Exploiting Images for Video Recognition: Heterogeneous Feature Augmentation via Symmetric Adversarial Learning

Feiwu Yu, Xinxiao Wu, *Member, IEEE*, Jialu Chen, and Lixin Duan

Abstract—**Training deep models of video recognition usually requires sufficient labeled videos in order to achieve good performance without over-fitting. However, it is quite labor-intensive and time-consuming to collect and annotate a large amount of videos. Moreover, training deep neural networks on large-scale video datasets always demands huge computational resources which further hold back many researchers and practitioners. To resolve that, collecting and training on annotated images are much easier. However, thoughtlessly applying images to help recognize videos may result in noticeable performance degeneration due to the well-known domain shift and feature heterogeneity. This proposes a novel symmetric adversarial learning approach for heterogeneous image-to-video adaptation, which augments deep image and video features by learning domain-invariant representations of source images and target videos. Primarily focusing on an unsupervised scenario where the labeled source images are accompanied by unlabeled target videos in the training phrase, we present a data-driven approach to respectively learn the augmented features of images and videos with superior transformability and distinguishability. Starting with learning a common feature space (called image-frame feature space) between images and video frames, we then build new symmetric generative adversarial networks (Sym-GANs) where one GAN maps image-frame features to video features and the other maps video features to image-frame features. Using the Sym-GANs, the source image feature is augmented with the generated video-specific representation to capture the motion dynamics while the target video feature is augmented with the image-specific representation to take the static appearance information. Finally, the augmented features from the source domain are fed into a network with fully connected layers for classification. Thanks to an end-to-end training procedure of the Sym-GANs and the classification network, our approach achieves better results than other state-of-the-arts, which is clearly validated by experiments on two video datasets, i.e., the UCF101 and HMDB51 datasets.**

Index Terms—**Heterogeneous domain adaptation, feature augmentation, symmetric GANs, image-to-video adaptation.**

Manuscript received December 9, 2018; revised April 23, 2019; accepted May 12, 2019. Date of publication May 24, 2019; date of current version August 22, 2019. This work was supported by the Natural Science Foundation of China (NSFC) under Grant 61673062 and Grant 61772118. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiaochun Cao. *(Corresponding author: Xinxiao Wu.)*

F. Yu, X. Wu, and J. Chen are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: yufeiwu@bit.edu.cn; wuxinxiao@bit.edu.cn; chenjialu@bit.edu.cn).

L. Duan is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: lxduan@uestc.edu.cn).

Digital Object Identifier 10.1109/TIP.2019.2917867

I. Introduction

VIDEO recognition is an active research topic in computer vision due to its wide applications such as video retrieval, intelligent video surveillance and smart robots system. Thanks to the great success of deep neural networks, the performance of classifying videos has been dramatically improved. However, training deep video classifiers requires collecting and labeling large amounts of videos to overcome over-fitting, which is particularly labor-intensive and time-consuming. Furthermore, training deep neural networks on such large-scale dataset usually consumes substantial computational and storable resources. Fortunately, it is much easier to collect and annotate images, and there are also many existing labeled image datasets that can be leveraged. In addition, images often highlight the discriminative static information within videos, such as the scenes, object appearances and human postures, which have complementary characteristics to videos. Therefore, it would be beneficial a lot to utilize images to train deep models for video recognition with much less computational cost.

However, directly applying the images trained classifier to videos might lead to the domain shift problem, where the variations in data distribution between source images and target videos will significantly degrade the classification performance at test time. To solve this problem, several recent methods [1]–[3] use images and video frames to train shared CNNs to learn the common feature between image and video domains. Li *et al.* [1] exploited class-discriminative spatial attention maps for transferring images to videos to make video classifiers trained on images suffer less from the domain shift. In these methods, each video is represented by a bag of images and the underlying temporal relationship between sequential frames may be lost during the knowledge transfer.

Different from the previous works aforementioned, we primarily focus on the heterogeneous domain adaptation from image to video, where the video is represented by spatiotemporal feature which totally differs from the image feature in both feature dimension and physical meaning. Inspired by recent advances of generative adversarial learning [4], we propose a novel symmetric generative adversarial learning approach to transfer knowledge from image to video by learning domain-invariant feature representation between them. Our method is under the unsupervised scenario where all the source images are labeled while all the target videos are unlabeled. Two generative adversarial networks (GANs) with symmetric architectures, called Sym-GANs, are built to learn the bidirectional mappings between source image feature and target video feature. Then the image feature from the source domain is augmented with the video-specific feature generated by the image-to-video mapping and the video feature from the target domain is augmented with the image-specific feature generated by the video-to-image mapping. The new augmented features of source images or target videos can be treated as domain-invariant features with superior transferability and representability by capturing both spatial appearance and temporal motion information.

Since there is no correspondence between source images and target videos, the video frames and their corresponding videos are utilized to train the Sym-GANs model. Due to the data distribution discrepancy between source images and video frames, we adopt the JAN model [5] to learn a common feature space between them, called image-frame feature space. Accordingly, the bidirectional mappings between image and video features are actually the ones between image-frame and video features. Thus, the image-specific part of the augmented feature is represented by the image-frame feature.

