Learning Using Privileged Information for Zero-Shot Action Recognition

Zhiyi Gao[1], Yonghong Hou[1(✉)], Wanqing Li[2], Zihui Guo[1], and Bin Yu[1]

[1] School of Electrical and Information Engineering, Tianjin University, China
{zhiyigao,houroy,gzihui,yubin_1449508506}@tju.edu.cn
[2] Advanced Multimedia Research Laboratory, University of Wollongong, Australia
wanqing@uow.edu.au

Abstract. Zero-Shot Action Recognition (ZSAR) aims to recognize video actions that have never been seen during training. Most existing methods assume a shared semantic space between seen and unseen actions and intend to directly learn a mapping from a visual space to the semantic space. This approach has been challenged by the semantic gap between the visual space and semantic space. This paper presents a novel method that uses object semantics as privileged information to narrow the semantic gap and, hence, effectively, assist the learning. In particular, a simple hallucination network is proposed to implicitly extract object semantics during testing without explicitly extracting objects and a cross-attention module is developed to augment visual feature with the object semantics. Experiments on the Olympic Sports, HMDB51 and UCF101 datasets have shown that the proposed method outperforms the state-of-the-art methods by a large margin.

Keywords: Action recognition · Zero-shot learning · Privileged information · Cross-attention.

1 Introduction

Research on action recognition from videos has made rapid progress in the past years with ceiling performance being reached on some datasets [1–5]. However, with the growing number of action categories, traditional supervised approach suffers from scalability problem. Moreover, annotating sufficient examples for the ever-growing new categories in real-world is cost-expensive and time-consuming. Therefore, extending a well-trained model to new/unseen classes, known as Zero-Shot Action Recognition (ZSAR), is gaining increasing interest recently.

In general, ZSAR assumes that both seen and unseen classes of actions share a common semantic space, e.g. a high dimensional vector space, and is achieved by learning a mapping from visual feature to the semantic space. Once the mapping is learned, an unseen sample is classified by the nearest neighbor search in the semantic space. There are a number of ways to define the semantic space. Typical early works [6–8] define the semantic space using hand-crafted attributes. Recently, the semantic space is often defined as the word embedding space of
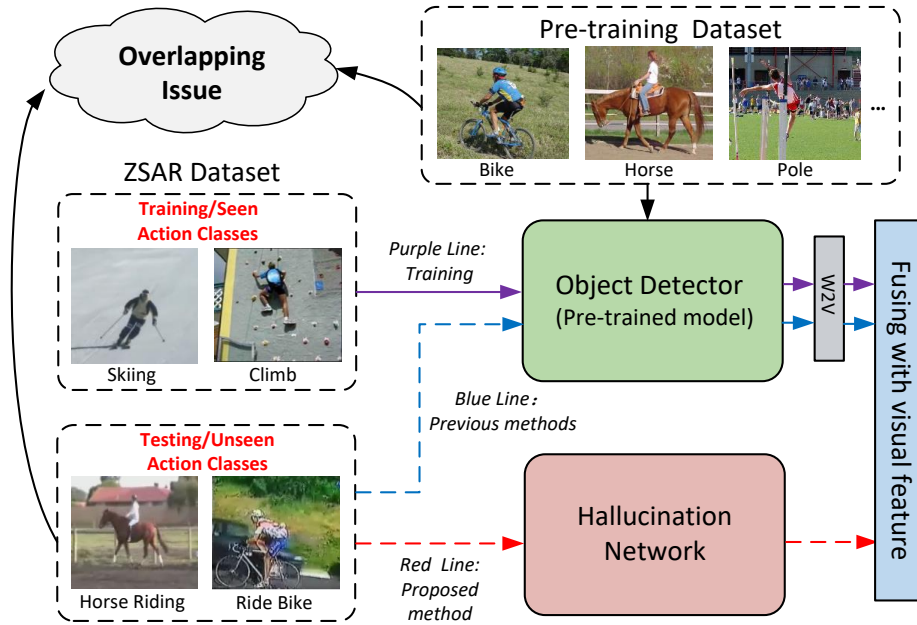
Fig. 1. Proposed method vs existing methods. An object detector pretrained on ImageNet is used to extract objects during training (purple line). When testing, the unseen actions are high-related to ImageNet classes. The previous methods (blue line) still require an object detector. Instead, the proposed method (red line) uses a hallucination network to extract object related information.

action names, i.e. word2vec, using a pretrained language model [9]. As for the mapping, direct learning from seen actions has been studied in [10]. However, visual observation and action names or hand-crafted attributes are two different modalities. The semantic gap between the two modalities has been challenging the effectiveness and robustness of learning such a mapping from visual feature to semantic representation.

