Aligned Dynamic-Preserving Embedding for Zero-Shot Action Recognition

Yi Tian, Yu Kong, *Member, IEEE,* Qiuqi Ruan, *Senior Member, IEEE,*Gaoyun An, *Member, IEEE,*
Yun Fu, *Fellow, IEEE*

***Abstract*—Zero-shot learning (ZSL) typically explores a shared semantic space in order to recognize novel categories in the absence of any labeled training data. However, traditional ZSL methods always suffer from serious domain shift problem in human action recognition. This is because: (1) Existing ZSL methods are specifically designed for object recognition from static images, which don't capture temporal dynamics of video sequences. Poor performances are always generated if those methods are directly applied to zero-shot action recognition. (2) Those methods always blindly project target data into a shared space using a semantic mapping obtained by source data without any adaptation, in which underlying structures of target data are ignored. (3) Severe inter-class variations exist in various action categories. Traditional ZSL methods don't take relationships across different categories into consideration. In this paper, we propose a novel aligned dynamic-preserving embedding model (ADPE) for zero-shot action recognition in a transductive setting. In our model, an adaptive embedding of target videos is learned exploring the distributions of both source and target data. An aligned regularization is further proposed to couple the centers of target semantic representations with their corresponding label prototypes in order to preserve the relationships across different categories. Most significantly, during our embedding, temporal dynamics of video sequences are simultaneously preserved via exploiting temporal consistency of video sequences and capturing temporal evolution of successive segments of actions. Our model can effectively overcome the domain shift problem in zero-shot action recognition. Experiments on Olympic sports, HMDB51 and UCF101 datasets demonstrate the effectiveness of our model.**

***Index Terms*—Zero-shot learning, Action recognition, Aligned dynamic-preserving embedding.**

I. INTRODUCTION

Human action recognition is one of the most popular research topics in computer vision field, which has been widely applied in a number of applications [15][14], such as human-computer interaction, video surveillance systems, etc. Traditional video based action recognition methods follow a standard framework of supervised pattern recognition [44][27][38][45][49], which can be summarized as three processing steps: feature extraction, video representation and classification. Over the past few years, many effective methods have been proposed to promote the development of human action recognition. Most of them mainly focus on exploring effective features [20][41][43] and discriminative representations [28][42] of action videos on the condition that sufficient labeled training videos are provided. With a rapid development of social networking sites (Facebook, Twitter, etc) and video capture devices, the number of videos, categories of actions and complexities of videos' content are explosively increasing. For the traditional supervised human action recognition, it's expensive to annotate the huge number of videos and manual annotations would cause large confusions as well. Besides, it's difficult to make a clear definition for each action as the categories grow more fine-grained and inter-related, such as actions: 'Kick' and 'Kick-ball'.

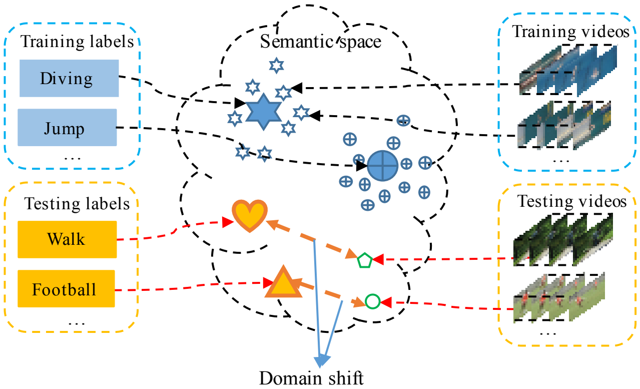


Fig. 1. Illustration of domain shift problem. Both labels and videos are embedded into a shared semantic space. The learned semantic representations of target videos are far away from their corresponding prototypes in the shared space due to the isolation between training and testing sets.

Zero-shot learning [4][19] offers an innovative solution to those scenarios, which aims to recognize *unseen* categories without any labeled training data. The essence of ZSL is to explore an intermediate semantic space [18], in which the shared knowledge across different categories can be transferred from *seen* categories to *unseen* ones. Novel categories can be recognized on the basis of those knowledge learned from seen categories. The semantic spaces can be predefined by a human user, e.g. attribute space [21], or learned using existing knowledge bases, e.g. word2vector space [24]. Take

Y. Tian is with the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China (e-mail: tianyi@bjtu.edu.cn).

Y. Kong is with B. Thomas Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA (e-mail: yu.kong@rit.edu).

Qiuqi Ruan and Ganyun An are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: qqruan@bjtu.edu.cn; gyan@bjtu.edu.cn).

Y. Fu is with the Department of Electrical and Computer Engineering, College of Engineering, and Khoury College of Computer and Information Sciences, Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

attribute space as an example, the attributes describe properties of each category (action) and are predefined manually. Each category is represented as a binary vector in terms of the presence or absence of certain attributes. After that, classifier of each attribute is learned using training samples. Each testing sample that is disjoint from training samples is specified as a novel category in terms of the scores measured by learned attribute classifiers. For example, in training stage, the labeled samples of categories 'Horse' and 'Bee' are provided. In testing stage, we need to recognize the samples that contain the unseen category 'Zebra'. We can infer the unseen category via utilizing the knowledge learned from 'Horse' and 'Bee', eg. owing horsehair and covered with stripes, even if we don't have any training samples of 'Zebra'. Obviously, ZSL takes advantage of a small quantity of existing labeled data to guide the recognition of disjoint unseen data, which greatly reduces manual annotation cost and makes visual recognition scalable. In recent years, it attracts lots of attentions and achieves impressive performance in object recognition [2][19].

As training set and testing set are totally disjoint, traditional ZSL methods always suffer from serious domain shift problem [46][5]. It causes a great discrepancy between a specified prototype position and representations of its class member instances in semantic space (Fig. 1). And it proves more severe in the task of zero-shot action recognition [46]. There are three main reasons: **Firstly**, traditional ZSL methods are specifically designed for object recognition from static images [12][50]. When they are extended to action recognition task, video sequences are firstly encoded into orderless representations, after which the same operations of ZSL in object recognition are executed. Obviously, those methods don't take temporal information of video sequences into consideration during their processing, which lead to poor performance of zero-shot action recognition. **Secondly**, most of existing ZSL methods use a semantic mapping for target domain[1] that is independently learned from source domain (by projecting source data onto a shared semantic space) without any adaptation. The blind using of training mapping only preserves underlying structures of source data but ignores the structures of target. Thus, target data are likely to be projected into irrelative regions, which are far away from their prototypes. **Thirdly**, there exist large inter-class variations in large-scale action datasets (e.g. HMDB51, UCF101). Traditional ZSL methods don't take those variations across categories into consideration, in which the knowledge transferred from source domain are insufficient to describe novel categories.

In this paper, we propose a novel aligned dynamic-preserving embedding model (ADPE) for zero-shot action recognition, which overcomes the above limitations of existing ZSL models. Our model can effectively alleviate the domain shift problem in zero-shot action recognition. Specifically, our model is implemented in a transductive setting [30] that assumes accessing to a full set of target data. Our goal is to explore an appropriate embedding that jointly projects source and target videos into a shared semantic space, where each target video is close to its corresponding true action label.

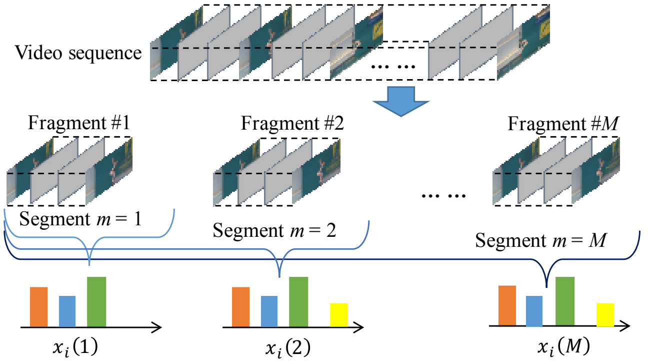


Fig. 2. Example of the division of a video sequence. The video is firstly divided into M uniform fragments. The m_{th} segment consists of the first m fragments. Each segment is encoded into a visual representation and the video is eventually represented as a set of M visual representations.

Different and superior to existing methods that encode each video sequence into an orderless representation [47][48][46], we divide each video into a series of successive segments [14], as shown in Fig. 2. In order to take advantages of temporal information of video sequences, our embedding model (ADPE) follows two mapping criterions. On the one hand, our model projects all the sequential segments of a specified video into the same semantic representation, which ensures temporal consistency of the video. On the other hand, our embedding ensures that the confidence of each segment being mapped to its true label is non-decreasing as much more frames are involved in the segment, which simulates temporal evolutions of actions. Thus, ADPE captures temporal dynamics of each video. In addition, we learn an adaptive embedding of target domain instead of directly applying the embedding learned from source data. With the consideration of both source and target data, much more semantic knowledge can be captured by our adaptive embedding. Furthermore, in order to handle large variations across complex action categories, an aligned regularization term is proposed. With this term, centers of target semantic representations and their corresponding label prototypes are aligned in the semantic space, thereby relationships between different action categories are preserved. In summary, our ADPE model can not only make full use of appearance information of source and target videos, but also capture temporal information of them. Thus, domain shift problem can be effectively reduced.

The main contributions of this work are summarized as follows:

- Our method is the early work to characterize temporal dynamics of each video during embedding processing, which describes action evolution and temporal consistency of videos.
- An adaptive embedding of target videos is learned to explore the underlying distributions of both source and target data, which aims to transfer shared semantic information from source domain to target one.
- An aligned regularization term is proposed to take variations across different action categories into consideration, which aligns the centers of target semantic representations with their corresponding label prototypes.

---

[1] In this paper, source domain equals to training set, and target domain equals to testing set.

Overview of our method. The framework of our method is shown in Fig. 3. At first, each video is divided into several successive segments (Fig. 2), each of which is encoded into a visual representation. The action labels are projected onto a semantic space manually (attribute space) or learned via existing knowledge (word2vector space).

At training stage, labels of videos are given. A naive visual-to-semantic embedding A^s is obtained via ridge regression. The embedding should project each segment of the video onto a same semantic representation, which keeps temporal consistency of the video. And the embedding also simulates action evolution of segment-by-segment sequences, which reflects that the segment involves more frames is more confident to being mapped to its true label.

At testing stage, instead of blindly using the mapping A^s learned from source data, an adaptive mapping A^t is learned for target domain under the guidance of A^s. In order to further overcome domain shift problem, centers of semantic representations of target data are enforced to align with their true labels in semantic space. During the learning of A^s, temporal consistency and action evolution of videos are also considered. Finally, each target video is classified to its nearest label prototype in the semantic space.