Finally, we design a classification network with fully connected layers which takes the augmented features as input and outputs the class label. A joint training method is presented to simultaneously learn the Sym-GANs model and the classification network to enhance the discriminative ability of the augmented feature. Three losses, i.e., the adversarial loss, the Correlation Alignment (CORAL) loss, and the cross-entropy loss, are effectively combined for training. The adversarial loss matches the distribution of generated features to the data distribution in the original domain. The CORAL loss minimizes the difference between the synthesized features and the original features in second-order statistics. The cross-entropy loss is responsible for the classification.

Overall, the main contributions are:

- We propose a novel symmetric generative adversarial learning approach to learn domain-invariant augmented feature with excellent transferability and distinguishability for heterogeneous image-to-video adaptation. It is worth emphasizing that the augmented feature preserves both static appearance and temporal motion information with superior descriptive ability which can be learned without any paired image-video training data in an unsupervised scenario.
- We build two generative adversarial networks with symmetric architecture (Sym-GANs) to learn the bidirectional mappings between heterogeneous two domains, and formulate the training of Sym-GANs and classification network in a joint learning manner to further enhance the discriminative ability of the augmented feature.
- Promising results on both the UCF101 and HMDB51 video datasets clearly evaluate the effectiveness of our method in leveraging images for video recognition.

The organization of the rest of this paper is given as follows. In Section II, we summarize the related works of learning from images to videos, domain adaptation and generative adversarial network. Section III describes the proposed method for heterogeneous adaptation from source images to target videos, including problem formulation, network architecture, learning and prediction. Section IV elaborates on the experimental result and analysis. The conclusion is made in Section V.

II. Related Work

A. Learning from Images to Videos

To leverage the information from large-scale images with annotations, several approaches have been proposed to take images as auxiliary domain for video recognition [6]–[8]. Yang *et al.* [6] presented a cross-media video tagging scheme to transfer tag knowledge from images to videos by exploring the intrinsic data structures of both images and videos. Duan *et al.* [7] leveraged a large number of loosely labeled images from different Web sources for recognizing events in videos, via proposing a multiple source domain adaptation method.

Recently, there have been some attempts based on deep learning to handle the image-to-video transfer problem [1]–[3], [9], [10]. Ma *et al.* [2] first collected a large scale image dataset from the Web, and then combined these Web images and frames of videos to train deep Convolutional Neural Network (CNN) for action recognition. Sun *et al.* [9] explored tagged images for temporal action localization in videos. They used pre-trained CNN for cross-domain transfer between video frames and Web images. Gan *et al.* [10] exploited both images and videos from the Web and proposed a mutually voting approach to select relevant images and video frames for effective transferring between images and video frames. Li *et al.* [1] proposed a method of adapting a CNN trained on Web images to videos with attention mechanism. It uses class-discriminative spatial attention maps to alleviate the domain shift problem. In all these methods, the shared CNN between image and video is trained using both images and video frames to learn the common deep features of both images and videos. In contrast, our method investigates on the heterogeneous image-to-video adaptation where the video is represented by spatial-temporal feature to capture the motion information that totally differs from the image representation.

B. Domain Adaptation

As a well-studied machine learning strategy, domain adaptation has gained increasing attention in various visual tasks [11]–[17]. How to learn domain-invariant transferable feature representation between different domains remains an important issue in domain adaptation. Earlier approaches [13], [14], [18], [19] resort to first estimating the weights of the source domain data and then training classifiers on the reweighted data with transferability. Some other approaches [20]–[22] learn domain-invariant feature representation by mapping function that aligns the source distribution to the target domain. In the literature, a few studies focus on heterogeneous domain adaptation where the source data and target data are represented by different types of feature representations. In [23]–[26], good common feature spaces are learned for connecting the heterogeneous source and target domains. In [27], an asymmetric feature transformation is proposed for knowledge transferring between source and target domains.

Deep neural networks have been widely exploited to learn more transferable features via integrating domain adaptation into the pipeline of deep learning [5], [28]–[32]. Tzeng *et al.* [29] introduced an adaptation layer into a traditional CNN architecture and designed a domain confusion loss to reduce the data bias between source and target domains. In [30], a deep adaptation network is proposed for reducing the domain discrepancy in higher task-specific layers using an optimal multi-kernel selection method for mean embedding matching. Later in [31] they extended the deep adaptation network to a residual module. To handle the domain shift in the joint distributions of input features and output labels, Long *et al.* [5] presented joint adaptation networks to align the joint distributions of multiple domain-specific layers using a Mean Maximum Distance criterion. Sun and Saenko [33] extended the CORAL [34] to a nonlinear transformation and used it for aligning correlations of layer activations in deep neural networks to reduce the domain shift. All these methods assume that the data from the source and target domains are represented by the same type of feature. In this paper, our deep neural networks deal with the heterogeneous domain adaptation, in which the feature representations of source and target data are totally different. Recently, few deep models have been proposed for heterogeneous adaptation. Chen *et al.* [35] proposed Transfer Neural Trees (TNT) as a novel Neural Network based architecture for semi-supervised heterogeneous domain adaptation. Different from TNT which requires the labeled target data for training, our method handles the heterogeneous domain adaptation in an unsupervised fashion without any supervision in the target domain.