To mitigate the semantic gap, several approaches have been reported recently. One approach is to define a semantic space using example images [11] e.g. considering the visual feature of the images as the semantic space. The second approach is to amend the visual feature with semantic information. For instance, in [12], a pretrained object detector is employed to recognize objects from action videos, and embedding of object names is treated as the visual representation without considering spatio-temporal information of the actions. In [13], objects are also extracted and the embedding of their names is concatenated to the feature representing poses to form the visual feature. Since such visual feature carries some amount of semantic information, the semantic gap is expected to be narrowed.

It has to be pointed out that both [12] and [13] require an object detector to extract objects and word-embed their names during testing. The detector is

pretrained on a large dataset such as the ImageNet [14], this practice has raised a question of validity of their methods being truly ZSAR because the large-scale dataset that is used to train the object detector likely contains images high-related to unseen action classes (see Fig. 1). For instance, images representing actions "Ride Bike" and "Horse Riding" are found in ImageNet [14] and these actions are likely to be considered as unseen actions in the random training-test split of ZSAR evaluation.

Nevertheless, using information of objects relevant to actions is a promising strategy. It not only narrows the semantic gap via word-embedding of the extracted object names, but also provides information in addition to the spatio-temporal visual information of the actions. As a true ZSAR, it is desirable that the object detector pretrained on a dataset should not be employed during testing for the sake of computation and avoiding potential licenses fee in commercial applications. This paper presents such a method based on the paradigm of learning using privileged information (LUPI) [15]. Specifically, objects are annotated or extracted offline from seen actions and their names are word-embedded into a vector in the visual space as privileged information (PI) in training. Unlike the methods in [12] and [13], our method *does not need the object detector during testing phase*, instead it uses a hallucination network to mimic the extraction of related semantic information. The output of the hallucination network is fused with the visual feature by a cross-attention module to narrow the semantic gap and assist the mapping from visual feature to the semantic space. Experiments on the widely used Olympic Sports [16], HMDB51 [17] and UCF101 [18] datasets show that the proposed method achieves the state-of-the-art results and outperforms the existing methods by a large margin.

The main contributions of this paper are as follows:

- A novel ZSAR method that implicitly leverages object information to narrow the semantic gaps and assists learning of a mapping from visual feature to a semantic space.
- A hallucination network with the privileged information of objects related to actions to "imitate" the object semantics during testing.
- A cross-attention module to fuse visual feature with object semantics or the feature from the hallucination network.
- Extensive evaluation on the widely used Olympic Sports, HMDB51 and UCF101 benchmarks achieves the state-of-the-art results.

2 Related work

2.1 Zero-Shot Action Recognition

In general, ZSAR is achieved by defining a common semantic space for seen and unseen actions and learning a mapping between the visual space and semantic space. In relation to the semantic space, early works define it based on hand-crafted attributes [6–8]. However, the definition and annotation of hand-crafted

attributes are subjective and labour-intensive. Recently, word-embedding of action names [10, 19, 20] is favorably adopted. In addition, textual descriptions of human actions or visual features extracted from still images are studied as a semantic space [11]. In [21, 22], semantically meaningful poses extracted from associated textual instructions of actions are also defined as a semantic space. Chen and Huang [23] use the textural description, called Elaborative Description (ED), of actions and embed them to define a semantic space. As for the visual space, it is generally defined through hand-crafted features [24] (e.g. the Dense Trajectory Features (DTF) [25] and the Improved Dense Trajectories (IDT) [26]) or deep features [27] extracted with the pre-trained 3D CNN-based Networks (e.g. Convolutional 3D Network (C3D) [28], Inflated 3D Network (I3D) [2]). After constructing semantic space and visual space, most existing works directly learn a mapping from the visual space to semantic space [10, 29]. However, this approach has been challenged by the semantic gap between the visual space and semantic space. Recently, some works attempt to amend the visual feature with object semantics in order to narrow the semantic gap between the visual space and semantic space. For example, Jain et al. [12] extract objects from action videos by a pretrained object detector and embed object names to form a visual representation. In [13], the object information and pose information are simultaneously extracted to form visual feature. However, both of them utilize the pretrained object detector at test time, which is questionable as illustrated in the introduction. In contrast, the proposed method in this paper does not use any object detector during testing.

