Transductive Zero-shot Action Recognition via Visually-connected Graph Convolutional Networks

Yangyang Xu, Chu Han, Jing Qin, Xuemiao Xu, Guoqiang Han, and Shengfeng He *Member, IEEE*

***Abstract*—With the explosive growth of action categories, zero-shot action recognition aims to extend a well-trained model to novel/unseen classes. To bridge the large knowledge gap between seen and unseen classes, in this paper, we visually associate unseen actions with seen categories in a visually-connected graph, and the knowledge is then transferred from the visual features space to semantic space via the Grouped Attention Graph Convolutional Networks (GAGCN). In particular, we extract visual features for all the actions, and a visually-connected graph is built to attach seen actions to visually similar unseen categories. Moreover, the proposed grouped attention mechanism exploits the hierarchical knowledge in the graph, so that the GAGCN enables propagating the visual-semantic connections from seen actions to unseen ones. We extensively evaluate the proposed method on three datasets, *i.e.*, HMDB51, UCF101, and NTU RGB+D. Experimental results show that the GAGCN outperforms state-of-the-art methods.**

***Index Terms*—zero-shot learning, graph convolutional network, action recognition**

I. Introduction

Human action recognition has been extensively explored due to its wide range of applications, *e.g.*, video surveillance, human-computer interaction, and robotics [37], [23], [5], [42], [21]. However, with the increasing demand for different applications and the explosive growth of action categories, the huge workload of manual labeling action data is unavoidable. Therefore, extending a well-trained model to novel/unseen classes is always challenging yet highly desired.

Zero-shot learning (ZSL) has been studied to overcome such a restriction [19], [37], [53], [44]. It aims to recognize the novel category by transferring the knowledge obtained from the seen classes to model the unseen ones. As a result, the core principle of ZSL is to find a good connection between seen and novel classes for an accurate knowledge transfer.

Previous works transfer the knowledge by connecting classes with various attributes [6], [34]. However, attributes are not the best way to describe an action, as an action contains both spatial and temporal information.

Yangyang Xu, Xuemiao Xu, Guoqiang Han and Shengfeng He are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Xuemiao Xu is also with State Key Laboratory of Subtropical Building Science, Ministry of Education Key Laboratory of Big Data and Intelligent Robot and Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information. E-mail: cnnlstm@gmail.com, xuemx@scut.edu.cn, csgqhan@scut.edu.cn, hesfe@scut.edu.cn.

Chu Han is with the Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China. E-mail: zq1992@gmail.com.

Jing Qin is with the Department of Nursing, Hong Kong Polytechnic University. E-mail: harry.qin@polyu.edu.hk.

Corresponding author: Xuemiao Xu and Shengfeng He.

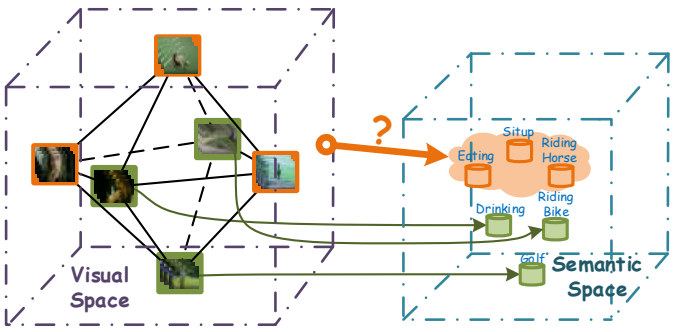


Fig. 1. The green color elements belong to the seen classes and the oranges belong to the unseen classes. We build the graph by linking all categories in visual space and learn a projection model between the visual space and semantic space. The semantic information of seen classes provides the supervision, and the projection model can predict the semantic information of the unseen classes.

An alternative solution is to transfer the knowledge from the seen classes to the unseen ones by adopting the semantic/textual representations as the auxiliary information [3], [49], [17]. Specifically, they connect seen and unseen classes by utilizing the relations hidden in the semantic/textual representations. However, the descriptions of action classes (usually a phrase or a sentence) contain far less information than the action sequences, this imbalance knowledge between two spaces makes the connections between seen and unseen classes inaccurate.

We observe that human can easily identify similar actions at a glance, and this process solely depends on the intrinsic visual similarity between actions. Comparing with attribute or textual information, spatial and temporal information contains the richest and the most accurate action representations. For example, actions "jump rope" and "rope climbing" may be close to each other in the semantic space, but they show totally different action movements.

As a consequence, we aim to connect seen and unseen categories with their intrinsic visual similarity. To this end, we propose a visually-connected graph convolutional network for transductive zero-shot action recognition. Transductive ZSL means the unseen data is available in the training phase [41], [52]. Our approach aims to transfer the intrinsic relationships in the visual space between the seen and unseen classes to their semantic space, which can be seen in Figure I. In particular, we extract the visual features on both the seen and the unseen actions. Then, a visually-connected graph is built by considering the distances among different categories

TABLE I
ADVANTAGES AND DISADVANTAGES OF THE RELATED ZERO SHOT ACTION RECOGNITION METHODS.

	Methods	RSC [17]	SER [47]	PDA [48]	MR [49]	TOM [20]	ZSECOC [33]	VDS [53]	BiDiLEL [43]	**Ours**
Advantages	Utilizes visual relationships	×	×	×	√	×	×	√	√	√
	Utilizes semantic relationships	×	×	×	×	×	×	√	√	×
	Domain adaptation	√	×	√	×	√	×	×	×	×
	Discriminative semantic representations	×	√	×	√	×	√	×	×	√
Disadvantages	Ignores visual relationships	√	√	√	×	√	√	×	×	×
	Ignores semantic relationships	√	√	√	√	√	√	×	×	√
	Requires auxiliary datasets or knowledge	×	√	√	√	×	√	×	×	×
	High complexity and computational costs	×	×	×	√	×	×	√	√	×

based on the discriminative features. We propose a grouped attention graph convolutional network to propagate the visual-semantic correlations of the seen actions to the unseen ones by exploring the hierarchical structure on the graph. As a result, unseen actions can be correctly recognized by matching the corresponding label vectors. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art approaches on three standard benchmarks.