C. Generative Adversarial Network

Inspired by the adversarial learning strategy, Generative Adversarial Networks (GANs) have achieved impressive progress in domain adaptation where an adversarial loss with respect to domain labels has become a popular solution to reduce the domain discrepancy [32], [36]–[40]. Ganin and Lempitsky [32], Ganin *et al.* [38] proposed domain adversarial neural networks to learn domain invariant features by an adversarial learning strategy between the feature extractor and the domain classifier. Tzeng *et al.* [36] first pre-trained a source CNN using labeled source data and then learned a target CNN to make a discriminator not correctly and reliably classify the encoded source and target samples into domains. Reference [39] proposes importance weighted adversarial networks, in which a weighting scheme is presented for detecting the samples from the outlier classes in the source domain to effectively reduce the domain shift. In [40], an adversarial image generation approach is presented to directly learn a joint feature space where the distance between source and target distributions are minimized. All these feature-level adversarial adaptation methods focus on modifications to the embedding discriminative feature space of homogeneous domains. Different from them, our method employs the adversarial learning to the heterogeneous domain adaptation. It learns bidirectional mappings between two domains by symmetric GANs to generate augmented features with powerful transferability for domain adaptation.

With the good performance of adversarial training in generative models, there has been a rich line of recent work to apply adversarial loss on pixel-level for domain adaptation [12], [41]–[45]. In [41], a GAN-based approach is presented to transform an image from one domain to the other in the pixel level using a task-specific loss and a new content-similarity loss. Taigman *et al.* [42] learned a domain transfer network for transferring a sample in one domain to analog sample in another domain via the combination of a multiclass GAN loss, an f-constancy component and a regularizing component. Liu and Tuzel [43] proposed coupled generative adversarial networks to learn a joint distribution of multi-domain images without any tuples of corresponding images and successfully applied it to unsupervised domain adaptation. Zhu *et al.* [44] combined an adversarial losses with a cycle consistency loss for learning to translate an image from source domain to target domain in the absence of paired samples. Russo *et al.* [45] introduced a bi-directional adaptive adversarial domain adaptation architecture that maps simultaneously source samples into the target domain and vice versa with the aim to learn and use both classifiers at test time. Different from these pixel-level adversarial adaptation methods which handle the image-to-image adaptation, our method uses the generative adversarial learning strategy to address the image-to-video adaptation on the heterogeneous feature level, where the source images and target videos are represented by different types of features.

III. Symmetric GANs for Heterogeneous Feature Augmentation

A. Problem Formulation

Our core idea of addressing heterogeneous domain adaptation from image to video is to learn a common feature representation with superior transferability between the two domains. Different from the exiting embedding space methods [23], [25], [46] and the asymmetric mapping method [27], we attempt to augment the original image and video features respectively with their corresponding complementary features to generate the domain-invariant features. Motivated by the good performance of generative model in GANs, we propose symmetric GANs (Sym-GANs) to build the bidirectional mappings between source image and target video for the feature augmentation. In Sym-GANs, one GAN maps image to video, and the other is responsible for the video-to-image mapping. Using Sym-GANs, the image feature from the source domain can be augmented with its projected representation in the video feature space and the video feature from the target domain can be augmented with its mapped representation in the image feature space. The augmented features of both domains are more powerful and descriptive by preserving the static information in image as well as the temporal motion in video. To improve the discriminative ability of