2.2   Learning Using Privileged Information

Vapnik and Vashist [15] introduce the paradigm of learning using privileged information (LUPI). It assumes that there are additional data available during training, referred to as privileged information (PI), but not available in testing. A number of approaches have been proposed for action recognition by using privileged information. Niu et al. [30] utilize textual features extracted from the contextual descriptions of web images and videos as privileged information to train a robust action detector. Motiian et al. [31] explore several types of privileged information such as motion information or 3D skeletons for improving model learning. Crasto et al. [32] regard optical flow as privileged information, along with RGB for training, but only RGB is used in test to avoid flow computation. Similarly, Garcia et al. [33] consider that depth data is not going to be always available when a model is deployed in real scenarios. Therefore, depth data is used as privileged information along with RGB data during training, while employing only RGB data at test time. Different from the above methods, the proposed method uses object semantics from seen actions as privileged information to assist learning visual feature as well as a mapping between the visual feature and semantic space. In addition, a hallucination network is proposed to implicitly extract semantics during testing instead of relying on a pretrained object detector.
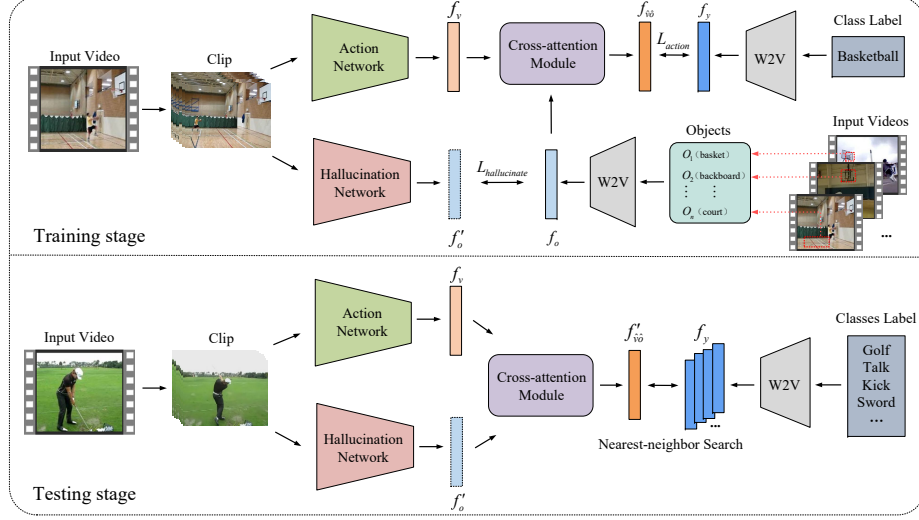
Fig. 2. The illustration of the proposed framework. It takes visual feature and the corresponding object semantics as input during training. A hallucination network is simultaneously trained to implicitly extract object related semantics and replace the object detector at test time. A cross-attention module is designed to fuse visual feature with object semantics or the feature from hallucination network. At test time, the nearest neighbor search is used for prediction in the semantic space. Gray blocks represent modules that are fixed during training. Colors (green, red, purple) blocks indicate modules trained.

3 Method

3.1 Overview

Let the training set $D_s = \{(v_s, y_s)|v_s \in \mathcal{V}_s, y_s \in \mathcal{Y}_s\}$ consist of videos v_s in video space $\mathcal{V}_s$ with labels from seen action classes $\mathcal{Y}_s$. Similarly, $D_u = \{(v_u, y_u)|v_u \in \mathcal{V}_u, y_u \in \mathcal{Y}_u\}$ is the test set, which consists of videos v_u with labels from unseen action classes $\mathcal{Y}_u$. Here, seen and unseen action classes are disjoint, i.e. $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$. The task of ZSAR is to train a model only on D_s, and predict the class labels of the unseen videos v_u. The proposed architecture is illustrated in Fig. 2, where the embedding space f_y of action names is used as the shared semantic space, visual feature f_v extracted by the action network is used as visual space. In addition, the object semantic f_o for each known action class is utilized as privileged information to assist learning of a mapping from visual feature to the semantic space. The object semantic f_o is the word embedding of object o which may be obtained offline using an object detector. Specifically, in the training stage, video features f_v and object semantics f_o are fed to a cross-attention module and fused into a joint representation $f_{\hat{v}\hat{o}}$ being matched with the semantic representations f_y in the semantic space. Meanwhile, a hallucination network

is trained to "imitate" the object semantics f_o. At test time, the trained hallucination network is used to implicitly extract object related semantic f'_o, which is fused with the visual feature f_v by the cross-attention module to obtain the joint representation $f'_{\hat{v}\hat{o}}$ as well. Finally, nearest-neighbor search method is used in the semantic space to recognize unseen action class. In the rest of the section, the five key components of the proposed framework including video and semantic embedding, object embedding, hallucination network, cross-attention module, and training and testing of the framework will be described in detail.

3.2   Video and Semantic Embedding

Video embedding. Given a video clip $v \in \mathbb{R}^{T \times H \times W}$ of T frames with a height and width of H and W respectively, the action network consists of a pretrained R(2+1)D [10] network followed by a fully connected layer for extracting a video spatio-temporal feature $f_v \in \mathbb{R}^{300}$

Semantic embedding. Following the majority of works [19], word2vec is applied to embed action names as semantic representation. Specifically, the skipgram language model [9] trained on 100 billion words from Google News articles is used to map each word into a 300 dimensional semantic space. For the class name containing multiple words, the vectors of all words are averaged to obtain the semantic embedding. That is, a class name consisting of n words $y = \{y_1, \cdots, y_n\}$ can be embedded as $f_y = \frac{1}{n} \sum_{i=1}^{n} \text{W2V}(y_i) \in \mathbb{R}^{300}$.