The contributions of our paper are summarized as follows:

- We propose a visually-connected graph for transductive zero-shot action recognition. It learns the correlations between the visual feature space and the semantic space, which can correctly propagate the seen classes knowledge to unseen ones.
- We present a grouped attention graph convolution network. It exploits the nodes in a hierarchical manner for the convolutional operation, leading to an accurate learning of visual-semantic mapping.
- The proposed model outperforms state-of-the-art approaches on three benchmarks, as well as different baselines.

II. RELATED WORK

A. General Action Recognition

Action recognition [22], [26] has been studied for years. This problem has already gained great progress since deep neural network showed its remarkable representability. In addition, the explosive growth of data and the increasing computational power push fully supervised action recognition to the peak. Simonyan *et al*. [37] propose a two-stream CNN for action recognition by incorporating spatial and temporal information from video. Inspired by the two-stream model, Wang *et al*. [42] present a Temporal Segmental Network (TSN) to understand the long-range temporal information in a video. Hara *et al*. [9] develop an effective approach for spatio-temporal features learning using deep 3-dimensional convolutional network. Recently, Temporal Relation Network (TRN) is proposed by Zhou *et al*. [55] to learn temporal dependencies in video frames at multiple time scales. Notwithstanding the demonstrated success of the existing action recognition methods, they are restricted by the demand of huge labeled data.

B. Zero-shot Learning

Zero-shot learning [24], [17], [52], [54], [10] aims to transfer the knowledge from the seen classes to the unseen ones. It has been drawn attention due to the explosive growth of unlabeled data and demand of classification for the unseen classes. There are two settings in ZSL, the conventional setting takes only the unseen videos as the test data, while the generalized ZSL setting takes both seen and unseen videos for testing.

For the generalized ZSL, the instances come from all classes are formed as the test set. Previous zero-shot learning methods mainly focus on still images. There are many approaches [2], [4], [7], [52], [13] in ZSL rely on various attributes to represent categories. These attributes are regarded as the side-information for learning an embedding vector for ZSL. Recent approaches use deep neural networks to map the visual space to the semantic space. Ba *et al*. [19] present a neural network to classify the unseen classes from their textual description. Recently, Wang *et al*. [45] propose a model to distill information via semantic embedding and knowledge graph using graph convolutional networks. Kampffmeyer *et al*. [14] further explore the knowledge graph and achieve remarkable performance on the ImageNet dataset. Both of them predict the classifiers of the unseen classes. Recently, Gao *et al*. [8] extend [45] and add another branch for predicting attribute-features of objects. There are also several methods utilize generative models for ZSL [53], [28], [46], [35], [12]. The basic idea is to use different conditional Generative Adversarial Networks (GAN) [27] for generating visual features of the unseen classes conditioned on the semantic representations. M^2GAN [12] fuses various types of semantic representations by a feature fusion network to generate pseudo visual features for alleviating the "heterogeneity gap". AgNet [13] aligns different modal data into the semantic space, which bridges the gap introduced by modality heterogeneity and ZSL.

There are only a few works in the field of zero-shot action recognition. Some of them utilize semantic or visual relations among all classes in the mapping process. Bidirectional latent embedding learning (BiDiLEL) [43] maps the visual features and semantic representations to a shared latent space while preserving the class relations. Visual Data Synthesis (VDS) [53] generate visual features by a GAN while using both semantic knowledge and visual distribution to build the connections. However, these two models involve complex optimization processes leading to large computational costs. On the other hand, domain adaptation is also introduced for aiding the mapping process, such as Regularised Sparse Coding (RSC) [17], Prioritized Data Augmentation (PDA) [48], and Two output Model (TOM) [20]. They apply the learned mapping model from the seen classes to the unseen classes