II. RELATED WORK

A. Action recognition

Over the past few years, many methods [44][27][38] have been proposed to promote the development of human action recognition in videos. Most of those methods mainly focus on exploring efficient features [20][41][43] and discriminative representations [28][42] of video sequences. Local space-time feature based approaches [27] have gained popularity and shown impressive results. Recently, deep learning methods have been introduced in this filed and got promising performance, such as LSTM model [8], C3D model [39], etc.

Those traditional methods have one property in common that sufficient labeled training videos are required to train their models. With a rapid development of social networking sites (Facebook, Twitter, etc) and video capture devices, the number of video data, categories of actions and complexities of videos' content are explosively increasing. The video datasets for action recognition that are mostly collected from Internet videos are constantly developing. The early datasets are the indoor and constrained ones, which focus on simple and isolated human actions performed by a single person, e.g. KTH (2004) [31] and Weizmann (2005) datasets [7]. Nowadays, the datasets are outdoor and wild datasets, which include complex backgrounds and diversified environment, e.g. OlympicSports (2010) [26], HMDB51 (2011) [17] and UCF101 (2012) datasets [32]. On the one hand, it's expensive to annotate the huge number of videos manually and manual annotations would cause large confusions as well. Besides, it's also difficult for humans to make a clear definition for each action as their categories grow more fine-grained and inter-related. On the other hand, it's incapable for traditional models to recognize the actions accurately without sufficient labeled training videos. Therefore, it's essential and meaningful to explore some methods that can handle the recognition problems of complex action categories with less labeled data or even without any labeled videos.

B. Zero-shot learning

The emerging zero-shot learning (ZSL) [4][19] appropriately meets the situation that mentioned above, which aims to recognize *novel* categories in the absence of any labeled training data. The key of ZSL is to learn a classifier $f : X \rightarrow Y$ that predicts novel values of Y that are omitted from training set. To achieve this, a shared semantic space is always defined in ZSL, which transfers shared characteristics from source domain to target domain. The most common semantic spaces used in ZSL are attribute space [19] and word2vector space [3]. In semantic space, the representations of action categories are also called 'prototypes'.

Attribute space: In attribute space, a series of attributes are predefined manually to describe shared properties across different categories [21][19], such as visual appearance (e.g. is-sharp), body parts (e.g. torso), and motion patterns (e.g. walking). Each category is denoted as a binary vector in term of the presence (1) or absence (0) of each attribute. Attributes are capable to describe novel action. However, they need to be predefined manually and attribute-categories associations should be annotated in advance. When a novel category is joined, the variety of attributes and annotations of all the categories should be modified. Such limitations make attribute-based methods improper to large-scale recognition problems.

Word2vector space: An alternative semantic embedding space named word2vector space exactly overcomes above limitations. In word2vector-based methods [3][6][47], descriptions of labels are firstly learned using *Wikipedia* articles or *WordNet* [24]. Regressors [47] are then learned to map source data onto the word2vector space. A zero-shot strategy is realized by adopting those regressors to unseen data and evaluating similarity between embedded representations and label prototypes. Superior to attribute-based methods, word2vector-based methods only require the name of each category, but have no use for additional manual definitions of attributes and attribute-categories associations. However, as attributes describe fine-grained properties of each action, they are always more informational than word2vectors.

To evaluate the generality of our model, we implement it on both attribute and word2vector spaces in this paper.

Domain shift problem: Domain shift problem of ZSL was firstly discussed in [5], which seriously decreases models' recognition performance. As source domain and target domain are totally disjoint, It causes a great discrepancy between a specified prototype position and representations of its class member instances in semantic space (Fig. 1). To solve the problem, lots of ZSL models have been proposed. Fu *et al.* [5] proposed a self-training strategy to modify classes prototypes and align multi-view semantic spaces using canonical correlation analysis. Recently, transductive ZSL becomes an emerging topic, which is good at solving domain shift problem. Transductive ZSL extends traditional ZSL into a semi-supervised setting [30], in which full of target data are given. Kodirov *et al.* [13] proposed an unsupervised domain

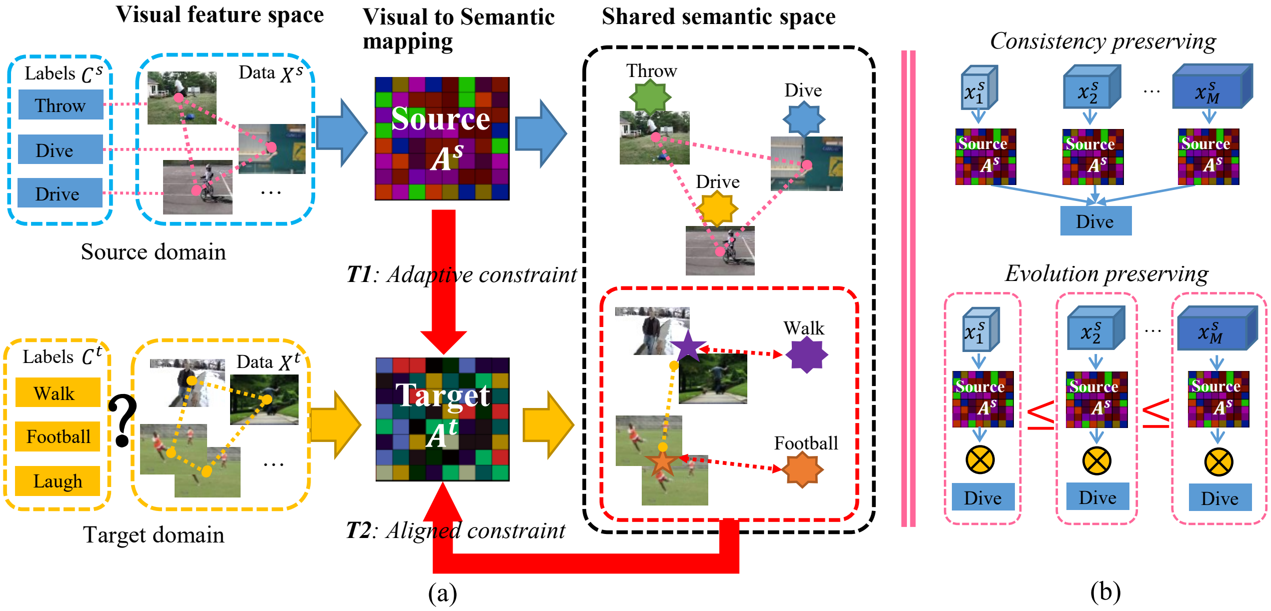


Fig. 3. Framework of our ADPE model. (a) In ZSL, source domain (X^s, C^s) and target domain (X^t, C^t) are disjoint, thereby causing domain shift problem. In our model, a target semantic mapping A^t is adaptively learned to solve the problem under the guidance of the mapping A^s learned from source domain. The alignment between target semantic representations and corresponding labels contributes to this learning as well. (b) In addition, our ADPE model takes temporal consistency and action evolution of videos into consideration. For each video, all its sequential segments are projected onto a same semantic representation. Meanwhile, the confidence of each segment being mapped to its true label is non-decreasing as much more frames are involved in the segment.

adaptation method, which uses target labels' projections to regularize the learned target domain projection with the aids of some auxiliary ZSL methods.

C. ZSL for action recognition

Despite promising progresses have been achieved in ZSL, it is relatively slow-growing for zero-shot action recognition. Liu *et al.* [21] firstly proposed human-specified attributes for actions, where classifier of each attribute was learned using source data, then novel categories were specified in terms of their scores of learned attribute classifiers. This work kicked off the research of zero-shot action recognition. Jain *et al.* [10] proposed a semantic embedding to classify actions in videos without using any video data or action annotations as prior knowledge, which is named as objects2action. It provided inspirations for the word2vector based methods for zero-shot action recognition. Xu *et al.* [47] firstly imported word2vector to describe action categories and embed videos and category labels to a common word2vector space via linear regression method. Gan *et al.* [6] aimed at exploiting inter-class relationships to facilitate zero-shot action recognition task. They directly leveraged semantic inter-class relations between seen and unseen actions followed by label transfer learning. In order to alleviate domain shift problem, Xu *et al.* [48] introduced a multi-task visual-semantic mapping, which could improve the generalization by constraining the semantic mapping parameters to lie on a low-dimensional manifold. Xu *et al.* [46] proposed a manifold regularized embedding for zero-shot action recognition in a transductive setting, which aims at preserving manifold structures of videos and further tackling domain shift problem. More recently, Qin *et al.* [29] adopted Error-Correcting Output Codes as label embedding for zero-shot action recognition, which implicitly captures semantic correlations among categories and intrinsic local structure of visual videos.

Especially, there always exist severer domain shift problem in video-based action recognition task [47] than traditional ZSL tasks. This is because that videos contains complex structure information, inter-class variations, especially abundant temporal information. However, most of the existing zero-shot action recognition methods could not take full advantage of videos' structures and relationships across categories. Furthermore, existing methods are always derived from objects recognition models, which don't take temporal information of video sequences into consideration.

D. Temporal models

Video sequences contain abundant *temporal information*, which is crucial to identity different action categories. To take advantage of those information, researchers proposed lots of temporal models in various applications. Hoai *et al.* [9] and Kong *et al.* [16][14] adopted Max-Margin structural SVM framework to capture temporal dynamics of video sequences to detect early event or predict actions. Niebles *et al.* [25] represented activities as temporal compositions of motion segments. Su *et al.* [36] proposed a hierarchical dynamic parsing model to parses the multi-layer dynamics of an action on different scales. Furthermore, temporal information is found important in sequences dimensionality reduction [33][34][35]. A max-min inter-sequence distance analysis method was proposed in [33], who aligns the sequences of the same class to temporal states and separates classes based on statistics of those ordered

TABLE I
PRIMARY SYMBOLS.

Symbols	Description
$X^s_m \in \mathbb{R}^{D_v \times N^s}$	N^s visual representations of the m_{th} segments of source videos.
$X^t_m \in \mathbb{R}^{D_v \times N^t}$	N^t visual representations of the m_{th} segments of target videos.
$Y^s \in \mathbb{R}^{D_w \times N^s}$	Semantic representations of N^s source data;
$Y^t \in \mathbb{R}^{D_w \times N^t}$	Semantic representations of N^t target data;
$P^s \in \mathbb{R}^{D_w \times K^s}$	K^s source action prototypes;
$P^t \in \mathbb{R}^{D_w \times K^t}$	K^t target action prototypes;
$A^s \in \mathbb{R}^{D_w \times D_v}$	Source visual-to-semantic embedding;
$A^t \in \mathbb{R}^{D_w \times D_v}$	Target visual-to-semantic embedding;
D_v	Dimension of the visual representations;
D_w	Dimension of the semantic representations;
$U \in \mathbb{R}^{N^t \times K^t}$	Proposed alignment matrix;
M	The number of temporal segments;
I	The number of iterations in Algorithm 1.

states. [34] learned linear sequence discriminant analysis models to project sequences to lower-dimensional subspace, which also address temporal dependency of sequences via extracting statistics of sequence classes. [35] proposed a latent temporal linear discriminant analysis method, which learns an abstract template for each class to discover the temporal structures via employing a modified DTW barycenter.