3.3   Object Embedding as PI

Objects are extracted by a pretrained object detector from all video clips with t frames uniformly sampled, and the top k objects with the highest probabilities are reserved. In order to associate the extracted objects with action classes, we calculate the frequency of the objects extracted from all clips of the same action class and preserve the top m objects $o = \{o_1, \cdots, o_m\}$. Then the objects are word-embedded as the privileged information of individual action class. Similarly, the object semantic of each action class can be presented by $f_o = \frac{1}{m} \sum_{i=1}^{m} \text{W2V}(o_i) \in \mathbb{R}^{300}$.

3.4   Hallucination Network

Since the ImageNet dataset from which the object detector is pretrained contains highly related categories with unseen action classes, we argue that the object detector should not be employed during testing phase for truly ZSAR. To replace the object detector, we introduce a hallucination network, which is trained to "imitate" the object semantics. Thus at test time, it can extract the object related semantics f'_o to fuse with visual features. The hallucination network consists of a fixed pretrained R(2+1)D [10] network and four fully connected layers.
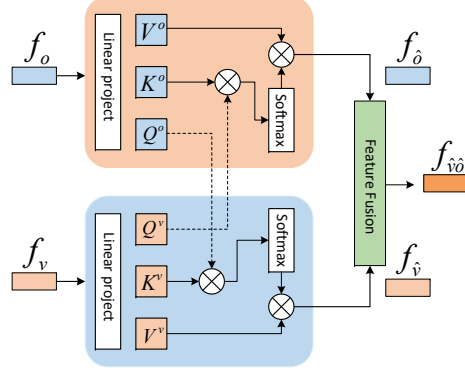
Fig. 3. Illustration of the proposed cross-attention module. It consists of a mutual-attention layer and feature fusion operation. Visual features and object semantics or features from hallucination network are fed to the attention layer to exchange the information. Then the features are fused together to obtain the joint representations.

3.5 Cross-attention Module

We propose a cross-attention module to fuse visual feature f_v with object semantics f_o (training stage) or the feature f_o' (testing stage) from the hallucination network using mutual-attentional mechanism. As shown in Fig. 3, the proposed module contains two parts, mutual-attention layer and feature fusion operation. Take the training phase as an example, mathematically, visual feature f_v and object semantic f_o are first mapped to a series vectors query (Q^v/Q^o) and key (K^v/K^o) of dimension d_k, and value (V^v/V^o) of dimension d_v using unshared learnable linear projections. Then the mutual-attention is applied to retrieve the information from vectors (key K^v and value V^v) of visual feature related to query vector Q^o of object semantics and vice the verse. The dot-product of the Q^o vector and K^v vector goes through a softmax function to obtain the attention weights, and have a weighted sum of V^v. The process can be expressed as

$$f_{\hat{v}} = \text{Softmax}\left(\frac{Q^o(K^v)^T}{\sqrt{d_k}}\right)V^v \tag{1}$$

Similarly, we can calculate the attention weights on each object semantic with respect to visual feature to obtain $f_{\hat{o}}$.

$$f_{\hat{o}} = \text{Softmax}\left(\frac{Q^v(K^o)^T}{\sqrt{d_k}}\right)V^o \tag{2}$$

Such a mutual-attention scheme takes cross-modal interactions into consideration. It calculates the visual feature attention in semantic modality and object semantic attention in visual modality, and produces corresponding cross-modal features. After mutual-attention layer, features from both modalities are fused via simply feature addition operation.

$$f_{\hat{v}\hat{o}} = f_{\hat{v}} + f_{\hat{o}} \tag{3}$$

3.6   Training and Testing

As illustrated in Fig. 2, in the training stage, the visual feature f_v and object semantic f_o are fused to obtain a joint representation $f_{\hat{v}\hat{o}}$ in the semantic space to match their equivalent semantic representation f_y. The loss function $\mathcal{L}_{action}$ is defined as

$$\mathcal{L}_{action} = \|f_y - f_{\hat{v}\hat{o}}\|_2^2 \tag{4}$$

In addition, a regression loss between the object semantics f_o and feature f_o' from the hallucination network is defined

$$\mathcal{L}_{hallucinate} = \|f_o - f_o'\|_2^2 \tag{5}$$

The full model is end-to-end trained, and the overall loss of the proposed model is defined as

$$\mathcal{L} = \mathcal{L}_{action} + \mathcal{L}_{hallucinate} \tag{6}$$

At test time, a test video goes through both the action network and the hallucination network to produce the visual feature f_v and object related semantics f_o'. After the cross-attention module, a joint representation $f_{\hat{v}\hat{o}}'$ is obtained. Finally, nearest-neighbor search is used in the semantic space to predict the label $\mathcal{P}$ of unseen action class as

$$\mathcal{P} = \underset{y \in D_u}{\operatorname{argmin}} \cos(f_{\hat{v}\hat{o}}', f_y) \tag{7}$$

where $\cos(\cdot)$ is the cosine distance.