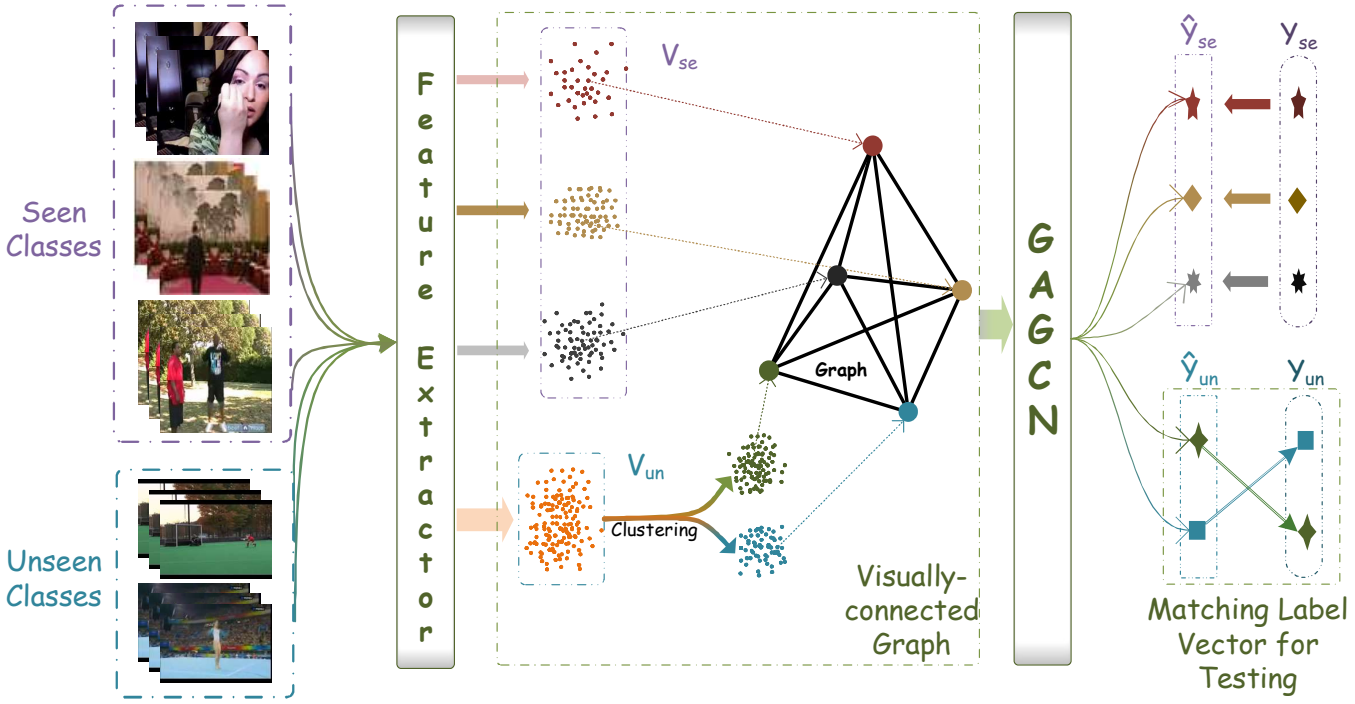


Fig. 2. Overview of our framework. It consists of four components: a feature extractor, a visually-connected graph, a grouped attention graph convolutional network, and a label vector matching for testing. The feature extractor produces visual features V_{se} and V_{un} as the input of the graph. For those seen classes, we simply calculate the center of each cluster of the class. For the unseen classes, we apply a k-means clustering. Then we build the graph by considering the distances among the action clusters. Our grouped attention graph convolutional network takes the center features of each cluster as the input and then predicts its label vector. In the testing phase, we match the predicted label vectors of the unseen classes for recognition.

with domain adaptation. But they ignore the intrinsic relations among different classes in two spaces. Some other works focus on proposing a new semantic space for extracting discriminative representations, such as word-vector or error-correcting output codes [33], [49], [47]. Like them, we use the discriminative word-vector as the semantic space. Unlike them, we explore the relations that hidden in the visual space by a novel graph convolutional network. We summarize the advantages and disadvantages of the related works in Table I.

III. Approach

In this section, we first formalize the problem of zero-shot learning, and then introduce four components of our zero shot learning framework. Overview of our framework is illustrated in Figure 2.

A. Problem Formalization

Zero shot learning aims to recognize the label of each unseen sample by transferring the knowledge obtained from the seen classes $S_{\text{se}} = \{(x_i, y_i)|i = 1, ..., I\}$, where x_i denotes the i-th seen video and $y_i \in Y_{\text{se}}$ is the corresponding label. We also denote Y_{un} as the set of unseen classes, and in total there are N categories in Y_{se} and M in Y_{un}. Note that in zero shot learning setting, Y_{se} and Y_{un} are disjoint, *i.e.*$Y_{\text{se}} \cap Y_{\text{un}} = \emptyset$.

B. Features Extractor

As we mentioned above, visual features show strong correlation to an action. In this paper, we aim to learn the correlations between the visual feature space and the semantic space. Thus, we first extract visual features of an action by a pre-trained network. Given a video sample x_i, the visual feature of this action is extracted by the features extractor $F(\cdot)$ and outputs a visual features vector v_i as follows:

$$v_i = F(x_i) \tag{1}$$

We train the network using the video samples of the seen classes. Then, we use this pre-trained network to extract feature vectors for both the seen and unseen data.

C. Visually-connected Graph

Before we introduce the proposed visually-connected graph, we first briefly review previous works that based on a semantically-connected graph. They transfer the knowledge by adopting the semantic/textual representations as the side information. In this way, they perform zero-shot classification by using the word embedding of the class labels and the semantic similarity to predict the classifier for each unseen class [44], [14]. We set this approach as the baseline in our experiment.

Comparing to semantic features, visual features contain the richest and the most accurate action representations. We build the visually-connected graph by connecting seen and unseen categories with their intrinsic visual similarity. Firstly, we group the feature vectors extracted from the same action class into a cluster. In the proposed graph, we define the mean vectors of a cluster as the graph node. Since the correlations between the labels and the visual features in seen classes are known, the cluster center of seen class n can be simply

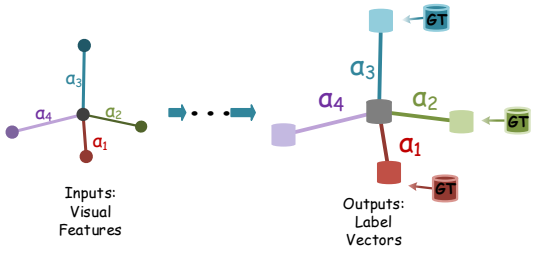


Fig. 3. A toy example of our graph-based inference model, each node is connected to 4 neighboring nodes in this example for simplicity. Each node represents a category. α_i represents the edge attention weight. The output is the label vector of one category. ***GT*** is the Ground Truth of the seen classes' label vectors, which provides the supervision information for training.

calculated the mean feature vectors with the same class as follows:

$$c_n = \frac{1}{w}\sum_{i=1}^{w} v_i^n, \tag{2}$$

where w is the number of feature vectors in class n. Then the cluster centers of all seen classes are defined as follows:

$$C_{\text{se}} = \{c_{\text{se}}^n | n = 1, ..., N\}. \tag{3}$$

Due to the unknown correlations between the unseen classes and their visual features, we initially apply the K-Means clustering to the feature vectors to get the clusters of unseen classes. Then the feature vectors V_{un} can be divided into M clusters, where M is the number of unseen classes. Here, we assume that all the features from the same cluster share the same label. Then we get the cluster centers of all unseen classes defined as follows:

$$C_{\text{un}} = \{c_{\text{un}}^m | m = 1, ..., M\}. \tag{4}$$

Now we obtain N cluster centers in C_{se} and M centers in C_{un}. These cluster centers are used to compose the nodes of the graph, as shown in Figure 2. For each node in the graph, we seek its k nearest neighbors in $C_{\text{un}} \cup C_{\text{se}}$. After that, a dense graph G of the visual features is conducted. Then the graph G can be simply represented by an adjacency matrix $A \in \mathbb{R}^{(N+M)\times(N+M)}$, where $\mathbb{R}$ indicates the set of real numbers. With this graph, All categories in seen and unseen classes are linked with k neighbors.