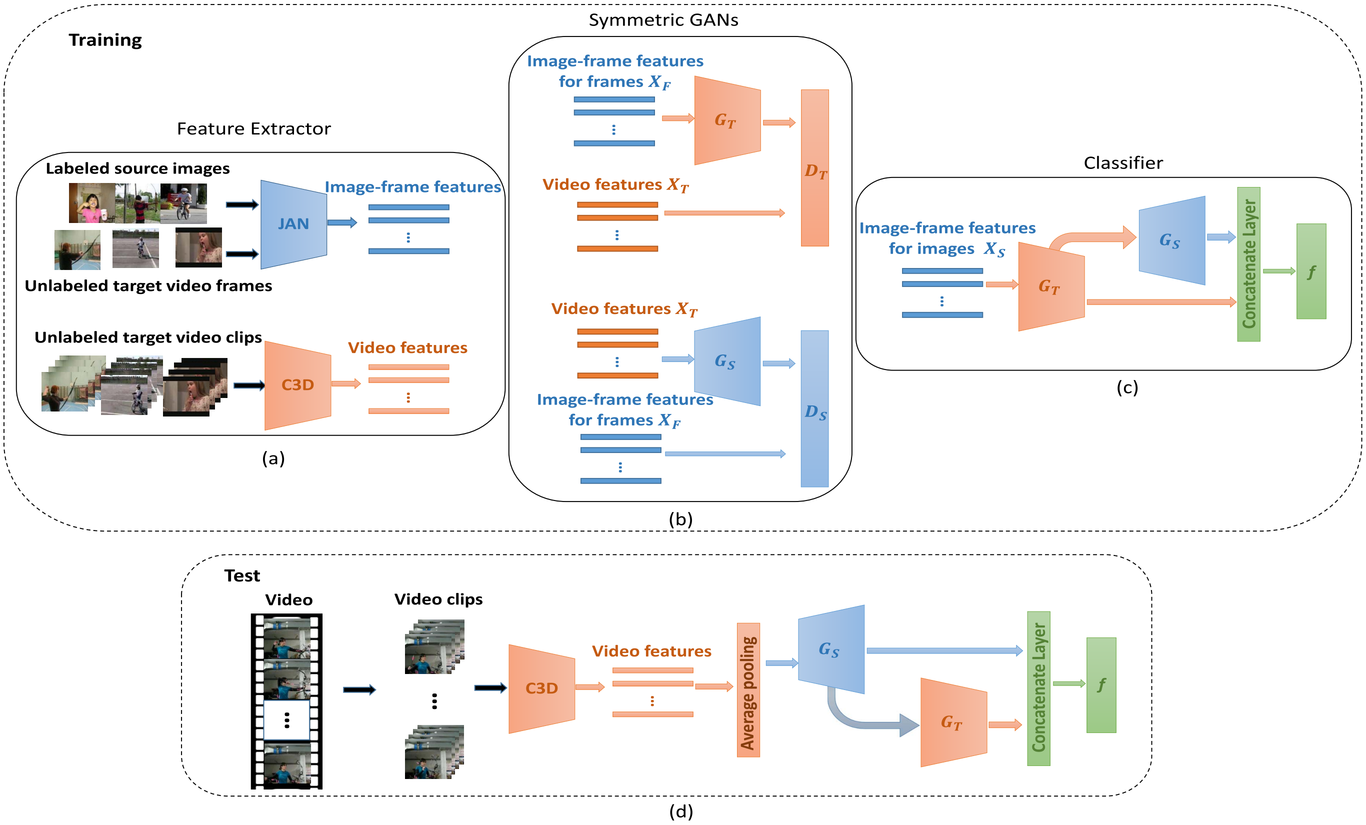


Fig. 1. The overall architecture of our method. There are three components: feature extractor (a), Symmetric GANs (b) and classifier (c), which are learned in the training phrase. The test procedure of target labels' inference is shown in (d).

the augmented feature, we present a joint optimization method to simultaneously learn the Sym-GANs and the classifier under the class label supervision from the source domain. A formal problem statement is given below.

Let $X_S = \{\mathbf{x}_s^i|_{i=1}^{n_s}\}$ denote n_s images from the source domain, where $\mathbf{x}_s^i \in \mathbb{R}^{d_s \times 1}$ represents the feature vector of the i-th image. Let $Y_S = \{y_s^i|_{i=1}^{n_s}\}$ denote the class labels of X_S where $y_s^i \in \{1, 2, ..., C\}$ is the label of $\mathbf{x}_s^i$ and C is the number of classes. For each video in the target domain, we divide it into several clips with the same length. Let $X_T = \{\mathbf{x}_t^i|_{i=1}^{n_t}\}$ denote n_t unlabeled video clips from the target domain, where $\mathbf{x}_t^i \in \mathbb{R}^{d_t \times 1}$ represents the feature vector of the i-th video clip. Note that in the heterogeneous domain adaptation problem, $d_s \neq d_t$. We now aim to learn the bidirectional mappings between X_S and X_T, defined by $G_T : \mathbf{x}_s \rightarrow \mathbf{x}_t$ and $G_S : \mathbf{x}_t \rightarrow \mathbf{x}_s$. Then the augmented feature $\hat{\mathbf{x}}_s$ of the original image feature $\mathbf{x}_s$ is given by $\hat{\mathbf{x}}_s = [\mathbf{x}_s; G_T(\mathbf{x}_s)] \in \mathbb{R}^{(d_s+d_t)\times 1}$. Similarly, the augmented feature $\hat{\mathbf{x}}_t$ of the original video clip feature $\mathbf{x}_t$ is formulated by $\hat{\mathbf{x}}_t = [G_S(\mathbf{x}_t); \mathbf{x}_t] \in \mathbb{R}^{(d_s+d_t)\times 1}$.

Since there is no direct corresponding between X_S and X_T (i.e. unpaired image and video samples) for training the two mappings, we assume that a video clip has a relationship with any frame in it (i.e., paired frame and video samples). On the other hand, video frames are actually a collection of images, which could be easily adapted to the source image domain. Thus for each video clip, we randomly select one frame and all the selected frames compose the unlabeled video frame data, denoted by $X_F = \{\mathbf{x}_f^i|_{i=1}^{n_t}\}$ where $\mathbf{x}_f^i \in \mathbb{R}^{d_s \times 1}$ indicates the feature vector of frame from the i-th video. Consequently, the mapping from video to image becomes $G_S : \mathbf{x}_t \rightarrow \mathbf{x}_f$ and the mapping from image to video is $G_T : \mathbf{x}_f \rightarrow \mathbf{x}_t$.

B. Proposed Architecture

The overall architecture of our method consists of three key components: feature extractor, symmetric GANs, and classifier, as shown in Figure 1. The symmetric GANs is a new contribution of our work.