4   Experiments

4.1   Setup

Datasets and Splits. The proposed method is extensively evaluated on the Olympic Sports [16], HMDB51 [17] and UCF101 [18] datasets. The Olympic Sports dataset consists of 783 videos, which are divided into 16 sports actions. The HMDB51 dataset contains 6766 realistic videos distributed in 51 actions. The UCF101 dataset has 101 action classes with a total of 13,320 videos, which are collected from YouTube. To compare the proposed method with the state-of-the-art methods, we follow the same 50/50 data splits as used in [29], i.e., videos of 50% categories are used for training and the rest 50% categories are considered as the unseen for testing. Specifically, experiments are conducted on the cases of 8/8, 26/25 and 51/50 splits for Olympic Sports, HMDB51 and UCF101, respectively. For each case, 30 independent splits for each dataset [34] are randomly generated and the average accuracy and standard deviation are reported for experimental evaluation.

ZSAR Settings. There are two ZSAR settings: inductive setting and transductive setting. The former assumes that only the labeled videos from the seen categories are available during training while the latter can use the unlabeled data of the unseen categories for model training. Specifically, in this work, we focus on inductive ZSAR [10, 13, 23, 35] and do not discuss the transductive approach [7, 29].

4.2 Implementation Details

In the experiments, we use the pretrained R(2+1)D network [10] followed by a fully connected layer to extract spatio-temporal features. The object detector is the BiT image model [36] pretrained on ImageNet21k [14]. When extracting objects, frame t of each video clip is set to 8 and top k of all action clips is set to 20, top m of each action class is set to 5. Both the action class and extracted objects are embedded using skip-gram word2vec model [9], which is pretrained on the Google News dataset. The word2vec model generates a 300-dimensional word vector representation for each word. For the input videos, we select a clip of 16 frames from each video, uniformly sampled in time, and each frame is cropped to 112 pixels × 112 pixels. The experiments are conducted on two TitanX GPUs. The networks are trained for 10 epochs on HMDB51 and UCF101 datasets, 20 epochs on the Olympic Sports dataset with a batch size of 16.

We use Adam optimizer [37] to train all networks. The learning rate is initially set to 1e-4 and is decayed by a factor of 0.5 every 5 epochs. The average Top-1 and Top-5 accuracy (%) ± standard deviation are reported.

Table 1. ZSAR performances on the three existing benchmarks compared with state-of-the-art methods. VE and SE represent Visual Embedding and Semantic Embeding. FV represents Fisher Vectors, obj represents objects. A represents the human annotated attribute vectors, W represents the word2vec embedding, and ED represents elaborative description. The average Top-1 accuracy (%) ± standard deviation is reported.

| Methods	VE	SE	Olympic Sports	HMDB51	UCF101
O2A[12]	Obj	W	N/A	15.6	30.3
MTE[38]	FV	W	44.3 ± 8.1	19.7 ± 1.6	15.8 ± 1.3
ASR[11]	C3D	W	N/A	21.8 ± 0.9	24.4 ± 1.0
GMM[34]	C3D	W	N/A	19.3 ± 2.1	17.3 ± 1.1
UAR[39]	FV	W	N/A	24.4 ± 1.6	17.5 ± 1.6
CEWGAN[40]	I3D	W	50.5 ± 6.9	30.2 ± 2.7	26.9 ± 2.8
TS-GCN[41]	GCN	W	56.5 ± 6.6	23.2 ± 3.0	34.2 ± 3.1
BD-GAN[35]	C3D	A	49.80 ± 10.72	23.86 ± 2.95	18.73 ± 3.73
E2E[10]	R(2+1)D	W	N/A	32.7	48
VDARN[13]	GCN	W	57.6 ± 3.4	21.6 ± 2.8	26.4 ±2.8
ER[23]	TSM	ED	60.2 ± 8.9	35.3 ± 4.6	51.8 ± 2.9
Ours	R(2+1)D	W	**61.9 ± 7.8**	**38.8 ± 4.6**	**52.6 ± 2.4**

4.3   Comparison with the State-of-the-art

The comparison results are shown in Table 1. Overall, the proposed method performs best against state-of-the-art methods on three widely used datasets [16–18]. When using the same word embedding methods, our approach outperforms recent VDARN [13], and TS-GCN [41] methods on three datasets. Compared with other attributes-based [35] and ED-based [23] methods which are not scalable and labour-intensive, the proposed method uses the word embedding of action names as semantic space and still obtains better performance than them. Compared with E2E [10] method which also uses the R(2+1)D network and directly maps from visual features to the semantic space, our method improves by 6.1 and 4.6 percentage points on the HMDB51 and UCF101 datasets respectively, indicating that amending the visual feature with semantic information can effectively narrow the semantic gap. Note that the O2A [12] and the VDARN [13] methods also use object information as part of the visual feature, but they use object detector during testing as well. Our method uses a hallucination network to extract object related information during testing and still performs better than them. In addition, the proposed method not only outperforms the very recent method ER [23] in classification accuracy, but also has a smaller standard deviation, which indicates that our method has a relatively stable performance under different training and testing data partitions.