D. Grouped Attention Graph Convolution Network

After building the graph, we transfer the knowledge from the visual feature space to semantic space via a Graph Convolutional Network (GCN). The original GCN was proposed for performing semi-supervised entity classification [16] and the original GCN can be presented as:

$$\hat{Y}_{ori} = \sigma(\widetilde{A}XW), \tag{5}$$

where $\hat{Y}_{ori}$ denotes the output of the original GCN, $\sigma(\cdot)$ denotes the non-linear activation function. $\widetilde{A}$ is the normalized version of A. W is the trainable weight, and X denotes the input.

The adjacency matrix A in the original GCN encodes the connections among all nodes, but ignores their discriminative distance relations. All edges are weighed equally in the original GCN. However, for a host node, the contributions of its neighbors should be different and dynamically determined during training. In order to learn the dynamic weights of neighboring nodes, we propose a Grouped Attention Graph Convolutional Network (GAGCN) for predicting the label vectors of unseen classes and the graph-based inference model is illustrated in Figure 3.

As we utilize *K-NN* algorithm to form the graph, in the visual feature space, a larger distance between two nodes explicitly indicates less similarity and less contribution. In GAGCN, we divided all edges in the graph into K groups (the same as the K in *K-NN* algorithm, but not the same k value in K-Means clustering above), the K-th group contains the K-th nearest neighbor of each node to form the adjacency sub-matrix A_{K}, and edges of the same group share the same attention weight. In this way, a group of global weight is assigned according to the local similarity conditions. A simplified grouped graph is shown in Figure 4. Hence, we replace the adjacency matrix A by K adjacency sub-matrices $\{A_1, A_2, ..., A_{\text{K}}\}$ for getting a discriminative connection graph. Algorithm 1 shows the pseudo-code for producing the K adjacency sub-matrices.

We define the learnable attention weight of group k as g_k and initialize g_k with Gaussian distribution. Then we normalize the group attention weights using a $softmax$ function $\alpha_k = \frac{\exp(g_k)}{\sum_{k=1}^{K}\exp(g_k)}$. The proposed GAGCN can be represented as follows:

$$\hat{Y} = \sigma(\sum_{k=1}^{K} \widetilde{A}_k X W_k \alpha_k), \tag{6}$$

where $\hat{Y}$ is the output of our GAGCN, W_k is the trainable weight of group k. $\widetilde{A_k}$ is the normalized version of A_k. Our GAGCN performs convolutions on the adjacency matrix A_k. These graph convolutional operations can be stacked one by one to form a network.

Our GAGCN takes the center features of categories as input $X = C_{\text{se}} \cup C_{\text{un}}$ and $X \in \mathbb{R}^{(N+M)\times C}$, here C is the dimension of a center feature. The sub-matrix A_k has the same dimension with the adjacency matrix A, which can be presented as $A_k \in \mathbb{R}^{(N+M)\times(N+M)}$. Once the network is properly trained, the grouped weights are adaptively learned for handling nodes with different levels of similarities.

E. Training and Testing

Our GAGCN predicts the label vector for each class, and the label vectors of the seen classes are provided as the ground truth for training. We use the mean-square error as the loss function, which can be computed as:

$$\mathbb{L} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i^{seen} - y_i^{seen})^2, \tag{7}$$

where $\hat{y}_i^{seen}$ is the predicted label vector of seen classes and $\hat{y}_i^{seen} \in \hat{Y}$, $y_{i,j}$ is the corresponding ground truth of label vector. N is the number of the seen classes.

In the testing period, our GAGCN predicts the label vectors of the unseen clusters $\hat{Y}_{\text{un}}$. For each predicted vector

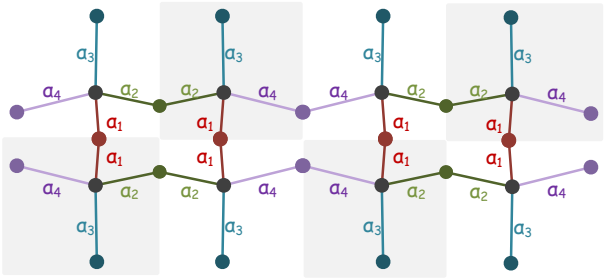


Fig. 4. Simplified example of our grouped graph. Each grid contains a host node and its neighboring nodes and edges. Edges of the same colors are grouped together and share the same weight.

Algorithm 1: Producing process for K adjacency sub-matrices

Input: Mean vector set $C_{se} \cup C_{un}$ for all classes and the number of nearest neighbors K;
Output: K adjacency sub-matrices;

```
for k = 1; k ≤ K; k++ do
    for Each vector c^n in C_se ∪ C_un do
        Seek the k_th nearest neighbor c_k^n in the space
          of C_se ∪ C_un − c_n by calculating the cosine
          similarity;
        Build a connection between c_n and c_k^n and add
          it into adjacency sub-matrix A_k.
    end
end
return {A_1, A_2, ..., A_K}.
```

$\hat{y}_{\mathrm{m}} \in \hat{Y}_{\mathrm{un}}$, we seek its nearest neighbor in the space of ground truth label vectors (includes ground truth label vectors of unseen classes Y_{un}) by calculating their cosine similarity. The matched vector corresponds to the predicted label, which can be presented as:

$$y_{\mathrm{m}} = \arg\max_{\mathrm{y_{un}}} \cos\left\langle \hat{y}_{\mathrm{m}}, y_{\mathrm{un}} \right\rangle, y_{\mathrm{un}} \in Y_{\mathrm{un}}. \tag{8}$$

In this way, every cluster (unseen class) is assigned a label with the highest label vectors similarity. As we discussed above, we assume that all samples from the same cluster share the same label, now we assign the label of each sample as its cluster label.

