method.

Visualization of the Confusion Matrices. Fig. 3 shows the confusion matrices of 3 random class splits on the NTU-120 dataset. Each matrix contains 10 unseen classes and the number in the matrix represents the classified samples. By reading the confusion matrices, the classification accuracy and misclassification of each class can be understood. We observe 1) short actions such as "jump up" or "yawn" are likely to be misclassified and the temporal constraint is less effective for them; 2) some actions such as "pick up"

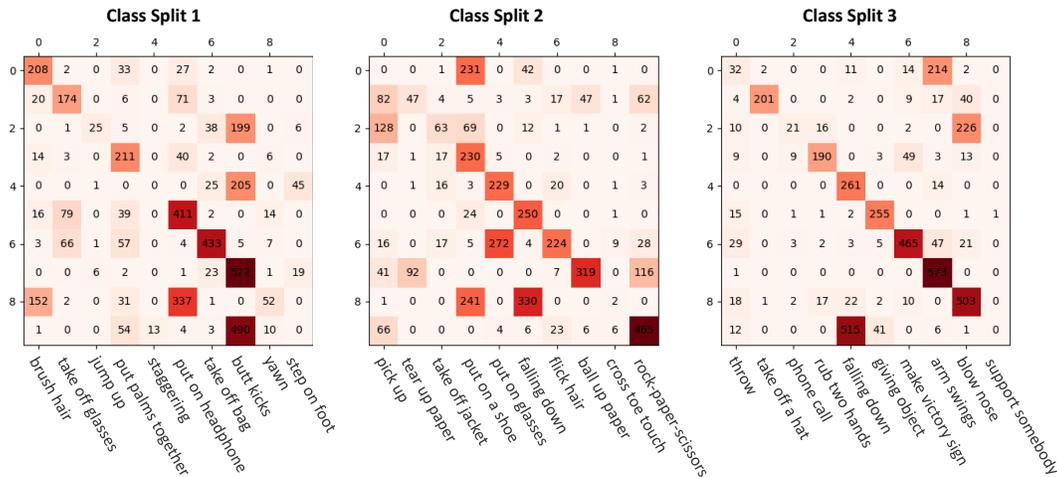


Figure 3: Confusion matrices for 3 randomly selected class splits on the NTU-120 dataset. The x-axis indicates the predicted class and the y-axis indicates the true class.

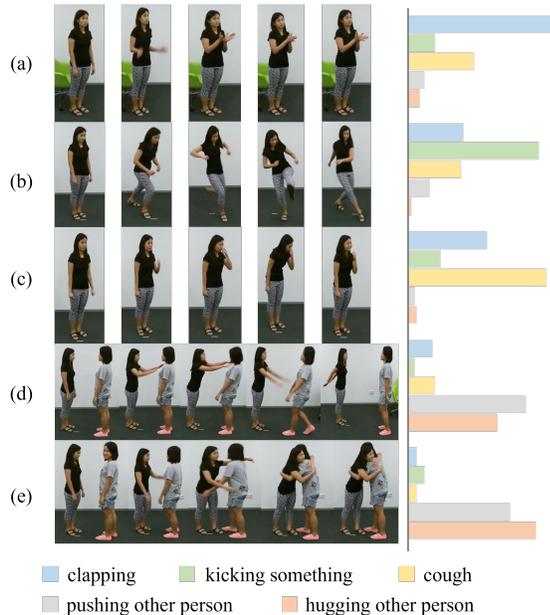


Figure 4: Visualization of the predicted results on NTU-60 dataset, where the estimated score of each class is represented by the corresponding bar length. Best viewed in color.

and "tear up paper" are hard to classify since they are mostly done with volunteers' hands. Thus, we will further investigate these two issues in our future work.