Motivated by references [16][14], we proposed a dynamic-preserving embedding model for zero-shot action recognition, in which temporal consistency and evolution of actions are captured. To the best of our knowledge, our model is the first one to characterize appearance information and temporal information of video sequences simultaneously in zero-shot action recognition task.

III. METHODOLOGY

Problem formulation and notations. A variable with a superscript s or t denotes that it is used in source (training) or target (testing) domain. Source domain is denoted as $G^s = \{X^s, Y^s, C^s, P^s\}$, $X^s = \{x^s_1, ..., x^s_{N^s}\}$ is the set of N^s training videos. All the training data belong to K^s action categories in the source domain $C^s = \{c^s_1, ..., c^s_{K^s}\}$, such as 'Walk', 'Drink', etc. Target domain is denoted as $G^t = \{X^t, Y^t, C^t, P^t\}$, $X^t = \{x^t_1, ..., x^t_{N^t}\}$ denote N^t target videos. The K^t novel categories $C^t = \{c^t_1, ..., c^t_{K^t}\}$ are disjoint from C^s: $C^t \cap C^s = \emptyset$. $P^s = \{p^s_1 ..., p^s_{K^s}\} \in \mathbb{R}^{D_w \times K^s}$ and $P^t = \{p^t_1 ..., p^t_{K^t}\} \in \mathbb{R}^{D_w \times K^t}$ are the D_w-dimensional semantic embeddings of seen and unseen labels, respectively, which are also called prototypes. $Y^s = \{y^s_1, ..., y^s_{N^s}\} \in R^{D_w \times N^s}$ and $Y^t = \{y^t_1, ..., y^t_{N^t}\} \in R^{D_w \times N^t}$ are the semantic representations of source and target data. In source domain, Y^s is known as the label of each source data is given. Y^s consist of P^s: $Y^s \supset P^s$. Our goal is to estimate the semantic embedding y^t_i of each target data x^t_i and infer its label c^t_i. For a better understanding, a list of primary symbols used in our model are summarized in Table I.

Visual representations of videos. Instead of encoding each video sequence into an orderless representation, which ignores temporal information of videos, we treat each video as several sequential segments in this work. As illustrated in Fig. 2, each video x^s_i (source and target domain) is uniformly divided into M ***fragments***. The length of each fragment is T/M. Note that for different videos, their length T may be different. Thus, the length of fragments of diverse videos may be different. Then, a set of successive ***segments*** are made up of those fragments. The m_{th} segment consists of the first m fragments, which is denoted as $x^s_i[1, m]$. In a similar fashion, the M_{th} segment is actual the whole video sequence, which contains all the M fragments. Eventually, the video x^s_i is denoted as a set of sequential segments: $x^s_i = \{x^s_i[1, 1], ..., x^s_i[1, m], ..., x^s_i[1, M]\}$. Each of the segment is then encoded into a D_v-dimensional visual vector via traditional video encoding methods. For a convenience, the visual representation of m_{th} segment $x^s_i[1, m]$ is briefly denoted as x^s_{im} in the followings.

Shared semantic embedding spaces. In our framework, all the action labels (C^s, C^t) are firstly embedded into a semantic space. These semantic projections of labels are also called label prototypes (P^s, P^t). In order to demonstrate the generalization of our model, both attribute and word2vector semantic spaces are selected in this paper. Performances of the two spaces will be discussed in experiments. For attribute space, P^s and P^t are sets of binary vectors. Each element of the attribute vector corresponds to a manual attribute, e.g. 'torso'. Its value (1/0) denotes presence/absence of the attribute in a specified action. For word2vector space, following [47], we use a word2vector neural network [23] trained on a 100 billion word corpus to map each word to a D_w-dimensional vector. For multi-words action labels, average vector [11] of multi embedded words is taken as a final description.

A. Aligned dynamic-preserving embedding

Once we get the semantic projections of labels, two kinds of visual-to-semantic mappings (A^s, A^t) for both source and target domain are then learned, respectively. For most of existing methods [47][48][46], the visual-to-semantic embedding A^s of source domain is learned via a ridge regression:

$$\min_{A^s} \|A^s X^s - Y^s\|^2_F + \beta \|A^s\|^2_F, \tag{1}$$

where $\|.\|_F$ is Frobenius norm of a matrix. $\|A^s\|^2_F$ is a smooth constraint and β controls its strength. A closed-form solution of A^s can be obtained:

$$A^s = Y^s (X^s)^T (X^s (X^s)^T + \beta I)^{-1}. \tag{2}$$

Target data X^t are then directly mapped onto the unified semantic space via the learned visual-semantic mapping A^s: $Y^t = A^s X^t$.

As discussed in Section I, there are three main limitations of the traditional framework, which leads to sever domain shift problem. Firstly, it doesn't capture temporal information of video sequences during its processing. Secondly, it applies the semantic mapping that is separately learned from source domain for target domain without any adaptation. Thirdly, it doesn't take variations across action categories into consideration, in which the knowledge transferred from source domain are insufficient to describe novel categories.

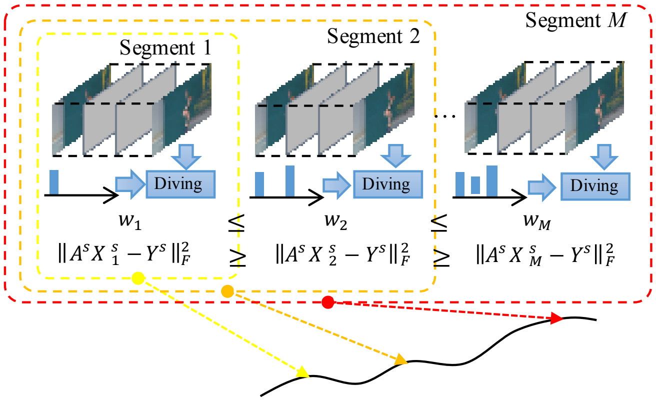


Fig. 4. Graphical illustration of the preserving of temporal consistency of each segment and temporal action evolution over time of each video in our ADPE model. On the one hand, each segment is projected onto the same semantic prototype. On the other hand, the segments with more observed frames get higher similarity with the ground-truth prototype.

In order to overcome those limitations and further alleviate domain shift problem in zero-shot action recognition, we propose an aligned dynamic-preserving embedding (ADPE) model. In the followings, we introduce our framework in terms of the learning of source and target domain, respectively.

*1) **Source domain**:* In our model, temporal information of video sequences are taken into consideration. Instead of encoding each video into an orderless vector and mapping it into a fixed semantic representation, we deal with each video segment by segment and propose a temporal dynamic-preserving embedding. The projection of source domain is formulated as follows:

$$\begin{aligned} &\min_{A^s} \sum_{m=1}^{M} w_m \|A^s X^s_m - Y^s\|_F^2 + \beta \|A^s\|_F^2 \\ &\text{s.t.} \;\; w_m = f(m), \sum_{m=1}^{M} f(m) = 1, \end{aligned} \tag{3}$$

X^s_m constitutes of the m_{th} segments of all source data. The quadric error ($\|A^s X^s_m - Y^s\|_F^2$) measures compatibility between the projected semantic embedding of m_{th} segment ($A^s X^s_m$) and the video's true label prototype (Y^s).

In order to take advantages of temporal information of video sequences, our embedding model follows two temporal mapping criterions.

- Firstly, we project all the sequential segments ($x^s_{im}, \forall m$) of a specified video onto the same semantic representation (y^s_i), as shown in Fig. 4. This criterion encourages temporal ***label consistency*** between segments and the corresponding full video. Context information of segments is also implicitly captured by enforcing the label consistency.
- Secondly, we ensure that the confidence of each segment being mapped to its true label is non-decreasing as much more frames are involved in the segment. In other words, the quadric error of m_{th} segment should be smaller than or equal to the error of $(m-1)_{th}$ segment: $\|A^s X^s_{m-1} - Y^s\|_F^2 \geq \|A^s X^s_m - Y^s\|_F^2$.
  To realize this criterion, we apply a temporal factor w_m to assign different weights to the mapping of different segments. w_m is defined as a positive non-decreasing function of m in this paper: $w_m = f(m), f(m_1) \geq f(m_2), \forall m_1 \geq m_2$. Recall our objective function (Eq. 3), which aims at minimizing a summation of all the segments' quadric errors. As w_m is larger than or equal to w_{m-1}, the quadric error of $\|A^s X^s_m - Y^s\|_F^2$ is expected to contribute a larger proportion of all the segments' quadric errors than $\|A^s X^s_{m-1} - Y^s\|_F^2$. Accordingly, larger penalty falls on the m_{th} segment than the $m-1_{th}$ one, which promotes it being more tend to own smaller quadric error ($\|A^s X^s_{m-1} - Y^s\|_F^2 \geq \|A^s X^s_m - Y^s\|_F^2$). Thus, the temporal factor w_m helps the segment with much more frames to be mapped into a more convincing semantic representation. Our embedding not only captures action appearance changing with segment increasing, but also elaborately characterizes the nature of sequentially arriving action data. The learned embedding therefore considers ***temporal action evolution*** of video sequence over time.

In summary, our dynamic-preserving model makes full use of temporal information of video sequence via considering temporal consistency of each segment and capturing temporal action evolution over time of video simultaneously.

We can derive a closed-form solution of A^s in Eq. 4. Let its partial derivative with respect to A^s equal to zero. We can easily get:

$$A^s = \sum_{m=1}^{M} w_m Y^s (X^s_m)^T \Big(\sum_{m=1}^{M} w_m X^s_m (X^s_m)^T + \beta I \Big)^{-1}, \tag{4}$$

where I is an unit matrix.

*2) **Target domain**:* Once A^s is obtained, most of existing ZSL methods directly apply it to target domain, resulting in so-called domain shift problem. Instead, we learn an adaptive mapping A^t using both source and target data and propose a novel aligned dynamic-preserving embedding (ADPE) method. ADPE not only exploits underlying structures of source and target data, but also takes inter-class variations into consideration. In target domain, we also divide each target video into a series of successive segments and follow the two temporal mapping criterions as used in source domain. The objective function of target domain is formulated as Eq. 5:

$$\begin{aligned} &\min_{A^t, Y^t, U} \sum_{m=1}^{M} w_m \|A^t X^t_m - Y^t\|_F^2 + \lambda_1 T_1 + \lambda_2 T_2 \\ &\text{s.t.} \;\; \|u_j\|_2^2 = 1, \; T_1 = \|A^s - A^t\|_F^2, \; T_2 = \|Y^t U - P^t\|_F^2, \end{aligned} \tag{5}$$

where $Y^t \in R^{D_w \times N^t}$ is the learned semantic embedding of target data. T_1 and T_2 are our proposed adaptive regularization term and aligned regularization term, respectively. U is an alignment matrix to couple learned semantic representations of target data with corresponding label prototypes. The detail interpretations of each component of ADPE are discussed as follows.