IV. Experiments

A. Datasets and Settings

We evaluate the proposed method on three benchmark datasets. (1) **HMDB51 dataset** [18] consists of 6,766 videos mostly from movies with 51 categories. These videos are RGB videos, so we use RGB frame modality only in our experiment. (2) **UCF101 dataset** [39] is collected from YouTube with 101 action categories, containing 13,320 video samples and 27 hours of video data in total. Same as the HMDB51 dataset, we only report the experiment result of RGB information. (3) **NTU RGB+D dataset** [36] is currently the largest dataset with 3D joints annotations for action recognition. This dataset contains 56,680 action samples in 60 action categories. This dataset was collected by Microsoft Kinect camera in the indoor environment, and it provides RGB videos, depth map sequences, 3D skeletal data, and infrared videos for each action sample. Since this dataset was collected from indoor with similar backgrounds for different categories, 3D skeletal data contains more discriminative information than RGB videos. In our experiment, we examine the proposed model with RGB frame modality, skeleton modality, and their fusion. To the best of our knowledge, we present the first attempt to perform zero-shot action recognition on this dataset.

We follow the setting of [49], which splits each dataset into seen and unseen classes evenly, *i.e.*30/30, 51/50, and 26/25, with regards to NTU RGB+D, UCF101 and HMDB51 datasets, respectively. For generalized ZSL, we follow the work of Mishra *et al*. [29] to split 20% data of seen classes for testing and the remains for training. Meanwhile, since our method needs to obtain cluster centers by applying clustering algorithm on the unseen data, we follow the work [38] that assumes whether the test instances belong to the seen or the unseen classes is known in advance. We generate 10 splits randomly for each dataset, and the average accuracy and standard deviation are reported.

B. Implementation Details

We use two types of features: RGB features and skeleton features. For the RGB modality, the visual features are extracted by the last convolution layer of Temporal Segmental Network (TSN) [42] with dimension of 1024. For the skeleton modality, we utilize Spatial Temporal Graph Convolutional Network (ST-GCN) proposed by Yan *et al*. [50]. The visual features of skeleton modality are produced by the last convolution layer of ST-GCN. Here, both TSN and ST-GCN are trained on the seen classes with default settings from the scratch. For getting the same dimension with RGB visual features, we modify the kernel numbers of the last convolution layer of original ST-GCN. We also report the fusion results of two modalities, we fuse two heterogeneous visual features to a new feature and take it as the input of GAGCN, our fusion methods including mean fusion, max fusion and concatenate fusion [5]. To be specific, mean fusion computes the mean value of two features at the same spatial location and channel. Max fusion takes the maximum of two features. Concatenate fusion stacks two feature maps at the same spatial location across channels. For the semantic representation, we use GloVe text model [32] trained on the Wikipedia dataset to encode each category into a 300-dimensions label vector. For the multi-word category, we accumulate the vector of each unique word.

When we utilized K-Means clustering to the feature vectors to get the clusters of unseen classes, the cluster centers are initialized randomly and always converge to the similar result experimentally. Our GAGCN consists of 2 layers with 2048 hidden units, and rectified linear units ($ReLU(\cdot)$) is used as the activation function. L_2 normalization is performed on the outputs of the network and the ground truth label vectors for

TABLE II
COMPARISONS TO THE STATE-OF-THE-ART ZSL ACTION RECOGNITION METHODS ON UCF101 AND HMDB51 DATASETS. AVERAGE ACCURACY AND STANDARD DEVIATION ARE REPORTED. **VISUAL** AND **SEMANTIC** DENOTE THE VISUAL FEATURES AND SEMANTIC REPRESENTATIONS RESPECTIVELY. MEANWHILE, **D** REPRESENTS THE DEEP VISUAL FEATURES AND **L** REPRESENTS THE LOW-LEVEL FEATURES. **WV** IS THE WORLD VECTOR, **ATT** IS THE ATTRIBUTES AND **ECOC** IS THE ERROR-CORRECTING OUTPUT CODES. RANDOM GUESS IS THE LOWER BOUND FOR EACH DATASET.

Methods	Visual	Semantic	HMDB51	UCF101
Random Guess	-	-	4.0	2.0
Baseline	D	WV	14.7±2.3	13.7±1.2
RSC [17]	L	WV+Att	-	14.0±1.8
SER [47]	L	WV	21.2±3.0	18.6±2.2
MR [49]	L	WV	24.1±3.8	22.1±2.5
PDA [48]	L	WV	24.8±2.2	22.9±3.3
TOM [20]	D	WV	-	26.8±4.4
ZSECOC [33]	L	ECOC	22.6±1.2	15.1±1.7
VDS [53]	D	WV	25.3±4.5	28.8±5.7
BiDiLEL [43]	D+L	WV	22.3±1.1	23.0±0.9
Ours	D	WV	**29.8±2.2**	**30.0±1.8**

TABLE III
COMPARISONS TO THE STATE-OF-THE-ART GENERALIZED ZSL ACTION RECOGNITION METHODS ON UCF101 AND HMDB51 DATASETS. AVERAGE ACCURACY AND STANDARD DEVIATION ARE REPORTED.

Methods	HMDB51	UCF101
SJE [1]	10.5±2.4	8.9±2.2
ConSE [30]	15.4±2.8	12.7±2.2
GA [29]	20.1±2.1	17.5±2.2
Objects2Action [11]	-	30.3
Ours	**32.5±2.5**	**35.6±2.1**

both training and testing. We train our GAGCN for 50000 epochs using Adam optimizer [15] with the learning rate of 0.001 and weight decay of 0.0005, and the experiments is performed with PyTorch using an Nvidia 1080Ti GPU. The source code is available at this link.

C. Compared Methods

For HMDB51 and UCF101 dataset, we compare our model with various zero-shot action recognition methods existing in Table I. For NTU RGB+D dataset, there are no zero-shot learning results have been reported so far. This dataset provides two standard splits for fully supervised classification, which includes cross-subject split and cross-view split [36], we compare with the fully supervised methods under different modalities in the cross-view split. We set the semantically-connected graph with the original GCN when $k = 10$ as the baseline on three datasets in our experimental, and other implementation details are set the same as the visually-connected graph.