1) Feature Extractor: The images and video clips are represented by heterogeneous features. We extract the C3D [47] feature to describe the video clips by capturing spatia-temporal information such as the temporal evolution of human postures and the continuously changing scenes. Considering the domain shift between source images and target video frames, we adopt a state-of-the-art deep domain adaptation method (i.e., JAN [5] model) to learn the common CNN feature invariant to both source image and video frame, using X_S and X_F as the training data. This learned domain-invariant feature is called **image-frame feature** in the rest of this paper. Accordingly, both $\mathbf{x}_s$ and $\mathbf{x}_f$ are represented by the image-frame feature. In contrast to the video feature, the image-frame feature tends to highlight the static information such as human body posture and appearance. Thus it would be beneficial to combine the complementary image-frame feature and video feature for video recognition.

2) Symmetric GANs: The Generative Adversarial Network (GAN) [4] has been successfully applied in many visual tasks, including image-to-image translation [44], [48]–[50], semantic segmentation [51], person re-identification [52] and object detection [53]. A traditional GAN consists of two competing networks: a generator G and a discriminator D. G and D compete with each other in a two-player minmax game. The generator G tries to produce samples as realistic as possible to confuse the discriminator D while the discriminator D aims to differentiate between the generated samples and the real ones as correctly as possible. More formally, G and D can be trained jointly by solving

$$\min_G \max_D \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))], \tag{1}$$

where $D(\mathbf{x})$ represents the probability that $\mathbf{x}$ comes from the real data distribution rather than the distribution modeled by the generator G.

In this paper, we build two GANs with symmetric structures to learn the bidirectional mappings between the image-frame feature space and the video feature space. Let G_T represent the mapping from the image-frame feature to the video feature, associated with the discriminator D_T. Given the paired training data X_F and X_T, to train the G_T and D_T, the loss function is formulated as

$$\begin{aligned}&\mathcal{L}_{GAN}(G_T, D_T, X_F, X_T)\\ &\quad = \mathbb{E}_{\mathbf{x}_t \sim P_T(\mathbf{x}_t)}[\log D_T(\mathbf{x}_t)]\\ &\qquad + \mathbb{E}_{\mathbf{x}_f \sim P_F(\mathbf{x}_f)}[\log(1 - D_T(G_T(\mathbf{x}_f)))].\end{aligned} \tag{2}$$

G_T attempts to generate video features $G_T(X_F)$ that resemble the real video features X_T, while D_T tries to distinguish $G_T(X_F)$ from X_T. In other words, G_T aims at minimizing the loss against an adversarial D_T that tries to maximize it:

$$\min_{G_T} \max_{D_T} \mathcal{L}_{GAN}(G_T, D_T, X_F, X_T). \tag{3}$$

Similarly, the mapping function G_S from the video feature to the image-frame feature and its associated discriminator D_S are jointly trained by the following loss:

$$\begin{aligned}&\mathcal{L}_{GAN}(G_S, D_S, X_T, X_F)\\ &\quad = \mathbb{E}_{\mathbf{x}_f \sim P_F(\mathbf{x}_f)}[\log D_S(\mathbf{x}_f)]\\ &\qquad + \mathbb{E}_{\mathbf{x}_t \sim P_T(\mathbf{x}_t)}[\log(1 - D_S(G_S(\mathbf{x}_t)))],\end{aligned} \tag{4}$$

and the optimization is given by

$$\min_{G_S} \max_{D_S} \mathcal{L}_{GAN}(G_S, D_S, X_T, X_F). \tag{5}$$

In the experiment, we replace the negative log likelihood objective in the loss $\mathcal{L}_{GAN}$ (Eq. 2 and Eq. 4) by a least square loss [54]:

$$\begin{aligned}&\mathcal{L}_{GAN}(G_T, D_T, X_F, X_T)\\ &\quad = \mathbb{E}_{\mathbf{x}_t \sim P_T(\mathbf{x}_t)}[D_T(\mathbf{x}_t)^2]\\ &\qquad + \mathbb{E}_{\mathbf{x}_f \sim P_F(\mathbf{x}_f)}[(1 - D_T(G_T(\mathbf{x}_f)))^2],\end{aligned} \tag{6}$$

$$\begin{aligned}&\mathcal{L}_{GAN}(G_S, D_S, X_T, X_F)\\ &\quad = \mathbb{E}_{\mathbf{x}_f \sim P_F(\mathbf{x}_f)}[D_S(\mathbf{x}_f)^2]\\ &\qquad + \mathbb{E}_{\mathbf{x}_t \sim P_T(\mathbf{x}_t)}[(1 - D_S(G_S(\mathbf{x}_t)))^2],\end{aligned} \tag{7}$$

which performs more stably during training and yields better results.

Besides the aforementioned generative adversarial loss, we also introduce CORAL loss [33] to minimize the difference between the generated features and the real features in second-order statistics. As a simple and effective criterion, the CORAL loss can be easily integrated into a deep neural network. Let $\mathbf{T} = [\mathbf{x}_t^1, \mathbf{x}_t^2, ..., \mathbf{x}_t^{n_t}] \in \mathbb{R}^{d_t \times n_t}$ denote the video feature matrix and $\mathbf{F} = [\mathbf{x}_f^1, \mathbf{x}_f^2, ..., \mathbf{x}_f^{n_t}] \in \mathbb{R}^{d_s \times n_t}$ denote the image-frame feature matrix of video frames. By the generator G_T, the synthesized video feature matrix from the video frames is formed by $\mathbf{T}_f = [G_T(\mathbf{x}_f^1), G_T(\mathbf{x}_f^2), ..., G_T(\mathbf{x}_f^{n_t})] \in \mathbb{R}^{d_t \times n_t}$. By the generator G_S, the generated image-frame feature matrix from the videos is formed by $\mathbf{F}_t = [G_S(\mathbf{x}_t^1), G_S(\mathbf{x}_t^2), ..., G_S(\mathbf{x}_t^{n_t})] \in \mathbb{R}^{d_s \times n_t}$.