Adaptive regularization term, $\|A^t - A^s\|_F^2$. Although source and target domains are totally disjoint, the mapping matrix A^s that is learned from source data carries latent semantic information across seen and unseen categories. On the one hand, we expect to preserve the shared information that

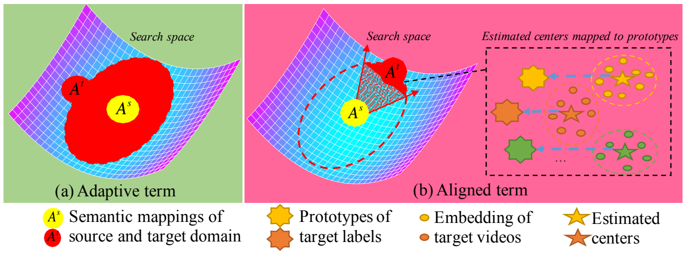


Fig. 5. Illustration of components in our objective function. The red shades denote the search space of A^t. (a) Adaptive term limits the learning of A^t in the neighborhood of A^s; (b) Aligned term ensures the alignment between semantic embedding of target data and label prototypes. It gives a more discriminative and narrow direction to the learning of A^t.

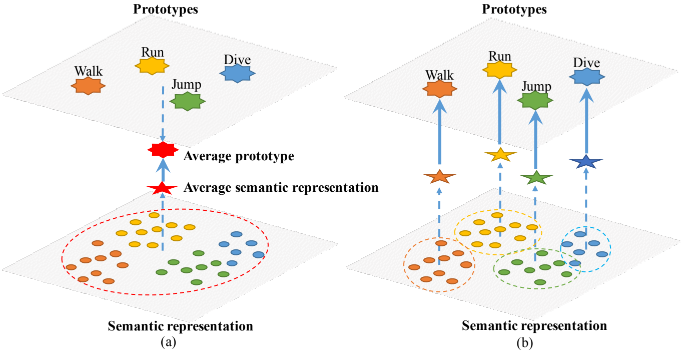


Fig. 6. Illustration of aligned term. (a) MMD term: Average vectors of action prototypes and all the target semantic representations are calculated, respectively. Distributions difference is shorten via prompting the two average vectors to be similar. (b) Our aligned term: Estimated centers of target semantic representations are aligned with corresponding action prototypes class wisely, during which variations across categories are preserved.

is carried in A^s. On the other hand, the underlying structure of target data also needs to be captured. Thus, the adaptive regularization term is proposed to enforce target mapping A^t to be close to source mapping A^s. As illustrated in Fig. 5(a), the adaptive term guides the searching of A^t to be limited in a neighborhood of A^s. In fact, the adaptive regularization term avoids A^t be any arbitrary or unreasonable values. It guarantees the reliability of our model.

Aligned regularization term, $\|Y^tU - P^t\|_F^2$. To further overcome domain shift problem and improve discriminability of learned representations Y^t, one may expect that the distributions of learned semantic representations Y^t of target data and the true label prototypes P^t should be similar to each other in the semantic space. This idea can be naturally implemented by reducing distributions difference between Y^t and P^t or minimizing some predefined distance measures. However, as the dimensions of Y^t and P^t are different, we can not directly measure biunique similarity between them. To solve the problem, Long *et al.* [22] adopted empirical Maximum Mean Discrepancy (MMD) to compare the two distributions, which computes distance between sample means of the learned semantic representations of testing videos and known action prototypes in the shared semantic space (Fig. 6(a)). Transplant to our task, the aligned regularization term could be formulated as Eq. 6 [37].

$$\begin{aligned} &T_{2(\mathrm{MMD})} = \|\tilde{Y^t} - \tilde{P^t}\|_F^2, \\ &\tilde{Y^t} = (\textstyle\sum_{i=1}^{N^t} y_i^t)/N^t, \tilde{P^t} = (\textstyle\sum_{j=1}^{K^t} p_j^t)/K^t, \end{aligned} \tag{6}$$

where N^t and K^t denote the number of testing samples and prototypes, respectively. The regularization term that formulated via MMD measurement actually aligns their average vectors instead of the two distributions (Y^t and P^t). However, for zero-shot action recognition, there exist large inter-class variations among different categories. MMD term ignores the inter-class variations and erases characters of each category.

In order to take the inter-class variations into account and learn an effective mapping A^t, we propose a nonparametric distance measurement $\|Y^tU - P^t\|_F^2$, which is called aligned regularization. In this term, an alignment matrix U is proposed to align the center of each category of target semantic representations with its estimated label prototype (Fig. 6(b)). U is defined in Eq. 7. Each element u_{ij} in U is the probability that each learned semantic embedding y_i^t belongs to a specified category p_j^t:

$$u_{ij} = \mathrm{p}(y_i^t|p_j^t) \subset U. \tag{7}$$

The semantic representations of target data Y^t multiplied by the j_{th} column of U can be derived in Eq. 8:

$$Y^tU_j = \sum_i y_i^t u_{ij} = \sum_i y_i^t \mathrm{p}(y_i^t|p_j^t). \tag{8}$$

It can be observed that Y^tU_j is actually the mean value of the semantic representations that are estimated to be the j_{th} category. In terms of our aligned regularization, the estimated center Y^tU_j of the j_{th} category is enforced to be close to its corresponding label prototype p_j^t. Thus, centers of semantic representations are aligned with label prototypes P^t class by class (Fig. 6(b)). As we deal with the alignment class-wisely, character of each category and inter-class variations among different categories are preserved. Compared to adaptive regularization term, aligned regularization term offers a discriminative direction to the optimization of A^t in the neighborhood of A^s, which guarantees the practicability of our model (Fig. 5(b)).

In summary, our ADPE model not only captures the commonality underlying both source and target data, but also preserves inter-classes variations across categories. Thus, domain shift problem can be significantly reduced via our model.

B. Optimization

Our objective function of target domain is not convex for Y^t, A^t and U simultaneously, but it is convex for each of them while the others are fixed. Thus, Alternating Direction Method (ADM) [1] is applied to solve the optimization problem.

Fixing A^t and U, update Y^t. Our objective function turns to be Eq. 9,

$$\min_{Y^t} \sum_{m=1}^{M} w_m \|A^tX_m^t - Y^t\|_F^2 + \lambda_2 \|Y^tU - P^t\|_F^2 \tag{9}$$

This is a least square problem and we can easily get a closed-form solution. A^t is initialized via A^s. U is initialized randomly. To obtain Y^t, the partial derivative of function Eq.

Algorithm 1 Aligned semantic embedding

```
Input:
Source data $X_m^s, m \in \{1, ..., M\}$; Source semantic representations $Y^s$ .
Target data $X^t, m \in \{1, ..., M\}$; Target labels prototypes $P^t$.
Output: Semantic embedding $Y^t$ and target embedding $A^t$
Source domain:
1. Calculate $A^s$ by Eq. 4
Target domain:
1. Initialize: $A^t$ by $A^s$; $U$ randomly
2. repeat
3.     Update $Y^t$ by Eq. 11;
4.     Update $A^t$ by Eq. 14;
5.     Update $U$ by Eq. 15 and Eq. 16;
6. until converge.
```

9 with respect to Y^t should be zero and could be expressed as Eq. 10.

$$\sum_{m=1}^{M} w_m A^t X_m^t - \sum_{m=1}^{M} w_m Y^t - \lambda_2 Y^t U U^T + \lambda_2 P^t U^T = 0. \tag{10}$$

Then, we can get a closed-form solution of Y^t,

$$Y^t = (\lambda_2 P^t U^T + \sum_{m=1}^{M} w_m A^t X_m^t)(\lambda_2 U U^T + \sum_{m=1}^{M} w_m I)^{-1}, \tag{11}$$

where I is an unit matrix.

Fixing Y^t and U, update A^t. Our objective function (Eq. 5) turns to be Eq. 12,

$$\min_{A^t} \sum_{m=1}^{M} w_m \|A^t X_m^t - Y^t\|_F^2 + \lambda_1 \|A^s - A^t\|_F^2. \tag{12}$$

We can also calculate the derivative of it with regard to A^t as Eq. 13 and get a closed-form solution as Eq. 14,

$$\begin{aligned} &\sum_{m=1}^{M} w_m A^t X_m^t (X_m^t)^T - \sum_{m=1}^{M} w_m Y^t (X_m^t)^T + \lambda_1 A^t \\ &-\lambda_1 A^s = 0. \end{aligned} \tag{13}$$

$$\begin{aligned} &A^t = \\ &(\lambda_1 A^s + \sum_{m=1}^{M} w_m Y^t (X_m^t)^T)(\sum_{m=1}^{M} w_m X_m^t (X_m^t)^T + \lambda_1 I)^{-1}. \end{aligned} \tag{14}$$

Fixing Y^t and A^t, update U. According to the definition of U, each element u_{ij} can be calculated on the basis of Bayes' theorem as Eq. 15.

$$u_{ij} = \mathrm{p}(y_i^t | p_j^t) = \frac{\mathrm{p}(p_j^t | y_i^t) \cdot \mathrm{p}(y_i^t)}{\sum_k \mathrm{p}(p_j^t | y_k^t) \cdot \mathrm{p}(y_k^t)}, \tag{15}$$

where $\mathrm{p}(p_j^t | y_i^t)$ is the probability of a specified semantic representation y_i^t belonging category $p_j^t, \forall j$. We assume x_i conforming to an uniform distribution. And $\mathrm{p}(p_j^t | y_i^t)$ can be calculated as Eq. 16.

$$\mathrm{p}(p_j^t | y_i^t) = \begin{cases} 1, & \text{if } j = \arg\min_j \|y_i^t - p_j^t\|_F^2 \\ 0, & \text{otherwise} \end{cases}. \tag{16}$$

It can be observed from Eq. 16 that if p_j^t is the nearest prototype of y_i^t, the probability of it belonging to category p_j^t is 1. We can also adopt some soft methods to estimate $\mathrm{p}(p_j^t | y_i^t)$ according to the similarity between y_i^t and each prototype. But the hard estimation is sufficient to demonstrate the effectiveness of our method. The whole optimization algorithm of our method is summarized in Algorithm 1.