The experimental results on UCF101 and HMDB51 datasets with conventional and generalized settings are shown in Table II and Table III respectively. From the tables we can have the following observations:

(1) All approaches significantly outperform the random guess, shows the effectiveness of zero-shot learning in action recognition.

TABLE IV
COMPARISONS TO FULLY SUPERVISED METHODS ON NTU RGB+D DATASET. AVERAGE ACCURACY AND STANDARD DEVIATION ARE REPORTED. RANDOM GUESS IS THE LOWER BOUND UNDER THE SETTING OF ZERO-SHOT LEARNING. **ZSL?**: ZERO-SHOT LEARNING MODEL OR NOT. **MODALITY**: THE USED MODALITY, INCLUDING **RGB**, **DEPTH**, AND SKELETON (**SKE**) VIDEOS.

Methods	ZSL?	Modality	Accuracy
Random Guess	√	-	3.3
Baseline	√	RGB+Ske	24.3±2.3
HON4D [31]	×	Depth	7.3
SNV [51]	×	Depth	13.6
Lie Group [40]	×	Ske	**52.8**
Ours	√	RGB+Ske	28.5±2.8

(2) Generally, the deep visual feature based methods outperform the low-level visual features, which indicates that the choice of visual representations is of vital importance for ZSL.

(3) Our approach beats the baseline method which links the categories by their semantic similarities, which shows the effectiveness of our visually-connected graph and the grouped attention method.

(4) Our approach outperforms state-of-the-art methods by a large margin on both UCF101 and HMDB51 datasets with different settings, which demonstrates the effectiveness of the proposed method.

The experimental results on NTU RGB+D dataset are shown in Table IV. We can see that the proposed zero-shot learning method is not comparable to the latest fully supervised approaches. It is apparent that zero-shot learning is a much harder task than fully supervised action recognition. On the other hand, the proposed approach still outperforms the two fully supervised methods HON4D [31] and SNV [51], even though we under the restrict zero-shot learning setting.

D. Ablation Studies

We conduct the ablation studies to further evaluate the effectiveness of different components in our framework. We report the performances of the proposed visually-connected graph, the grouped attention mechanism, and the parameters of k in the *K-NN* algorithm for building the graph (k edges for each node). For the sake of simplicity, we only report the results on HMDB51 and UCF101 datasets. From the results shown in Table V, we can draw the following observations:

(1) The visually-connected graph outperforms the semantically-connected graph both with the original GCN and GAGCN with respect to different k values, which shows that the proposed visually-connected graph models the relationship between seen and unseen classes much more accurate than the semantically-connected graph does. This is because our visually-connected graph is more accurate for representing the relations among different action classes. More qualitative results will be presented in Section IV-E.

(2) GAGCN outperforms the original GCN, which demonstrates the effectiveness of the attention mechanism in GAGCN. This is mainly because the attention mechanism in GAGCN can strengthen the weights of the nearest neighbors

TABLE V
EXPERIMENTAL RESULTS WITH THE DIFFERENT VALUES OF k FOR THE VISUALLY-CONNECTED AND SEMANTIC-CONNECTED GRAPH WITH ORIGINAL GCN AND GAGCN, AVERAGE % ACCURACY AND STANDARD DEVIATION FOR HMDB51 AND UCF101 DATASET.

K	HMDB51				UCF101			
	Semantically-connected		Visually-connected		Semantically-connected		Visually-connected	
	OriGCN	GAGCN	OriGCN	GAGCN	OriGCN	GAGCN	OriGCN	GAGCN
5	17.4±2.9	23.2±4.5	25.6±2.8	29.7±2.3	13.3±1.7	14.6±1.8	26.9±1.7	29.4±2.8
10	14.7±2.3	17.0±3.2	23.4±1.8	**29.8±2.2**	13.7±1.2	14.8±6.5	24.3±1.8	**30.0±1.8**
15	17.2±3.5	25.3±3.1	22.7±2.4	28.6±1.8	12.2±0.9	12.9±0.5	22.8±1.6	29.7±1.9

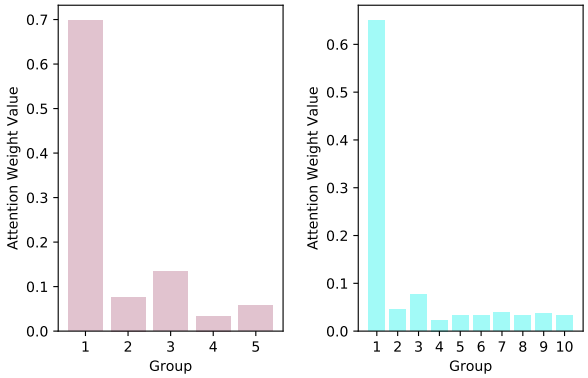


Fig. 5. Attention weights of the trained GAGCN when $k = 5$ (left) and $k = 10$ (right) on the UCF101 dataset.

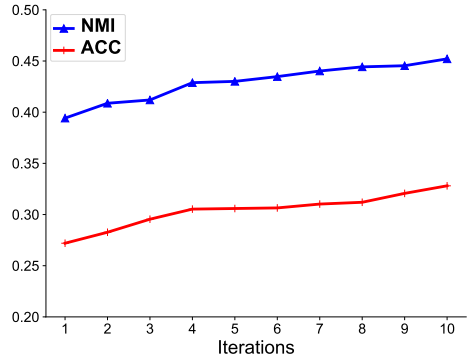


Fig. 6. NMI scores ans ZSL accuracies with respect to different iterations in K-Means algorithm on a split of HMDB51 dataset.

while weakens the negative impact from the unnecessary edges.

(3) As the value of k increases, the performance of the original GCN gets worse, but GAGCN can keep a steady performance. This reveals that the attention mechanism in our GAGCN can reduce the influence of unnecessary edges. Also, the best value of k is 10 for both datasets, and we set $k = 10$ in all the following experiments.

We also analyze the attention weight of the trained GAGCN with respect to $k = 5$ and $k = 10$, on the UCF101 dataset, As shown in Figure 5, we can see that when the attention weight of group 1 is significantly larger than the other groups. This indicates that the knowledge mainly propagated among the nearest neighbors. Meanwhile, as the k increases, the latter groups are assigned with little attention weights. We believe this is the key reason for GAGCN can keep a steady performance with a large k, while the original GCN weights all the edges equally.