Table 2. Ablation study of the proposed object semantics and hallucination network on HMDB51 and UCF101 datasets. "Baseline" refers to directly mapping visual features to semantic space; "OS (Object Semantics)" refers to extracting object semantics through an object detector to fuse with visual features during training, but the detector is not used in the test stage; "HN (Hallucination Network)" refers to implicitly extracting object semantics to fuse with visual features at test time. The average Top-1 and Top-5 accuracy (%) $\pm$ standard deviation are reported.

Baseline	OS	HN	HMDB51		UCF101	
			Top-1	Top-5	Top-1	Top-5
✓	×	×	32.7 ± 3.0	56.3 ± 3.4	47.9 ± 2.2	72.4 ± 3.4
✓	✓	×	36.7 ± 4.3	65.6 ± 5.0	49.8 ± 2.5	76.9 ± 2.7
✓	✓	✓	$\mathbf{38.8 \pm 4.6}$	$\mathbf{66.9 \pm 3.7}$	$\mathbf{52.6 \pm 2.4}$	$\mathbf{77.1 \pm 2.9}$

4.4   Ablation Studies

Impact of Object Semantics and Hallucination Network To verify the effectiveness of the proposed object semantics and hallucination network, ablation studies are conducted on HMDB51 and UCF101 datasets. As shown in the second row of Table 2, leveraging object semantics to augment visual feature only at the training stage can significantly improve classification accuracy on
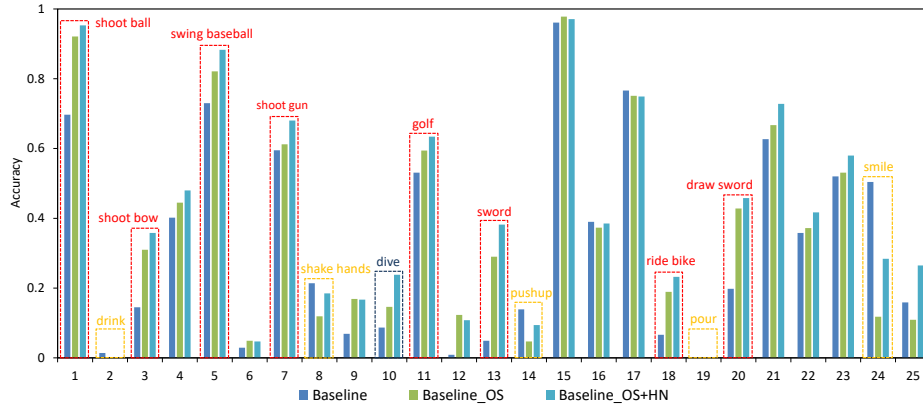
Fig. 4. Accuracy comparison of each unseen action class on the HMDB51 dataset between the proposed method and baseline. Here, OS represents Object Semantics, HN represents Hallucination Network. The experimental settings of OS and HN are the same as Table 2.

the two datasets. The hallucination network contributes further improvement by 4.9/2.1/2.8 percentage points for Top-1 accuracy and 2.1/1.3/0.2 percentage points for Top-5 accuracy on the three datasets respectively (see the last row of Table 2). Fig. 4 further shows the accuracy comparison of each unseen action. It can be seen that the recognition accuracy of our method is significantly higher than the baseline in the red-marked categories, such as "shooting ball", "shooting bow", "sword", "ride bike" etc. These actions involve interaction with objects thus the hallucination network extracts the semantics of the related object to augment the visual feature and improve the recognition. In addition, some actions such as "dive" (marked in blue) occur in a specific scene, and information on the scene is extracted by the hallucination network to augment the visual information. However, for some actions such as "smile" and "shake hands" (marked in yellow), these actions do not involve any interaction with the surrounding environment. In this case, the information extracted by the hallucination network may interfere with the visual feature. The interference can be minimized by the choice of the objects in the training (See below section 4.4 Impact of Object Number).

Impact of Fusion Strategy In order to verify that the proposed cross-attention module is better than other fusion strategies, we perform another ablation experiment on the HMDB51 and UCF101 datasets, the results are shown in Table 3. Compared with multiplication, concatenation and addition fusing schemes, the proposed fusion strategy achieves the best performance. The reason is that other fusion strategies neglect the modality discrepancy. The proposed cross-attention block can effectively mitigate such an issue, in which the visual features and object semantics can exchange information through the mutual-attention layer

Table 3. Ablation study of the proposed cross-attention module with different fusion manners on HMDB51 and UCF101 datasets. The average Top-1 and Top-5 accuracy (%) ± standard deviation are reported.