Zero-shot classification. Once the semantic representations Y^t of target data are obtained, classification is performed in the learned semantic space. The most common classifier is nearest prototype classifier. Given an obtained embedding y_i^t, its label is the nearest prototype as Eq. 17.

$$c(y_i^t) = \arg\min_k \|y_i^t - p_k^t\|. \tag{17}$$

Computational complexity. We analyze computational complexities of our method in source and target domains, respectively. We assume the simplest and worst matrix multiplication algorithm is applied. 1) Source domain: The complexity of calculating A^s (Eq. 4) is $O(MD_w N^s D_v + MN^s D_v^2 + D_w D_v^2 + D_v^3)$. As M is a low-dimensional fixed constant, we can omit its influence. The complexity can thus be simplified as $O(\mathbb{C}^2 N^s + \mathbb{C}^3)$, where $\mathbb{C}$ represents the dimension of visual or semantic representations (D_v or D_w).

2) Target domain. We also omit the influence of M and K^t as they are low-dimensional fixed constants. The complexity of updating Y^t (Eq. 11) is $O(D_w K^t N^t + MD_w D_v N^t + K^t {N^t}^2 + D_w {N^t}^2 + {N^t}^3)$, whose simplified form is $O({N^t}^3 + {N^t}^2 + \mathbb{C}{N^t}^2 + \mathbb{C}^2 N^t)$; The complexity of updating A^t (Eq. 14) is $O(MD_w N^t D_v + MN^t D_v^2 + D_w D_v^2 + D_v^3)$, whose simplified form is $O(\mathbb{C}^2 N^t + \mathbb{C}^3)$; The complexity of updating U (Eq. 16) is $O(N^t K^t)$, whose simplified form is $O(N^t)$. In general, the worst case of total computational complexity in target domain is I times of the summation of three-step updating, whose simplified form is $O(I({N^t}^3 + {N^t}^2 + \mathbb{C}{N^t}^2 + \mathbb{C}^2 N^t + N^t + \mathbb{C}^3))$. Therefore, the computation of our model is a three-order computation.

IV. EXPERIMENTS

A. Datasets and settings

We apply our novel zero-shot action recognition method (ADPE) on three challenging action datasets, Olympic sports [26], HMDB51 [17] and UCF101 [32], which have been popularly used in human action recognition task.

Olympic sports dataset consists of 783 videos, which are divided into 16 sports actions. HMDB51 dataset is a large collection of realistic videos from various sources, mostly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos. It contains 6766 realistic videos distributed in 51 actions. UCF101 is also a set of realistic action videos, collected from YouTube. With 13320 videos from 101 action categories, UCF101 gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, etc.

Because there is no existing zero-shot learning evaluation protocol for action recognition. We follow the rule of split in [47], in which 50% categories of videos are used for training and the other 50% for testing. We randomly generate 10 independent splits for each dataset. And the mean accuracy (%) and standard deviation are reported for performance.

Visual representations. In this work, we uniformly divide a full video into 5 fragments. Then 5 successive segments

TABLE II
Comparison between our method and state-of-the-art methods on Olympic, HMDB51 and UCF101 datasets. Label embedding: Attribute (A) or Word2Vector (W). Trans: Accessible to unseen videos. Splits: The number of splits; 'S': The applied splits are the same with ours; 'D': The applied splits are different with ours. '*': The results reported in their original papers. Accuracy (%) and standard deviation are reported for performance.

Methods	**Reference**	**Embed**	**Splits**	**Trans**	**Olympic**	**HMDB51**	**UCF101**
Random guess	—-	—-	—-	—-	12.5	4.0	2.0
HAA [21]	CVPR 2011	A	10 / S	No	52.74 ± 8.16	N/A	9.76 ± 1.58
IAP [19]	TPAMI 2014	A	10 / S	No	50.87 ± 7.11	N/A	9.95 ± 1.26
DAP [19]	TPAMI 2014	A	10 / S	No	52.48 ± 9.76	N/A	10.39 ± 1.56
RR [47]	ICIP 2015	A	10 / S	No	55.18 ± 8.87	N/A	15.61 ± 1.12
SIR [6]	AAAI 2015	A	10 / S	No	46.45 ± 11.58	N/A	7.57 ± 1.77
UDA* [13]	ICCV 2015	A	10 / D	Yes	N/A	N/A	13.20 ± 1.90
MTE* [48]	ECCV 2016	A	5 / D	No	55.60 ± 11.30	N/A	18.30 ± 1.70
MTE [48]	ECCV 2016	A	10 / S	No	55.23 ± 8.93	N/A	11.80 ± 1.00
MR* [46]	IJCV 2017	A	30 / D	Yes	53.50 ± 11.90	N/A	20.20 ± 2.20
MR [46]	IJCV 2017	A	10 / S	Yes	55.15 ± 9.17	N/A	15.10 ± 1.07
RR* [47]	ICIP 2015	W	30 / D	No	N/A	13.00 ± 2.70	10.90 ± 1.50
RR [47]	ICIP 2015	W	10 / S	No	30.93 ± 5.76	13.40 ± 1.11	11.68 ± 0.99
MTE* [48]	ECCV 2016	W	5 / D	No	**44.30 ± 8.10**	19.70 ± 1.60	15.80 ± 1.30
MTE [48]	ECCV 2016	W	10 / S	No	31.21 ± 5.80	13.22 ± 2.30	11.80 ± 1.00
MR* [46]	IJCV 2017	W	30 / D	Yes	38.60 ± 10.60	19.10 ± 3.80	**18.00 ± 2.70**
MR [46]	IJCV 2017	W	10 / S	Yes	30.77 ± 5.45	15.90 ± 2.15	11.66 ± 0.98
UDA* [13]	ICCV 2015	A+W	10 / D	Yes	N/A	N/A	14.00 ± 1.80
ZSECOC* [29]	CVPR 2017	ECOC	10 / D	No	59.80 ± 5.60	**22.60 ± 1.20**	15.10 ± 1.70
ADPE	Ours	A	10	Yes	**60.13 ± 6.01**	N/A	**20.56 ± 1.09**
ADPE	Ours	W	10	Yes	38.55 ± 5.11	17.72 ± 1.26	14.03 ± 0.66

are made up of those fragments as introduced in Section III. For example, the 2_{th} segment consists of the 1_{th} and 2_{th} fragments, the 3_{th} segment consists of the 1_{th}, 2_{th} and 3_{th} fragments, and so on. For each segment, we extract the C3D features [39]. After that, each segment is denoted as a 4096-dimensional mean vector of a set of extracted C3D features.

Label semantic embedding. We apply both word2vector and attribute spaces in our experiments. For attribute space, each label is represented as a binary attribute vector. Since the attributes' annotations are not available for HMDB51 dataset, we conduct those experiments on Olympic and UCF101 datasets. 40 and 115 hand-annotated attributes and attribute-categories associations are provided for the two datasets, respectively. For word2vector space, the skip-gram neural network [23] trained on Google News dataset is adopted to encode each action label to a 300-dimensional vector. For multi-words label, we obtain a fused vector by averaging them.

B. Comparison with the state-of-the-arts

To demonstrate the capabilities and generalization of the proposed method, we firstly compare it with several contemporary and related zero-shot learning methods using both word2vector and attribute spaces: (1) Human Actions by Attributes (HAA) method [21]; (2) Direct/Indirect Attribute Prediction (DAP/IAP) methods [19]; (3) Original Ridge Regression (RR) method [47]; (4) Semantic Inter-Class Relationships (SIR) method [6]; (5) Unsupervised Domain Adaptation (UDA) method [13]; (6) Multi-Task Embedding (MTE) method [48]. (7) Manifold Regression (MR) method [46]; (8) Error-Correcting Output Codes (ZSECOC) method [29].

We notice that some compared methods (e.g. RR, MR and MTE) developed some self-training and data augmentation techniques to improve their final performance. As our goal focuses on learning effective visual-to-semantic embedding, we therefore compare with their results without adopting those self-training and data augmentation settings. We mark their native results copied from corresponding papers with '*' in Table II. Besides, for fair comparisons, we re-implement all the compared methods using the same settings with ours (eg. 10 splits and C3D features) and fine tune their parameters to obtain the best performance.

*1) **Attribute space:*** Firstly, we compare our model with several state-of-the-art methods using attribute space. Average accuracies (%) and standard deviation of those methods are reported in Table II. It shows that our method achieves the best performance on Olympic and UCF101 datasets. HAA [21] learns a classifier for each attribute, after which unseen categories are determined in terms of scores of learned attribute classifiers. This method treats all the attributes independently and completely ignores relationships across videos. Additionally, its complexity is in direct proportion to the number of attributes, which is inapplicable for large datasets. IAP and DAP [19] methods also apply the attribute classifiers to recognize unseen categories. Those methods directly use the attribute classifiers learned from source data on target domain without any adaptation, and thus perform poorly on complex action datasets. Ridge regression (RR) [47] is the

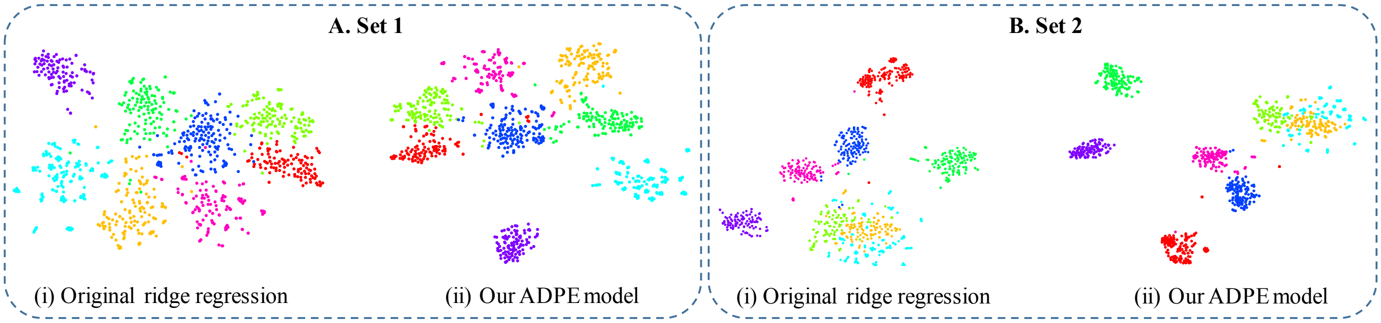


Fig. 7. Visualization of semantic representations in an attribute-based space via (i)traditional ridge regression and (ii) our aligned dynamic-preserving embedding (ADPE) model. Two random sets of target classes are presented for diversity: A. Set 1 and B Set 2. Dots denote semantic representations of samples and different colours indicate different action categories (Better in colour).