We further analyze the effectiveness on the purity of the cluster to the final ZSL performance. The purity of the clustering result is measured by the Normalized mutual information (NMI) scores, and NMI score is defined as:

$$NMI(T, P) = \frac{I(T, P)}{\sqrt{E(T)E(P)}}, \tag{9}$$

where $I(\cdot)$ is the mutual information and $E(\cdot)$ denotes the entropy, T is the truth label result and P is the clustering result. NMI ranges from $[0, 1]$ and a larger NMI score indicates a purer clustering result. In Figure 6, we report the experimental result (accuracy and NMI) with a different number of iterations in the K-Means algorithm on a split of HMDB51 dataset. We can see that as with more iterations, K-Means clustering obtains an increased NMI score indicating a more purity result. On the other hand, unsurprisingly, the accuracy has a positive correlation with the purity, which implies that a better cluttering result can achieve a better recognition performance.

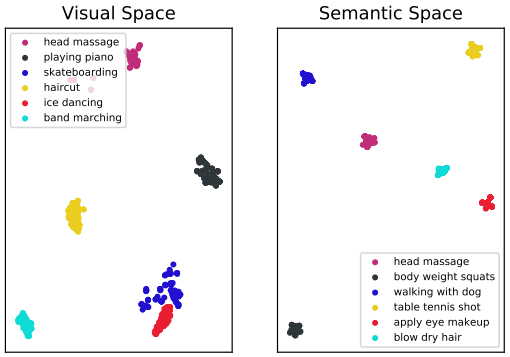
TABLE VI
EVALUATION WITH RESPECT TO DIFFERENT MODALITIES AND THEIR FUSION RESULTS ON NTU RGB+D DATASET. MODALITIES: **RGB** VIDEO AND SKELETON (**SKE**) VIDEO FOR NTU RGB+D DATASET. FUSION METHODS INCLUDE MAX FUSION (**MAX**), MEAN FUSION (**MEAN**), AND CONCATENATE FUSION (**CONCAT**). THE RESULTS OF ORIGINAL GCN (**ORIGCN**), **GAGCN** ARE ALSO REPORTED.

Modalites	OriGCN	GAGCN
RGB	23.4±1.6	27.6±2.6
Ske	24.2±4.1	28.1±3.3
Max	24.3±2.9	27.8±3.1
Mean	24.3±2.3	**28.5±2.8**
Concat	24.6±2.3	27.9±2.6

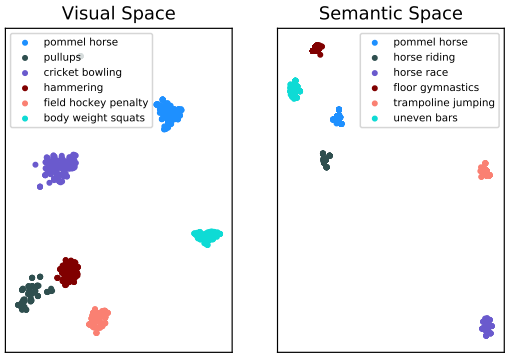
We also conduct an experiment shown in Table VI for the comparison of single modality input, RGB, and skeleton features respectively. We can observe that the skeleton modality demonstrates a better result than RGB modality. In the meanwhile, we also compare the performance of three different fusion methods, *i.e.*, mean fusion, max fusion, and concatenate fusion. Mean fusion achieves the best result among three fusion methods.

E. Analysis on Visually-connected Graphs

In this subsection, we aim to explore the differences between the original semantically-connected graph and our visually-connected graph by visualizing action samples in different spaces. We provide some actions and their neighbors both in the visual space and the semantic space, their t-SNE [25] visualizations can be seen in Figure 7. For action "head massage" in Figure 7(a), its top-5 nearest neighbors



(a) Action **"Head Massage"** and its Top-5 Neighbors in Two Spaces



(b) Action **"Pommel Horse"** and its Top-5 Neighbors in Two Spaces

Fig. 7. t-SNE visualizations for two actions and their neighbors in different space. Each colored cluster represents an action category in visual and semantic spaces.

in the visual space are "playing piano", "skateboarding", "haircut", "ice dancing", and "band marching". The hand movement information is involved both in "head massage" and "playing piano", our visually-connected graph takes "playing piano" as its nearest neighbors. Meanwhile, "haircut" is highly related with the "head", hence, we take it as the third nearest neighbor of "head massage". There are also some negative neighbors are selected, *i.e.*, "skateboarding" and "band marching", but the group attention mechanism in GAGCN can reduce its contribution dynamically. On the contrary, the top-5 neighbors in the semantic space are "body weight squats", "walking with dog", "table tennis shot", "apply eye makeup", and "blow dry hair", which are less relevant to those in the visual space.

A more interesting example is shown in Figure 7(b). The neighbors of action "pommel horse" in the semantic space is "horse riding", "horse race", "floor gymnastics", "trampoline jumping", and "uneven bars", but "pommel horse" and "horse riding" show totally different movement. We believe that the word "horse" in both classes leads the negative neighbors. In contrast, its neighbors in the visual space are "pullups", "body weight squats", "field hockey penalty", "fencing", and "cricket bowling". Both of them have similar movements, making an accurately connected graph.

F. Limitations

Due to our transductive setting, our model requires to process unseen videos (without labels) and thus cannot be applied in an inductive setting. However, in practice, the most difficult to obtain are the labels of unseen videos, while unseen videos themselves are massively generated day-by-day. As a result, transductive setting is a practical solution under the current circumstance. Besides, similar to other transductive methods [52], [41], the category number of unseen data is also required in our model because K-Means clustering is performed in advance. This problem might be mitigated to use the clustering algorithm that without the need of the number of clusters, such as DBSCAN or Mean-Shift clustering.