Methods	HMDB51		UCF101	
	Top-1	Top-5	Top-1	Top-5
Multiplication	34.1 ± 4.6	56.6 ± 3.6	49.4 ± 3.4	73.2 ± 3.6
Concatenation	35.3 ± 4.7	58.6 ± 5.4	51.8 ± 2.8	75.4 ± 3.4
Addition	37.1 ± 4.3	64.5 ± 4.4	51.9 ± 2.5	76.9 ± 3.5
Cross-attention	$\mathbf{38.8 \pm 4.6}$	$\mathbf{66.9 \pm 3.7}$	$\mathbf{52.6 \pm 2.4}$	$\mathbf{77.1 \pm 2.9}$

to reduce the feature variations. Then the cross-modal features are fused to produce the conjoint feature representation. The results also show that the proposed cross-attention module has the smallest standard deviation on the UCF101 dataset.

Impact of Object Number Fig. 5 shows the impact of the object number used for each action class. It can be seen that the performance first increases as the number of objects increases and then decreases when the number of objects is over five for the HMDB51 dataset. This is probably due to that each action class is related to only a limited number of objects, and irrelevant objects will become noise interference and degrade performance. However, the performance degradation is insignificant (only about 0.3 percentage points) even if the object number is increased to nine.
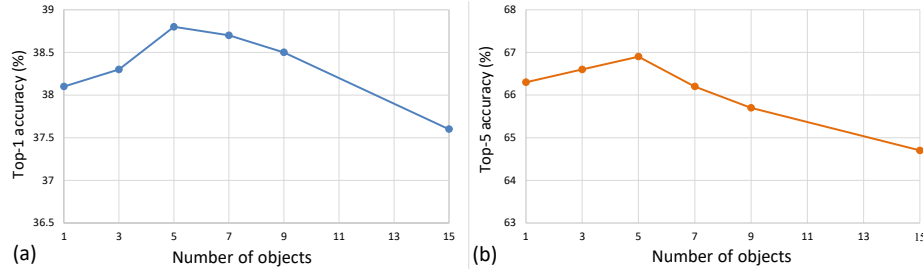
Fig. 5. Evaluate the proposed method with different number of objects on HMDB51 dataset. The average Top-1 (a) and Top-5 (b) accuracy (%) are illustrated.

4.5   Qualitative Results

We further analyze the effectiveness of the proposed method via the qualitative visualizations of visual feature distribution in the semantic space [42], as
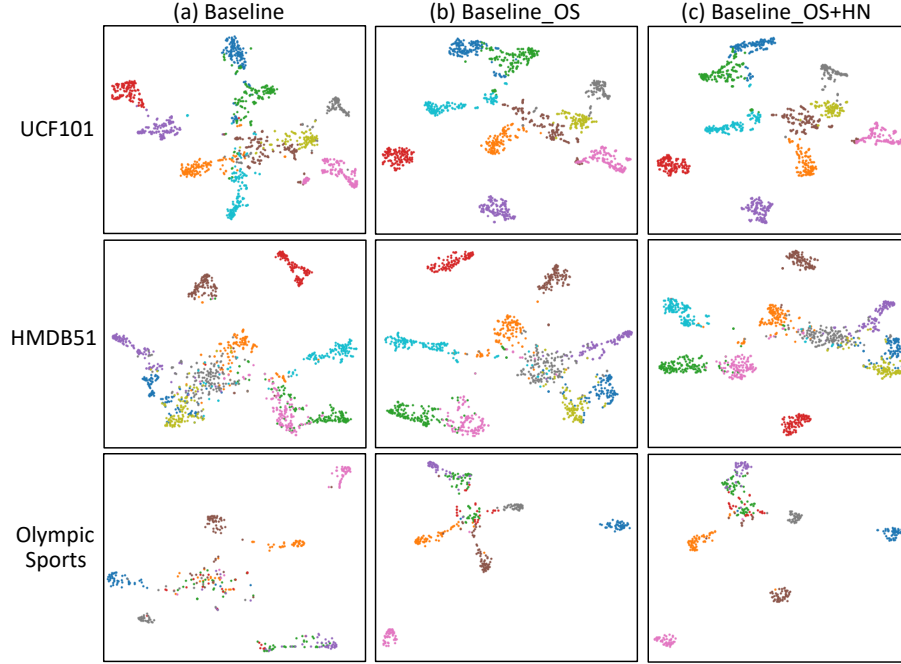
Fig. 6. The t-SNE visualization of the visual feature distribution of unseen classes of the three datasets. Different classes are shown as dots in different colors. Here, OS represents Object Semantics, HN represents Hallucination Network. The experimental settings of OS and HN are the same as section 4.4.

shown in Fig. 6. For the sake of visualizations, we randomly sample 8 unseen classes from the Olympic Sports dataset and 10 unseen classes from HMDB51 and UCF101 datasets. As can be seen from Fig. 6(a) and Fig. 6(b), with the aid of the object detector during training, the baseline network yield improved feature representations. In Fig. 6(c), we further use a hallucination network to mimic the object semantics to fuse with visual features at test time, which can obtain the tightest intra-class clusters and separable inter-class clusters in the semantic space. The visualization results prove that the proposed method can effectively use object semantics to amend visual representation, thus alleviating the semantic gap.