traditional ZSL method that directly adopts the regressor learned by source data on target domain. It always suffers from severe domain shift problem. Different from it, our method adaptively learns a semantic mapping via exploring structures of both source and target data. This allows us to efficiently solve the domain shift problem and achieve significant improvements. SIR [6] exploits semantic inter-class relationships between actions to transfers knowledge from unseen categories to seen ones. However, the inter-class relationships couldn't describe the diversity of unseen actions. In contrast, the aligned regularization term in our ADPE model is typically designed to collect inter-class information across various unseen categories, which achieves better performance than SIR on both Olympic and UCF101 datasets. Kodirov [13] exploits an unsupervised domain adaptation (UDA) to overcome domain shift problem, which is similar with our method. Their model is actually a two-step framework, in which some auxiliary ZSL methods need to be adopted to pre-compute the probability of each target data being labeled as each novel category. Thus, its performance is easily affected by the selections of different auxiliary ZSL methods. In contrast, our ADPE model is an unified optimization framework, which not only self-adaptively learns a visual-semantic mapping of target data without any auxiliary methods, but also takes inter-class variations into consideration. MTE model [48] uses a multi-task semantic mapping to improve generalization power of visual-to-semantic mapping by constraining its parameters to lie on a low-dimensional manifold. Compared to MTE, our method gains significant improvements on the two datasets, respectively. Manifold regression (MR) method [46] uses a manifold regularization to capture structures of target data. However, it does't establish the connections between semantic representations of target data and label prototypes. Besides, inter-class variations are always ignored during their processing. Superior to MR, our method ensures that the distributions of learned representations of target data Y^t are similar to their corresponding label prototypes'. Moreover, we preserve relationships across different categories. ZSECOC [29] is a recent method, which adopts Error-Correcting Output Codes as label embedding for zero-shot action recognition. It implicitly captures semantic correlations among categories and intrinsic local structure of visual videos. ZSECOC achieves competitive results with ours on Olympic and HMDB51 datasets.

However, all the mentioned methods in Table II are originally desired for objects recognition in images, which ignore the long-term temporal information underlying video sequences. Expressly, superior to them, our model characterizes temporal information of video sequences via preserving temporal consistency and temporal evolutions of actions.

*2) **Word2vector space**:* Our embedding method is also efficient in recognizing unseen actions using word2vector space. We further compare it with several popular word2vector-based methods on the three datasets. Results are shown in Table II. Our model also achieves better performance than the baseline RR method using word2vector space. The performances of MR and MTE methods reported in their papers using word2vector space are better than ours on the three datasets. Nevertheless, when we re-conduct their experiments using the same settings (eg. same splits and visual representations), we can obtain improvements over them on the three datasets.

In summary, our embedding model is applicable for both word2vector and attribute semantic spaces.

TABLE III
COMPARISONS BETWEEN THE EMBEDDINGS LEARNED FROM SOURCE (A^s) AND TARGET (A^t) DOMAIN ON OLYMPIC SPORT, HMDB51 AND UCF101 DATASETS. LABEL EMBEDDING: ATTRIBUTE (A) OR WORD2VECTOR (W). ACCURACY (%) AND STANDARD DEVIATION ARE REPORTED FOR PERFORMANCE.

Methods	**Olympic (A)**	**HMDB51 (W)**	**UCF101 (A)**
RR-A^s	55.18 ± 8.87	13.40 ± 1.11	15.61 ± 1.12
Ours-A^s	55.22 ± 9.00	15.56 ± 3.44	15.79 ± 1.26
Ours-A^t	**60.13 ± 6.01**	**17.72 ± 1.26**	**20.56 ± 1.09**

*3) **Comparisons between** A^s **and** A^t:* To better demonstrate that our model can better tackle domain shift problem, we directly apply the projection A^s learned in the source domain to the target domain. The results on Olympic sport, HMDB51 and UCF101 datasets are reported in Table III.

To avoid confusion, 'RR-A^s' denotes the visual-to-semantic embedding learned in source domain via Original Ridge Regression (RR) method. Compared to 'RR-A^s', 'Ours-A^s' is the source embedding obtained via our ADPE method (Eq. 3, Eq. 4), which takes the temporal information of source videos into consideration. 'Ours-A^t' is the target embedding learned via ADPE method (Eq. 5, Eq. 14), which not only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2908487, IEEE Transactions on Circuits and Systems for Video Technology

captures temporal information of videos, but also incorporates the proposed adaptive and aligned constraints. All the three embeddings are adopted to target domain, respectively, whose performance is reported in Table III. From the experiments we can find that, ‘Ours-A^s’ gets higher accuracies than ‘RR-A^s’ on the three datasets, which demonstrates the importance of temporal information. Superior to ‘Ours-A^s’, ‘Ours-A^t’ achieves significant rises of accuracies on the three datasets. The source embedding A^t learned via our ADPE model takes full advantage of underlying structures of both source and target videos and takes the relationships across different categories into account, which effectively alleviates the domain shift problem and improves the performance of zero-shot action recognition. In the followings, we will discuss the properties of our method in detail.

4) ***Qualitative Visualization:*** In this section, we further analyze the effectiveness of our APDE model via qualitative visualizations. For the visualizations, we randomly sample 8 target classes from UCF101 dataset and project all samples from these classes into an attribute semantic space by (i) traditional ridge regression (RR) model and (ii) our ADPE model. The semantic representations are visualized in 2D using t-SNE toolbox [40]. Two random sets of target classes are presented for diversity, as shown in Fig. 7. Data instances are shown as dots. Different colours indicate different action categories. We can observe from Fig. 7 that our ADPE model yields better visual semantic projections than the baseline method (RR). ADPE not only obtains much tighter clusters in the semantic space, but also yields more separable clusters. In other words, ADPE can reduce intra-classes distances and increase inter-classes distances, which alleviates domain shift problem effectively.

TABLE IV

COMPARISON BETWEEN THE MODELS THAT CAPTURES TEMPORAL INFORMATION (RR-WITHTIME AND ADPE) AND THOSE WITHOUT TEMPORAL PROPERTY (RR AND ADPE-NOTIME). ACCURACY (%) AND STANDARD DEVIATION ARE REPORTED FOR PERFORMANCE.

Methods	Label embedding	UCF101
RR	Attribute	15.61 ± 1.12
RR-WithTime	Attribute	**16.00 ± 1.13**
RR	Word2vector	11.68 ± 0.99
RR-WithTime	Word2vector	**11.94 ± 1.13**
ADPE-NoTime	Attribute	18.18 ±1.42
ADPE (Ours)	Attribute	**20.56 ± 1.09**
ADPE-NoTime	Word2vector	12.90 ± 0.72
ADPE (Ours)	Word2vector	**14.03 ± 0.66**

C. Temporal dynamic-preserving property

Superior to existing methods, our ADPE model captures temporal information of video sequences via preserving temporal dynamics of video sequences. In order to demonstrate the superiority of dynamic-preserving property of our ADPE model, we evaluate it in two aspects.

1) ***Evaluation of dynamic-preserving property:*** Firstly, we compare our model with several variants that don’t take temporal information of videos into consideration. By removing the dynamic-preserving property, the objective function of our ADPE model degenerates into Eq. 18.

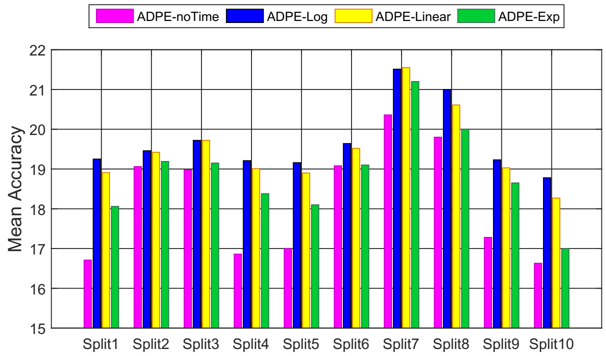


Fig. 8. Evaluation of different dynamic-preserving functions of each split on UCF101 dataset using attribute space.

$$
\begin{aligned}
&\textit{Source domain:}\\
&\min_{A^s} \|A^s X^s - Y^s\|_F^2 + \beta \|A^s\|_F^2,\\
&\textit{Target domain:}\\
&\min_{A^t, Y^t, U} \|A^t X^t - Y^t\|_F^2 + \lambda_1 \|A^s - A^t\|_F^2 + \lambda_2 \|Y^t U - P^t\|_F^2\\
&\text{s.t. } \|u_j\|_2^2 = 1,
\end{aligned}
\tag{18}
$$

where each video sequence is presented as a D_v-dimensional vector. Compared to our ADPE model, the degenerated model (denoted as ADPE-NoTime in the followings) ignores temporal dynamics of actions during their visual-to-semantic embedding. The results of ADPE-NoTime and ADPE (ours) are reported in Table IV. Besides, to better explain the superiority of dynamic-preserving property, we apply our time model on the baseline RR method [47] (RR-WithTime), and compare it with original RR in Table IV. Obviously, no matter what semantic space we use (attribute or word2vector), the models that consider temporal dynamics of videos achieve better performance than those models that ignore these information. It demonstrates that temporal information of video sequences is helpful to alleviate domain shift problem and important to improve the performance of zero-shot action recognition. Our embedding model can exactly capture these information via preserving temporal consistency of videos and temporal evolutions of actions

2) ***Evaluations of dynamic-preserving functions:*** As explained in Section III, we apply a temporal factor w_m to control the quality of visual-to-semantic projection of each segment, which is defined as an positive non-decreasing function of m in this paper ($w_m = f(m)$). In the experiment, we evaluate the performance of our model using different non-decreasing functions, namely linear function (ADPE-Linear), logarithmic function (ADPE-Log) and exponential function (ADPE-Exp). For those functions, w_m can be expressed as $w_m = m/(\sum_{m=1}^{M} m)$, $w_m = \log^m/(\sum_{m=1}^{M} \log^m)$ and $w_m = e^m/(\sum_{m=1}^{M} e^m)$, respectively. We record the result of each split on UCF101 dataset using those functions in Fig 8. As shown in the figure, the logarithmic function outperforms the linear and

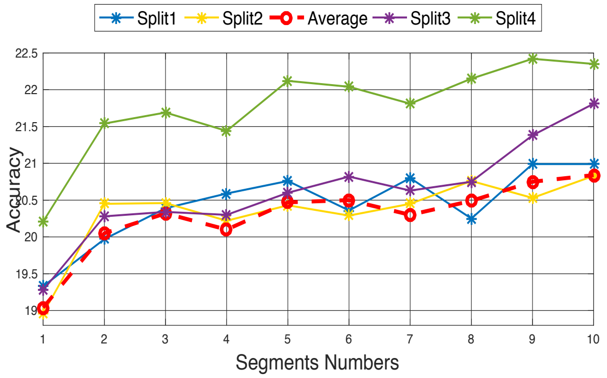


Fig. 9. Evaluations of the number of segments on UCF101 dataset using attribute space.

exponential functions. Nevertheless, no matter which function we use, our ADPE model with dynamic-preserving property performs better than the one that doesn't take temporal information into consideration (ADPE-noTime). It further proves that the temporal dynamic information benefits ZSL action recognition task.