For G_T associated with D_T, the CORAL loss is given by

$$\mathcal{L}_{CORAL}(X_T, G_T(X_F)) = \frac{1}{4d_t^2}\|\mathbf{C_T} - \mathbf{C}_{\mathbf{T}_f}\|_F^2, \tag{8}$$

where $\|\cdot\|_F^2$ is the squared matrix Frobenius norm, measuring the distance between the second-order statistics (covariance) of X_T and $G_T(X_F)$. $\mathbf{C_T}$ and $\mathbf{C}_{\mathbf{T}_f}$ are the feature covariance matrices of $\mathbf{T}$ and $\mathbf{T}_f$, respectively, calculated by

$$\mathbf{C_T} = \frac{1}{n_t - 1}(\mathbf{T}^\top\mathbf{T} - \frac{1}{n_t}(\mathbf{1}^\top\mathbf{T})^\top(\mathbf{1}^\top\mathbf{T})), \tag{9}$$

$$\mathbf{C}_{\mathbf{T}_f} = \frac{1}{n_t - 1}(\mathbf{T}_f^\top\mathbf{T}_f - \frac{1}{n_t}(\mathbf{1}^\top\mathbf{T}_f)^\top(\mathbf{1}^\top\mathbf{T}_f)), \tag{10}$$

where $\mathbf{1} \in \mathbb{R}^{d_t \times 1}$ has all elements equal to 1.

For G_S associated with D_S, we introduce a similar CORAL loss as well:

$$\mathcal{L}_{CORAL}(X_F, G_S(X_T)) = \frac{1}{4d_s^2}\|\mathbf{C_F} - \mathbf{C}_{\mathbf{F}_t}\|_F^2, \tag{11}$$

where the covariance matrices $\mathbf{C_F}$ and $\mathbf{C}_{\mathbf{F}_t}$ are given by

$$\mathbf{C_F} = \frac{1}{n_t - 1}(\mathbf{F}^\top\mathbf{F} - \frac{1}{n_t}(\mathbf{1}^\top\mathbf{F})^\top(\mathbf{1}^\top\mathbf{F})), \tag{12}$$

$$\mathbf{C}_{\mathbf{F}_t} = \frac{1}{n_t - 1}(\mathbf{F}_t^\top\mathbf{F}_t - \frac{1}{n_t}(\mathbf{1}^\top\mathbf{F}_t)^\top(\mathbf{1}^\top\mathbf{F}_t)). \tag{13}$$

3) Classifier: With the learned mapping G_T, we can augment the original image-frame feature of the source image $\mathbf{x}_s$ with its projected feature $G_T(\mathbf{x}_s)$ in the video space. Similarly, the original video feature in the target domain $\mathbf{x}_t$ can also be augmented with its projected feature $G_S(\mathbf{x}_t)$ in the image-frame space with G_S. In order to further improve the discriminative ability of the augmented features, we transform the projected video feature $G_T(\mathbf{x}_s)$ and the projected image-frame feature $G_S(\mathbf{x}_t)$ back to the image-frame space and the video space, respectively, to generate new features $G_S(G_T(\mathbf{x}_s))$ and $G_T(G_S(\mathbf{x}_t))$. We believe that the generated features $G_S(G_T(\mathbf{x}_s))$ and $G_T(G_S(\mathbf{x}_t))$ are more discriminative than their corresponding original features $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively, since the generators G_S and G_T are jointly learned with the classifier with the supervision information from the labeled source data. Actually, the augmented feature in the source domain is represented by $\hat{\mathbf{x}}_s = [G_S(G_T(\mathbf{x}_s)); G_T(\mathbf{x}_s)]$ and the augmented feature in the target domain is represented by $\hat{\mathbf{x}}_t = [G_S(\mathbf{x}_t); G_T(G_S(\mathbf{x}_t))]$. These augmented features are invariant to different domains, on which the classifier trained with source images can be adapted well to the target videos. Moreover, by capturing both static appearance and dynamic motion information, this hybrid representation would significantly benefit improving the recognition performance.

Consequently, the input to the classifier is the augmented feature and the output is the probability distribution of category labels. We build a network with fully connected layers to construct the classifier, denoted by f. Given the labeled augmented source data $\hat{X}_S = \{\hat{\mathbf{x}}_s^i|_{i=1}^{n_s}\}$ with its corresponding class labels $Y_S = \{y_s^i|_{i=1}^{n_s}\}$, we use cross-entropy loss to train f, defined as

$$\mathcal{L}_{class}(f, \hat{X}_S, Y_S) = -\mathbb{E}_{(\hat{\mathbf{x}}_s, y_s)\sim P_{data}(\hat{\mathbf{x}}_s, y_s)} \log(f(\hat{\mathbf{x}}_s)_{y_s}), \quad (14)$$

where $f(\hat{\mathbf{x}}_s)_{y_s}$ indicates the probability assigned by the classifier f for the input $\hat{\mathbf{x}}_s$ to the class y_s.