Explanation of Contrastive Learning in SMIE. The global alignment module uses contrastive pairs to estimate the mutual information between visual and semantic features[12, 26]. It can be motivated by aligning the two distributions of $p(v)$ and $p(a)$ to better connect the visual and semantic spaces. To this end, the joint distribution $p(v, a)$ and the multiplication of marginal distributions $p(v)p(a)$ should be as different as possible. Thus, we maximize the mutual information between v and a , which is the KL divergence between $p(v, a)$ and $p(v)p(a)$ as shown in Eq. 2.

Contrastive learning is used to maximize the approximated estimator of the KL divergence as in Eq. 4. We can obtain that $I(V, A) \geq \log(S) - L_S$ where

$$L_S = -E_V \left[\log \frac{f(v_i, a_i)}{\sum_{x_j \in V} f(v_j, a_j)} \right]. \quad (12)$$

Here $\{v_j, a_i\}$ ($j \neq i$) is a negative pair and S is the number of all the negative pairs. If we set $f = \exp\left(-\frac{\|v_i - c\|_2^2}{\tau}\right)$, then L_S is equal to the contrastive loss. It can be concluded that contrastive loss is a lower bound of mutual information.

For the contrastive learning in our experiment, (v, a) are paired visual and semantic features that serve as positive pairs while (v', a) are negative pairs.

5 CONCLUSION

In this work, we present a Skeleton-based Mutual Information Estimation and maximization framework (SMIE) for zero-shot action recognition, which consists of a global alignment module and a temporal constraint module. The global alignment module captures the complex statistical correlations between the visual space and the semantic space by applying mutual information as the similarity measure. The temporal constraint module exploits the inherent temporal information to keep the mutual information increasing with the number of observed frames. Extensive experiments on three skeleton benchmarks including NTU-60, NTU-120, and PKU-MMD datasets verify the effectiveness of our proposed SMIE model.

6 ACKNOWLEDGEMENTS

This work was supported by Shanghai AI Laboratory, the National Natural Science Foundation of China No. 61976206 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZ YJH012019100020098, Beijing Academy of Artificial Intelligence (BAAI), the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China 21XNLG05, and Public Computing Cloud, Renmin University of China.