*3) **Evaluation of segmentation:*** In order to further evaluate the temporal property of our model, we divide each video into different number of segments ($M = 1, 2, ..., 9, 10$). We randomly select several splits of UCF101 dataset and record their accuracy in the case of different number of temporal segments (Fig. 9). From the figure we can conclude that, when we treat the whole video as one segment, the accuracy of each split drops off. It further proves that each video contains abundant temporal information, which plays important roles in recognizing human actions. If we treat a video as one segment, its internal temporal information will be disregarded. When we divide each video into several segments and capture temporal dynamics and consistency of them, the performance is significantly improved. It can be also concluded from the figures that with more temporal segments that a video has been divided, the higher recognition accuracy we can obtain. A possible reason is that we can preserve detailed temporal information when we divide videos into refined segments. In order to get a tradeoff between accuracy and efficiency, we divide each video (Olympic sports, HMDB51 and UCF101 dataset) into 5 temporal segments in our experiments.

*4) **Analysis of intermediate results and visualization:*** In this section, we analyze the middle results of our model on Olympic sports and HMDB51 datasets. We apply the learned embedding A^t on temporal segments of testing videos, each of which is projected to a corresponding segmental semantic representation.

Firstly, we category each segmental semantic representation via nearest prototype classifier (Eq. 17) and calculate the average accuracy. Results are recorded in Table V. Results show that the latter segments containing more frames achieve higher accuracies than their former ones. This echoes our temporal mapping criterion that the confidence of each segment being mapped to its true label is non-decreasing as much more frames are involved in it.

TABLE V
ACCURACY (%) OF EACH TEMPORAL SEGMENT OF OLYMPIC SPORT AND HMDB51 DATASETS. LABEL EMBEDDING: ATTRIBUTE (A).

Segments	**m = 1**	**m = 2**	**m = 3**	**m = 4**	**m = 5**
Olympic (A)	53.63	56.14	57.93	59.24	60.13
HMDB51 (W)	7.96	15.46	16.95	17.23	17.72

Secondly, we randomly select a testing video of Olympic dataset, whose first frame of each temporal segment and corresponding segmental semantic representations are shown in Fig. 10. The bottom row visualizes the intermediate frames of action 'Snatch', which displays the whole dynamic evolution of the action. The upper row visualizes the segmental semantic representations of the video, whose horizontal axis denotes the indexes of attributes and vertical axis denotes the values of attributes. To get a better visualization, we adjust the values of attributes to the range$[-1, 1]$. It shows that the envelopes of segmental semantic representations are fitting gradually with corresponding prototype along with increasing fragments. It couples with the non-decreasing accuracies of temporal segments.

D. Evaluation of regularization terms

To further analyze each component in our model, the comparisons between our full model (ADPE) and sub-models are implemented on UCF101 dataset using attribute space. We remove adaptive regularization term (ADPE-noAD) or aligned regularization term (ADPE-noAL) from the full model. The average accuracy of each split of those models is reported in Fig. 11. Results clearly show that our full model achieves superior performance compared to its sub-models. Thus, both of the two regularization terms play important roles in our framework.

*1) **Adaptive term:*** If we remove the adaptive term from our model (ADPE-noAD), the average accuracy of each split is reduced. It demonstrates that the source mapping matrix A^s, which carries some shared semantic information between source and target data, provides a beneficial guidance to the adaptation of target mapping A^t.

TABLE VI
COMPARISON BETWEEN THE MODELS USING ALIGNED TERM OR MMD TERM ON UCF101 DATASET. ACCURACY (%) AND STANDARD DEVIATION ARE REPORTED FOR PERFORMANCE.

Methods	**RR**	**MMD**	**ADPE-noT**	**ADPE**
UCF101	15.61±1.12	15.76±1.24	18.18±1.42	20.56±1.09

*2) **Aligned term:*** Without our aligned term (ADPE-noAL), the accuracy of each split significantly decreases, which strongly proves the effectiveness of the novel aligned term. With the aligned regularization term, distributions of learned semantic representations Y^t and label prototypes P^t are enforced to be close to each other in the semantic space. Meanwhile, relationships between different categories are captured via matching the center of each group of semantic representations with each label prototype. Thus, domain shift problem can be alleviated via this term. In order to further demonstrate the effectiveness of aligned term, we compare it with MMD constraint. We replace our aligned term with

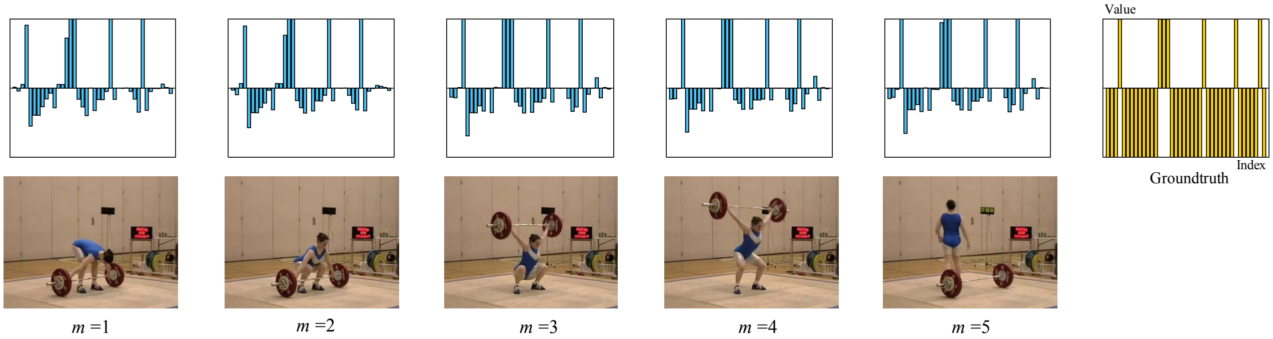


Fig. 10. The visualizations of the first frame of each temporal segment and corresponding semantic representations of a random video in Olympic dataset.

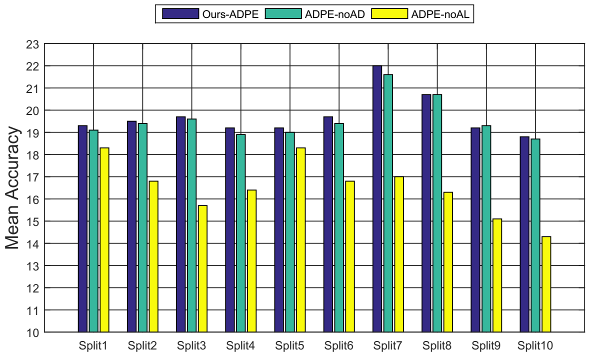


Fig. 11. Evaluations of regularization terms on UCF101 dataset using attribute space.

MMD measurement ($T_{2(\mathrm{MMD})}$ in Eq. 6) and record the results on UCF101 dataset using attribute space in Table VI. Explicitly, the aligned term that takes the inter-class variations into account helps learn an more effective visual-to-semantic embedding.

In summary, adaptive term restricts the learning of A^t in a neighborhood of A^s, while aligned term gives a discriminative direction to its learning. These two components complement each other, and contribute to the superior performance of our model.

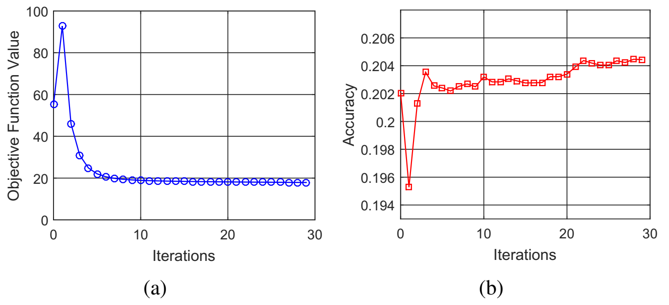
E. More discussions

*1) **Generalized ZSL:*** In generalized ZSL setting, the search space of target videos contains both training and testing classes ($P^s \bigcup P^t$). In other words, the testing video could belong to either seen classes or unseen ones. To further show the potential of our method in generalized ZSL setting, we randomly sample 50% of the original source videos for testing. Therefore, the generalized target set contains original target videos and 50% selected source videos. The generalized source set consists of the remaining original source videos. We learn the visual-to-semantic mappings (GA^s, GA^t) for both generalized source and target domains via our ADPE model (Eq. 3, Eq. 5), and record the performance on the unseen and seen videos in Table VII, respectively.

We can find that the source mapping GA^s could only recognize the seen actions containing in testing set. Most of unseen videos are mis-judged as seen ones, whose accuracy is even inferior to random guess. That's mainly because the source mapping GA^s only considers the structures of seen videos, which results in severe domain shift problem for unseen videos in generalized ZSL setting.

By contrast, our target mapping GA^t has the potential to help us recognize both seen and unseen videos. On the one hand, we can get considerable performance of unseen videos via GA^t (58.94±6.97, 15.72±4.63 and 13.70±2.02). On the other hand, the recognition of seen videos doesn't shift to unseen categories entirely. Our proposed ADPE embedding GA^t not only transfers shared semantic information from seen categories to unseen ones, but also characterizes inter-class variations of unseen videos, which is effective to solve domain shift problem. However, the accuracies on the three datasets in generalized ZSL setting are lower than those in standard setting, which indicates the strictness and difficulties of generalized ZSL. We will further study this problem in our future work.

*2) **Convergence:*** We apply Alternating Direction Method (ADM)[1] to solve our optimization problem and a local minima can be reached. To better illustrate the convergence of our model, we record value variations of our objective function of a random split on UCF101 dataset, as shown in Fig. 12. Results show that the objective value reduces quickly in the first 5 iterations and the model takes about 10 iterations to converge to a local minima. Accordingly, the accuracy boosts in the first 5 iterations and increases slowly in the following iterations.



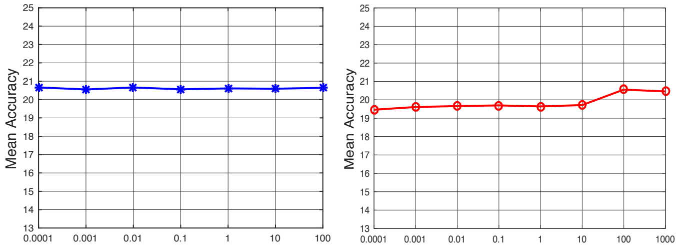
(a) (b)

Fig. 12. Evaluations of convergence of our model. (a) The value variations of objective function along iterations. (b) The recognition accuracy along iterations.