C. Learning

We have thus far described a deep domain adaptation method, which combines adversarial objective, CORAL constraint and cross-entropy loss to learn domain-variant feature representation between heterogeneous domains with superior transferable, descriptive and discriminative abilities.

Taken together, all the loss functions mentioned above form the complete objective:

$$\begin{aligned}
&\mathcal{L}(f, X_S, Y_S, X_F, X_T, G_T, D_T, G_S, D_S) \\
&\quad = \mathcal{L}_{GAN}(G_T, D_T, X_F, X_T) + \mathcal{L}_{GAN}(G_S, D_S, X_T, X_F) \\
&\qquad + \lambda_1 \mathcal{L}_{CORAL}(X_T, G_T(X_F)) \\
&\qquad + \lambda_2 \mathcal{L}_{CORAL}(X_F, G_S(X_T)) \\
&\qquad + \mathcal{L}_{class}(f, \hat{X}_S, Y_S) \\
&\qquad + \mathcal{L}_{reg}(G_T, D_T) + \mathcal{L}_{reg}(G_S, D_S) + \mathcal{L}_{reg\_f}(f), \quad (15)
\end{aligned}$$

where $\mathcal{L}_{reg}(G_T, D_T)$, $\mathcal{L}_{reg}(G_S, D_S)$ and $\mathcal{L}_{reg\_f}(f)$ are the regularization terms to prevent the learned parameters of G_T, D_T, G_S, D_S and f from overfitting. They are defined by

$$\mathcal{L}_{reg}(G_T, D_T) = = \sum_{i=1}^{n_G} \|W_{G_T}^i\|_F^2 + \sum_{i=1}^{n_D} \|W_{D_T}^i\|_F^2, \quad (16)$$

$$\mathcal{L}_{reg}(G_S, D_S) = \sum_{i=1}^{n_G} \|W_{G_S}^i\|_F^2 + \sum_{i=1}^{n_D} \|W_{D_S}^i\|_F^2, \quad (17)$$

$$\mathcal{L}_{reg\_f}(f) = \sum_{i=1}^{n_f} \|W_f^i\|_F^2, \quad (18)$$

where $W_{G_T}^i$, $W_{D_T}^i$, $W_{G_S}^i$, $W_{D_S}^i$ and W_f^i represent the layer-wise parameters of G_T, D_T, G_S, D_S and f, respectively. n_G, n_D and n_f denote the layer numbers of the generator, the discriminator and the classifier, respectively. This ultimately corresponds to solving for the bidirectional mappings between source image and target video (G_T, D_T, G_S, D_S) as well as the cross-domain classifier f according to the optimization problem:

$$\begin{aligned}
&(f^*, G_T^*, D_T^*, G_S^*, D_S^*) \\
&\quad = \arg \min_{G_T, G_S, f} \max_{D_T, D_s} \mathcal{L}(f, X_S, Y_S, X_F, X_T, G_T, D_T, G_S, D_S). \quad (19)
\end{aligned}$$

As in the standard GAN framework, we solve this minimax problem iteratively by first training D_T and D_S with the fixed G_T, G_S and f, and then training G_T, G_S and f with the learned D_T and D_S. The detailed training procedures are illustrated in Algorithm 1.

D. Prediction

In the testing phrase as shown in Figure 1, an input video is firstly divided into several video clips and the feature of each clip $\mathbf{x}_t$ is extracted by the C3D model [47]. Each video clip feature is then mapped into the image-frame feature space via the generator G_S to generate the image-frame feature $G_S(\mathbf{x}_t)$. Next, $G_S(\mathbf{x}_t)$ is transformed back to the video feature space via the generator G_T to produce the new discriminative video clip feature $G_T(G_S(\mathbf{x}_t))$. Thus, the augmented feature of each video clip is given by $[G_S(\mathbf{x}_t); G_T(G_S(\mathbf{x}_t))]$. Next, average across the augmented features of all the video clips for generating the final augmented feature of the whole video. At last, the augmented video feature is fed to the classification network to predict the action class label.

E. Discussion

Our Sym-GANs is most related to HiGAN [55] and SBADA-GAN [45]. The HiGAN and our Sym-GANs focus on the same task of exploiting images for video recognition. Although HiGAN and Sym-GANs use the same JAN and C3D to extract the image and video features, respectively, they resort to totally different network architectures to solve the heterogeneous image-to-video domain adaptation. The main differences between HiGAN and Sym-GANs are: (1) HiGAN maps target video to source image for adaptation, without capturing the motion information within the videos for classification. While Sym-GANs learns bidirectional mapping between source image and target video to generate the domain-invariant augmented feature which can represent both static appearance and dynamic motion information for video classification; (2) In HiGAN, the two level GANs are learned in a step-by-step manner and the MKL method is used to train classifiers. Different from the separate learning of the two-level GANs and classifiers in HiGAN, the symmetric GANs and the classification network are jointly trained in an end-to-end manner in Sym-GANs, which can learn more discriminative and transferable feature for domain adaptation. (3) Thanks to the augmented feature learned in an end-to-end manner, our Sym-GANs achieves better results than HiGAN, as shown in Table III.