V. Conclusion and Future Work

In this paper, we propose a novel framework to learn a propagation between visual space and semantic space for zero-shot action recognition. We first propose an visually-connected graph to accurately link the seen and unseen categories. Then a GAGCN is proposed to exploit the hierarchical knowledge in the graph, meanwhile predicting the label vectors of unseen categories. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art methods on three benchmark datasets.

The proposed methods mainly address the problem of the graph connection and the information flow of graph convolutional networks, which can be applied to different applications in the future. Also, we plan to study the correlations between graph connections and the architecture of graph convolutional networks.

Acknowledgement

This work is supported by NSFC (Grant No.61702194, 61972162, 61772206, U1611461, 61472145), Guangdong R&D Key Project of China (Grant No. 2018B010107003), Guangdong High-level Personnel of Special Support Program (Grant No. 2016TQ03X319), Guangdong Natural Science Foundation (Grant No. 2017A030311027), Guangzhou Key Project in Industrial Technology (Grant No. 201802010027, 201802010036), Key-Area Research and Development Program of Guangdong Province, China (2020B010165004, 2020B010166003, 2018B010107003), Guangdong High-level personnel program (Grant No. 2016TQ03X319), Guangdong NSF (2017A030311027), and Guangzhou key project in industrial technology (201802010027), a grant from the Hong Kong Polytechnic University (Project no. YBZE), and the CCF-Tencent Open Research fund (CCF-Tencent RAGR20190112). The first two authors contribute equally in this work.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 6

[2] Y. Cheng, X. Qiao, X. Wang, and Q. Yu. Random forest classifier for zero-shot learning based on relative attribute. *IEEE TNNLS*, 29(5):1662–1674, 2017. 2

[3] M. L. Chuang Gan, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic interclass relationships (sir) for zero-shot action recognition. In *AAAI*, pages 3769–3775, 2015. 1

[4] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis. Attributes2classname: A discriminative model for attributebased unsupervised zero-shot learning. In *ICCV*, 2017. 2

[5] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 1, 5

[6] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *IEEE TPAMI*, 36(2):303–316, 2014. 1

[7] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 37(11):2332–2345, 2015. 2

[8] J. Gao, T. Zhang, and C. Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 2

[9] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, June 2018. 2

[10] D. Huynh and E. Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *CVPR*, June 2020. 2

[11] M. Jain, J. C. van Gemert, T. Mensink, and C. G. Snoek. Objects2action: Classifying and localizing actions without any video example. In *ICCV*, pages 4588–4596, 2015. 6

[12] Z. Ji, K. Chen, J. Wang, Y. Yu, and Z. Zhang. Multi-modal generative adversarial network for zero-shot learning. *Knowledge-Based Systems*, page 105847, 2020. 2

[13] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han. Attribute-guided network for cross-modal zero-shot hashing. *IEEE TNNLS*, 2019. 2

[14] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing. Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496, 2019. 2, 3

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[17] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, pages 2452–2460, 2015. 1, 2, 6

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 5

[19] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, pages 4247–4255, 2015. 1, 2

[20] Y. Li, S.-h. Hu, and B. Li. Recognizing unseen actions in a domain-adapted embedding space. In *ICIP*, pages 4195–4199, 2016. 2, 6

[21] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao. Simple to complex transfer learning for action recognition. *IEEE TIP*, 25(2):949–960, 2016. 1

[22] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, pages 1996–2003, 2009. 2

[23] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *CVPR*, pages 3671–3680, 2017. 1

[24] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE TPAMI*, 40(10):2498–2512, 2018. 2

[25] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[26] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, pages 2929–2936, 2009. 2

[27] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[28] A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *CVPR*, pages 2188–2196, 2018. 2

[29] A. Mishra, V. K. Verma, M. S. K. Reddy, S. Arulkumar, P. Rai, and A. Mittal. A generative approach to zero-shot and few-shot action recognition. In *WACV*, pages 372–380, 2018. 5, 6

[30] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 6

[31] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, pages 716–723, 2013. 6

[32] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 5

[33] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang. Zero-shot action recognition with error-correcting output codes. In *CVPR*, page 6, 2017. 2, 3, 6

[34] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, pages 1641–1648, 2011. 1

[35] M. B. Sariyildiz and R. G. Cinbis. Gradient matching generative networks for zero-shot learning. In *CVPR*, pages 2168–2178, 2019. 2

[36] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 5, 6

[37] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014. 1, 2

[38] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song. Transductive unbiased embedding for zero-shot learning. In *CVPR*, pages 1024–1033, 2018. 5

[39] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[40] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, pages 588–595, 2014. 6

[41] Z. Wan, D. Chen, Y. Li, X. Yan, J. Zhang, Y. Yu, and J. Liao. Transductive zero-shot learning with visual structure constraint. In *NeurIPS*, pages 9972–9982, 2019. 1, 8

[42] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 1, 2, 5

[43] Q. Wang and K. Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124(3):356–383, 2017. 2, 6

[44] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866, 2018. 1, 3

[45] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, June 2018. 2

[46] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018. 2

[47] X. Xu, T. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *ICIP*, pages 63–67, 2015. 2, 3, 6

[48] X. Xu, T. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV*, pages 343–359, 2016. 2, 6

[49] X. Xu, T. Hospedales, and S. Gong. Transductive zero-shot action recognition by word-vector embedding. *IJCV*, 123(3):309–333, 2017. 1, 2, 3, 5, 6

[50] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 5

[51] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, pages 804–811, 2014. 6

[52] Y. Yu, Z. Ji, J. Guo, and Y. Pang. Transductive zero-shot learning with adaptive structural embedding. *IEEE TNNLS*, 29(9):4116–4127, 2017. 1, 2, 8

[53] C. Zhang and Y. Peng. Visual data synthesis via gan for zero-shot video classification. In *IJCAI*, pages 1128–1134, 2018. 1, 2, 6

[54] H. Zhang, H. Mao, Y. Long, W. Yang, and L. Shao. A probabilistic zero-shot learning method via latent nonnegative prototype synthesis of unseen classes. *IEEE TNNLS*, pages 1–15, 2019. 2

[55] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 2