REFERENCES

- [1] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. 2020. Rethinking Zero-Shot Video Classification: End-to-End Training for Realistic Applications. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.
- [3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [4] Ke Cheng, Yifan Zhang, Xiangyu He, Weihai Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-Based Action Recognition With Shift Graph Convolutional Network. *CVPR (2020)*, 180–189.
- [5] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical recurrent neural network for skeleton based action recognition. *IEEE Conference on Computer Vision and Pattern Recognition (2015)*.
- [6] Andrea Frome, S. Gregory Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. *NIPS (2013)*.
- [7] Chuang Gan, Ming Lin, Yi Yang, Yueting Zhuang, and G. Alexander Hauptmann. 2015. Exploring Semantic Inter-Class Relationships (SIR) for Zero-Shot Action Recognition. *AAAI (2015)*, 3769–3775.
- [8] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. 2022. Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-supervised Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 762–770.
- [9] Pranay Gupta, Divyanshu Sharma, and Kiran Ravi Sarvadevabhatla. 2021. Syntactically Guided Generative Embeddings for Zero-Shot Skeleton Action Recognition. *2021 IEEE International Conference on Image Processing (ICIP) (2021)*, 439–443.
- [10] Zongyan Han, Zhenyong Fu, and Jian Yang. 2020. Learning the redundancy-free features for generalized zero-shot object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12865–12874.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- [12] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670 (2018)*.
- [13] Guoliang Hua, Hong Liu, Wenhao Li, Qian Zhang, Runwei Ding, and Xin Xu. 2022. Weakly-supervised 3D Human Pose Estimation with Cross-view U-shaped Graph Convolutional Network. *IEEE Transactions on Multimedia (2022)*.
- [14] Bhavan Jasani and Afshaan Mazagonwalla. 2019. Skeleton based zero shot action recognition in joint pose-language semantic space. *arXiv preprint arXiv:1911.11344 (2019)*.
- [15] Qihong Ke, Mohammed Bannamoun, Senjian An, Ahmed Ferdous Sohel, and Farid Boussaid. 2017. A New Representation of Skeleton Sequences for 3D Action Recognition. *CVPR (2017)*.
- [16] Soo Tae Kim and Austin Reiter. 2017. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2017)*, 1623–1631.
- [17] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. 2015. Unsupervised Domain Adaptation for Zero-Shot Learning. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 2452–2460. <https://doi.org/10.1109/ICCV.2015.282>
- [18] Kai Li, Renqiang Martin Min, and Yun Fu. 2019. Rethinking Zero-Shot Learning - A Conditional Visual Classification Perspective. *ICCV (2019)*.
- [19] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3D Human Action Representation Learning via Cross-View Consistency Pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4741–4750.
- [20] Lilang Lin, Sijie Song, Wenhao Yang, and Jiaying Liu. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2490–2498.
- [21] Jun Liu, Amir Shahroudy, Lisboa Mauricio Perez, Gang Wang, Ling-Yu Duan, and Kot Alex Chichung. 2019. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)*.
- [22] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. 2020. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–24.
- [23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems (2013)*.
- [24] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. *NIPS (2016)*.
- [25] Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, and Hui Xiong. 2021. Robust local preserving and global aligning network for adversarial domain adaptation. *IEEE Transactions on Knowledge and Data Engineering (2021)*.
- [26] Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, and Hui Xiong. 2022. Interventional contrastive learning with meta semantic regularizer. In *International Conference on Machine Learning*. PMLR, 18018–18030.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [28] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP/IJCNLP (2019)*.
- [29] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. 2019. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. *CVPR (2019)*.
- [30] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *CVPR (2016)*.
- [31] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12026–12035.
- [32] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. 2011. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*. Ieee, 1297–1304.
- [33] Tristan Sylvain, Linda Petrini, and Devon Hjelm. 2020. Locality and Compositionality in Zero-Shot Learning. *ICLR (2020)*.
- [34] Chenwei Tang, Xue Yang, Jiancheng Lv, and Zhenan He. 2020. Zero-shot learning by mutual information estimation and maximization. *Knowledge-Based Systems (2020)*.
- [35] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. 2021. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th ACM international conference on multimedia*. 1655–1663.
- [36] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.
- [37] Hubert Yao-Hung Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning Robust Visual-Semantic Embeddings. *ICCV (2017)*.
- [38] Michael Tschannen, Josip Djolonga, K. Paul Rubenstein, Sylvain Gelly, and Mario Lucic. 2020. On Mutual Information Maximization for Representation Learning. *ICLR (2020)*.
- [39] Jiexin Wang, Yujie Zhou, Wenwen Qiang, Ying Ba, Bing Su, and Ji-Rong Wen. 2023. Spatio-Temporal Branching for Motion Prediction using Motion Increments. *arXiv preprint arXiv:2308.01097 (2023)*.
- [40] Qian Wang and Ke Chen. 2017. Zero-Shot Visual Recognition via Bidirectional Latent Embedding. *International Journal of Computer Vision (2017)*.
- [41] Michael Wray, Diane Larlus, Gabriela Csuska, and Dima Damen. 2019. Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings. *International Conference on Computer Vision (2019)*.
- [42] Xun Xu, M. Timothy Hospedales, and Shaogang Gong. 2016. Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation. *ECCV (2016)*, 343–359.
- [43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, Vol. 32.
- [44] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1112–1121.
- [45] Zhengyou Zhang. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia* 19, 2 (2012), 4–10.
- [46] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. 2023. Self-supervised Action Representation Learning from Partial Spatio-Temporal Skeleton Sequences. *arXiv preprint arXiv:2302.09018 (2023)*.
- [47] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. 2018. Towards Universal Representation for Unseen Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.