*3) **Evaluation of parameters:*** There are two important parameters in our model: factors of adaptive regu-

TABLE VII
EVALUATIONS OF OUR ADPE MODEL IN GENERALIZED SETTING. LABEL EMBEDDING: ATTRIBUTE (A) OR WORD2VECTOR (W). ACCURACY (%) AND STANDARD DEVIATION ARE REPORTED FOR PERFORMANCE.

Methods	Olympic (A)		HMDB51 (W)		UCF101 (A)	
	Unseen	**Seen**	**Unseen**	**Seen**	**Unseen**	**Seen**
Random guess	12.50	12.50	4.0	3.85	2.0	2.0
Ours-GA^s	0.61 ± 1.30	**91.07 ± 2.78**	0.04 ± 0.09	**73.83 ± 3.43**	0.87 ± 0.65	**94.63 ± 1.18**
Ours-GA^t	**58.94 ± 6.97**	41.19 ± 11.85	**15.72 ± 4.63**	35.12 ± 5.19	**13.70 ± 2.02**	62.90 ± 3.91



(a) (b)

Fig. 13. Evaluations of parameters on UCF101 datasets. (a) Fix λ_2, vary λ_1; (b) Fix λ_1, vary λ_2.

larization term λ_1 and aligned regularization term λ_2. In the above experiments, they are empirically set to 0.1 and 0.0001 on Olympic dataset, 0.1 and 100 on UCF101 dataset and 0.1 and 0.1 on HMDB51 dataset, respectively. In this section, we analyze the impacts of these parameters on UCF101 dataset. Firstly, we fix λ_2 to 100 and vary λ_1 in the range of $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$. Then we fix λ_1 to 0.1 and vary λ_2 in the range of $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. The variations of average accuracies of 10 splits are shown in Fig. 13. From Fig 13 (a) we can observe that our model is insensitive to the adaptive parameter on UCF101 dataset. Fig 13 (b) shows that our model achieves a slightly better performance when the aligned parameter gets larger. In general, our model is not very sensitive to all the parameters.

V. CONCLUSION

This paper studies the zero-shot action recognition problem. The challenge of this task is to overcome the serious domain shift problem and preserve temporal dynamic information of videos simultaneously. A novel aligned dynamic-preserving embedding model (ADPE) was proposed in this work, which not only captures dynamic evolution of actions and ensures temporal consistency of videos, but also takes the variations across different action categories into consideration. Experiments on three challenging action datasets have demonstrated the effectiveness of our model.

ACKNOWLEDGMENT

This work was supported partly by the National Natural Science Foundation of China (61772067, 61472030, 61471032), NSF IIS award 1651902.

REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[2] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *Proceedings of 24th International Conference on Artificial Intelligence*, pages 2234–2240, Buenos Aires, Argentina, July 2015.

[3] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2584–2591, 2013.

[4] A. Farhadi, A. F. . I. E. . D. H. . D. F. I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2009.

[5] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *Proceedings of European Conference on Computer Vision*, pages 584–599, 2014.

[6] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3769–3775, 2015.

[7] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proceedings of Tenth IEEE International Conference on Computer Vision*, pages 1395–1402, 2005.

[8] A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra. Robust human action recognition via long short-term memory. In *Proceedings of International Joint Conference on Neural Networks*, pages 1–8, 2013.

[9] M. Hoai and F. D. la Torre. Max-margin early event detectors. In *International Journal of Computer Vision*, volume 107, pages 191–202, 2014.

[10] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *2015 IEEE International Conference on Computer Vision*, pages 4588–4596, 2015.

[11] M. Jain, J. C. van Gemert, T. Mensink, and C. G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4588–4596, 2015.

[12] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, pages 3464–3472, 2014.

[13] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2452–2460, 2015.

[14] Y. Kong and Y. Fu. Max-margin action prediction machine. *IEEE Trans Pattern Anal Mach Intell*, 38(9):1844 – 1858, 2016.

[15] Y. Kong, Y. Jia, and Y. Fu. Interactive phrases: semantic descriptions for human interaction recognition. *IEEE Trans Pattern Anal Mach Intell*, 36(9):1775–1788, 2014.

[16] Y. Kong, D. Kit, and Y. Fu. A discriminative model with multiple temporal scales for action prediction. In *European Conference on Computer Vision*, pages 596–611, 2014.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2556–2563, 2011.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2009.

[19] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, March 2014.

[20] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2):107–123, 2005.

[21] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, June 2011.

[22] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu. Transfer sparse coding for robust image representation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 407–414, 2013.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[24] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[25] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. A discriminative model with multiple temporal scales for action prediction. In *European Conference on Computer Vision*, pages 392–405, 2010.

[26] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proceedings of European Conference on Computer Vision*, pages 392–405, 2010.

[27] X. Peng, L. Wangb, X. Wangc, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.

[28] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of European Conference on Computer Vision*, pages 143–156, 2010.

[29] 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1042–1051, 2017.

[30] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Proceedings of 26th Advances in Neural Information Processing Systems*, pages 46–54, 2013.

[31] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 32–36, 2004.

[32] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

[33] B. Su, X. Ding, and C. Liu. Discriminative transformation for multi-dimensional temporal sequences. In *IEEE Transactions on Image Processing*, volume 26, pages 3579–3593, 2017.

[34] B. Su, X. Ding, H. Wang, and Y. Wu. Discriminative dimensionality reduction for multi-dimensional sequences. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pages 77–91, 2018.

[35] B. Su and Y. Wu. Learning low-dimensional temporal representations. In *International Conference on Machine Learning*, pages 4768–4777, 2018.

[36] B. Su, J. Zhou, X. Ding, and Y. Wu. Unsupervised hierarchical dynamic parsing and encoding for action recognition. In *IEEE Transactions on Image Processing*, volume 26, pages 5784–5799, 2017.

[37] Y. Tian, Q. Ruan, and G. An. Zero-shot action recognition via empirical maximum mean discrepancy. In *The 14th IEEE International Conference on Signal Processing*.

[38] Y. Tian, Q. Ruan, G. An, and Y. Fu. Action recognition using local consistent group sparse coding with spatio-temporal structure. In *Proceedings of the ACM on Multimedia Conference*, pages 317–321, 2016.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. *The IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[40] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, 15(1):3221–3245, 2014.

[41] H. Wang, A. Kluser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

[42] H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3):219–238, 2016.

[43] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of IEEE International Conference on Computer Vision*, pages 3551–3558, 2013.

[44] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, and C. Sun. Action recognition using nonnegative action component representation and sparse basis selection. *IEEE Trans Image Process*, 23(2):570–581, 2013.

[45] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen. Temporal pyramid pooling-based convolutional neural network for action recognition. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 27, pages 2613–2622, 2017.

[46] X. Xu, T. Hospedales, and S. Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, pages 1–25, 2017.

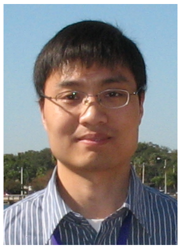
[47] X. Xu, T. M. Hospedales, and S. Gong. Semantic embedding space for zero-shot action recognition. In *Proceedings of IEEE International Conference on Image Processing*, pages 63–67), September 2015.

[48] X. Xu, T. M. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *Proceedings of European Conference on Computer Vision*, pages 343–359, 2016.

[49] X. Zhen, L. Shao, D. Tao, and X. Li. Embedding motion and structure features for action recognition. In *IEEE Transactions on Circuits and Systems for Video Technology*, volume 23, pages 1182–1190, 2013.

[50] V. S. Ziming Zhang. Zero-shot learning via semantic similarity embedding. In *Proceedings of IEEE International Conference on Computer Vision*, pages 4166–4174, 2015.

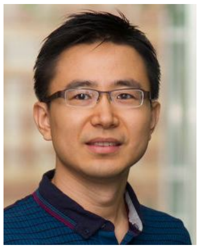
Yi Tian received the B.E. degree from Beijing Jiaotong University, China, in 2011, and PH.D degree in Institute of Information Science, Beijing Jiaotong University, China, in 2018. She was a visiting student in the Electrical and Computer Engineering Department, Northeastern University, Boston, Massachusetts during the period of 2016.10-2017.10. Her research interests mainly focus on human action recognition.



Yu Kong received B.Eng. degree in automation from Anhui University in 2006, and PhD degree in computer science from Beijing Institute of Technology, China, in 2012. He is now a tenure-track Assistant Professor in the College of Computing and Information Sciences at Rochester Institute of Technology. Prior to that, he was a postdoc in the Department of ECE, Northeastern University. He visited Department of Computer Science and Engineering, University at Buffalo, SUNY, and the National Laboratory of Pattern Recognition (NLPR), Chinese Academy of Science. Dr. Kong's research interests include computer vision, social media analytics, and machine learning. He is a member of the IEEE.

Qiuqi Ruan (SM96) received the B.S. and M.S. degree from Northern Jiaotong University, China in 1969 and 1981 respectively. From January 1987 to May 1990, he was a visiting scholar at the University of Pittsburgh, Pittsburgh, PA, and at the University of Cincinnati, Cincinnati, OH. Subsequently, he has been a Visiting Professor in the U.S. for several times. He is currently a Professor and a Doctorate Supervisor at the Institute of Information Science, Beijing Jiaotong University, Beijing. He is IEEE Beijing Section Chairman. He is the author of two books and more than 100 papers. He is the holder of a national patent. His main research interests include digital signal processing, computer vision, pattern recognition, and virtual reality

Gaoyun An Gaoyun An received the B.S. degree in Biological Engineering and Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University in 2003 and 2008 respectively, Beijing, China. Currently, he is an associate professor in Institute of Information Science, Beijing Jiaotong University, Beijing, China. His main research interests include image processing, computer vision and pattern recognition.



Yun Fu (S'07-M'08-SM'11-F'19) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated with College of Engineering and the Khoury College of Computer and Information Sciences at Northeastern University since 2012. His research interests are Machine Learning, Computational Intelligence, Big Data Mining, Computer Vision, Pattern Recognition, and Cyber-Physical Systems. He has extensive publications in leading journals, books/book chapters and international conferences/workshops. He serves as associate editor, chairs, PC member and reviewer of many top journals and international conferences/ workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; nine Best Paper Awards from IEEE, IAPR, SPIE, SIAM; many major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an Associate Editor of the IEEE Transactions on Neural Networks and Leaning Systems (TNNLS). He is fellow of IEEE, IAPR, OSA and SPIE, a Lifetime Distinguished Member of ACM, Lifetime Member of AAAI and Institute of Mathematical Statistics, member of ACM Future of Computing Academy, Global Young Academy, AAAS, INNS and Beckman Graduate Fellow during 2007-2008.