5   Conclusions

In this paper, we propose a novel method for ZSAR. It is built upon the paradigm of learning using privileged information (LUPI). Specifically, the object semantics are used as privileged information to narrow the semantic gaps and assist learning of a mapping from visual feature to a semantic space. A hallucination

network is trained to implicitly extract object related semantics at test time instead of using a pretrained object detector. Moreover, a cross-attention module is designed to fuse visual feature with object semantics or the feature from hallucination network. Empirical experiments on the Olympic Sports, HMDB51 and UCF101 benchmarks have shown that the proposed method achieves the state-of-the-art results.

References

1. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1933–1941
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 6299–6308
3. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4768–4777
4. Wang, Y., Long, M., Wang, J., Yu, P.S.: Spatiotemporal pyramid network for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1529–1538
5. Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., Li, M.: A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567 (2020)
6. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2011) 3337–3344
7. Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z., Gong, S.: Transductive multi-view embedding for zero-shot recognition and annotation. In: Proceedings of the European Conference on Computer Vision, Springer (2014) 584–599
8. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. International Journal of Computer Vision **124** (2017) 356–383
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., Chalupka, K.: Rethinking zero-shot video classification: End-to-end training for realistic applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2020) 4613–4623
11. Wang, Q., Chen, K.: Alternative semantic representations for zero-shot human action recognition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (2017) 87–102
12. Jain, M., Van Gemert, J.C., Mensink, T., Snoek, C.G.: Objects2action: Classifying and localizing actions without any video example. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 4588–4596
13. Su, Y., Xing, M., An, S., Peng, W., Feng, Z.: Vdarn: Video disentangling attentive relation network for few-shot and zero-shot action recognition. Ad Hoc Networks **113** (2021) 102380
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Ieee (2009) 248–255

15. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22** (2009) 544–557
16. Niebles, J.C., Chen, C.W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Proceedings of the European Conference on Computer Vision, Springer (2010) 392–405
17. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proceedings of the IEEE International Conference on Computer Vision, IEEE (2011) 2556–2563
18. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
19. Xu, X., Hospedales, T., Gong, S.: Semantic embedding space for zero-shot action recognition. In: 2015 IEEE International Conference on Image Processing, IEEE (2015) 63–67
20. Bishay, M., Zoumpourlis, G., Patras, I.: Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. arXiv preprint arXiv:1907.09021 (2019)
21. Zhou, L., Li, W., Ogunbona, P., Zhang, Z.: Semantic action recognition by learning a pose lexicon. Pattern Recognition **72** (2017) 548–562
22. Zhou, L., Li, W., Ogunbona, P., Zhang, Z.: Jointly learning visual poses and pose lexicon for semantic action recognition. IEEE Transactions on Circuits and Systems for Video Technology **30** (2019) 457–467
23. Chen, S., Huang, D.: Elaborative rehearsal for zero-shot action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2021) 13638–13647
24. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2009) 951–958
25. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 3551–3558
26. Wang, H., Oneata, D., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. International Journal of Computer Vision **119** (2016) 219–238
27. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6450–6459
28. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 4489–4497
29. Xu, X., Hospedales, T., Gong, S.: Transductive zero-shot action recognition by word-vector embedding. International Journal of Computer Vision **123** (2017) 309–333
30. Niu, L., Li, W., Xu, D.: Visual recognition by learning from web data: A weakly supervised domain generalization approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2774–2783
31. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Information bottleneck learning using privileged information for visual recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 1496–1505
32. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented rgb stream for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7882–7891

33. Garcia, N.C., Morerio, P., Murino, V.: Learning with privileged information via adversarial discriminative modality distillation. IEEE Transactions on Pattern Analysis and Machine Intelligence **42** (2020) 2581–2593
34. Mishra, A., Verma, V.K., Reddy, M.S.K., Arulkumar, S., Rai, P., Mittal, A.: A generative approach to zero-shot and few-shot action recognition. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. (2018) 372–380
35. Mishra, A., Pandey, A., Murthy, H.A.: Zero-shot learning for action recognition using synthesized features. Neurocomputing **390** (2020) 117–130
36. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: Proceedings of the European Conference on Computer Vision, Springer (2020) 491–507
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
38. Xu, X., Hospedales, T.M., Gong, S.: Multi-task zero-shot action recognition with prioritised data augmentation. In: Proceedings of the European Conference on Computer Vision, Springer (2016) 343–359
39. Zhu, Y., Long, Y., Guan, Y., Newsam, S., Shao, L.: Towards universal representation for unseen action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 9436–9445
40. Mandal, D., Narayan, S., Dwivedi, S.K., Gupta, V., Ahmed, S., Khan, F.S., Shao, L.: Out-of-distribution detection for generalized zero-shot action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9985–9993
41. Gao, J., Zhang, T., Xu, C.: I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 8303–8311
42. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008)