Although the bi-directional mapping between source image and target video in our Sym-GANs is similar to the bi-directional image transformation between different image domains in SBADA-GAN to some extent, Sym-GANs and SBADA-GAN focus on different tasks using different network architectures with different losses. The main differences between Sym-GANs and SBADA-GAN are: (1) SBADA-GAN addresses the image-to-image adaptation on the pixel level where the source and target images are with the same size. In contrast, Sym-GANs investigates the heterogeneous image-to-video adaptation on the feature level where the source

Algorithm 1 Symmetric GANs for Image-to-Video Adaptation

```
Input: Source images represented by the image-frame features $X_S = \{\mathbf{x}_s^i\}_{i=1}^{n_s}$ with labels $Y_S = \{y_s^i\}_{i=1}^{n_s}$
    Target video clips represented by the video features $X_T = \{\mathbf{x}_t^i\}_{i=1}^{n_t}$
    Target video frames represented by the image-frame features $X_F = \{\mathbf{x}_f^i\}_{i=1}^{n_t}$
Output: The mapping from the image-frame feature to the video feature $G_T$ associated with the discriminator $D_T$
    The mapping from the video feature to the image-frame feature $G_S$ associated with the discriminator $D_S$
    The cross-domain classifier $f$
Initialize $G_T, D_T, G_S, D_S, f$
repeat
    Update $D_T, D_S$ with fixed $G_T, G_S, f$ using the following optimization:
        $\max_{D_T, D_S} \mathcal{L}_{GAN}(G_T, D_T, X_F, X_T) + \mathcal{L}_{GAN}(G_S, D_S, X_T, X_F) + \mathcal{L}_{reg}(G_T, D_T) + \mathcal{L}_{reg}(G_S, D_S)$
        • $\mathcal{L}_{GAN}(G_T, D_T, X_F, X_T) = \frac{1}{n_t} \sum_{i=1}^{n_t} \log D_T(\mathbf{x}_t^i)^2 + \log(1 - D_T(G_T(\mathbf{x}_f^i))^2)$
        • $\mathcal{L}_{GAN}(G_S, D_S, X_T, X_F) = \frac{1}{n_t} \sum_{i=1}^{n_t} \log D_S(\mathbf{x}_f^i)^2 + \log(1 - D_S(G_S(\mathbf{x}_t^i))^2)$
        • $\mathcal{L}_{reg}(G_T, D_T) = \sum_{i=1}^{n_D} \|W_{D_T}^i\|_F^2$
        • $\mathcal{L}_{reg}(G_S, D_S) = \sum_{i=1}^{n_D} \|W_{D_S}^i\|_F^2$
    Update $G_T, G_S, f$ with fixed $D_T, D_S$ using the following optimization:
        $\min_{G_T, G_S, f} \mathcal{L}_{GAN}(G_T, D_T, X_F, X_T) + \mathcal{L}_{GAN}(G_S, D_S, X_T, X_F) + \lambda_1 \mathcal{L}_{CORAL}(X_T, G_T(X_F)) + \lambda_2 \mathcal{L}_{CORAL}(X_F, G_S(X_T))$
        $+ \mathcal{L}_{class}(f, \hat{X}_S, Y_S) + \mathcal{L}_{reg}(G_T, D_T) + \mathcal{L}_{reg}(G_S, D_S) + \mathcal{L}_{reg\_f}(f)$
        • $\mathcal{L}_{GAN}(G_T, D_T, X_F, X_T) = \frac{1}{n_t} \sum_{i=1}^{n_t} \log(1 - D_T(G_T(\mathbf{x}_f^i))^2)$
        • $\mathcal{L}_{GAN}(G_S, D_S, X_T, X_F) = \frac{1}{n_t} \sum_{i=1}^{n_t} \log(1 - D_S(G_S(\mathbf{x}_t^i))^2)$
        • $\mathcal{L}_{CORAL}(X_T, G_T(X_F))$ is calculated by Eq. 8
        • $\mathcal{L}_{CORAL}(X_F, G_S(X_T))$ is calculated by Eq. 11
        • $\mathcal{L}_{class}(f, \hat{X}_S, Y_S) = -\frac{1}{n_s} \sum_{i=1}^{n_s} \sum_{k=1}^{C} \mathbb{I}_{k=y_s^i} \log f([G_S(G_T(\mathbf{x}_s^i)); G_T(\mathbf{x}_s^i)])_{y_s^i}$
        • $\mathcal{L}_{reg}(G_T, D_T) = \sum_{i=1}^{n_G} \|W_{G_T}^i\|_F^2$
        • $\mathcal{L}_{reg}(G_S, D_S) = \sum_{i=1}^{n_G} \|W_{G_S}^i\|_F^2$
        • $\mathcal{L}_{reg\_f}(f) = \sum_{i=1}^{n_f} \|W_f^i\|_F^2$
until $G_T, D_T, G_S, D_S, f$ converge
```

image and target video are represented by different types of features with different feature dimensions; (2) SBADA-GAN linearly combines the outputs of the source and target classifiers for image classification on the target domain and the combination weights need to be chosen on a validation set in each setting. While Sym-GANs proposes feature augmentation to automatically learn domain-invariant feature on which the source classifier can adapt well to the target domain.

IV. Experiments

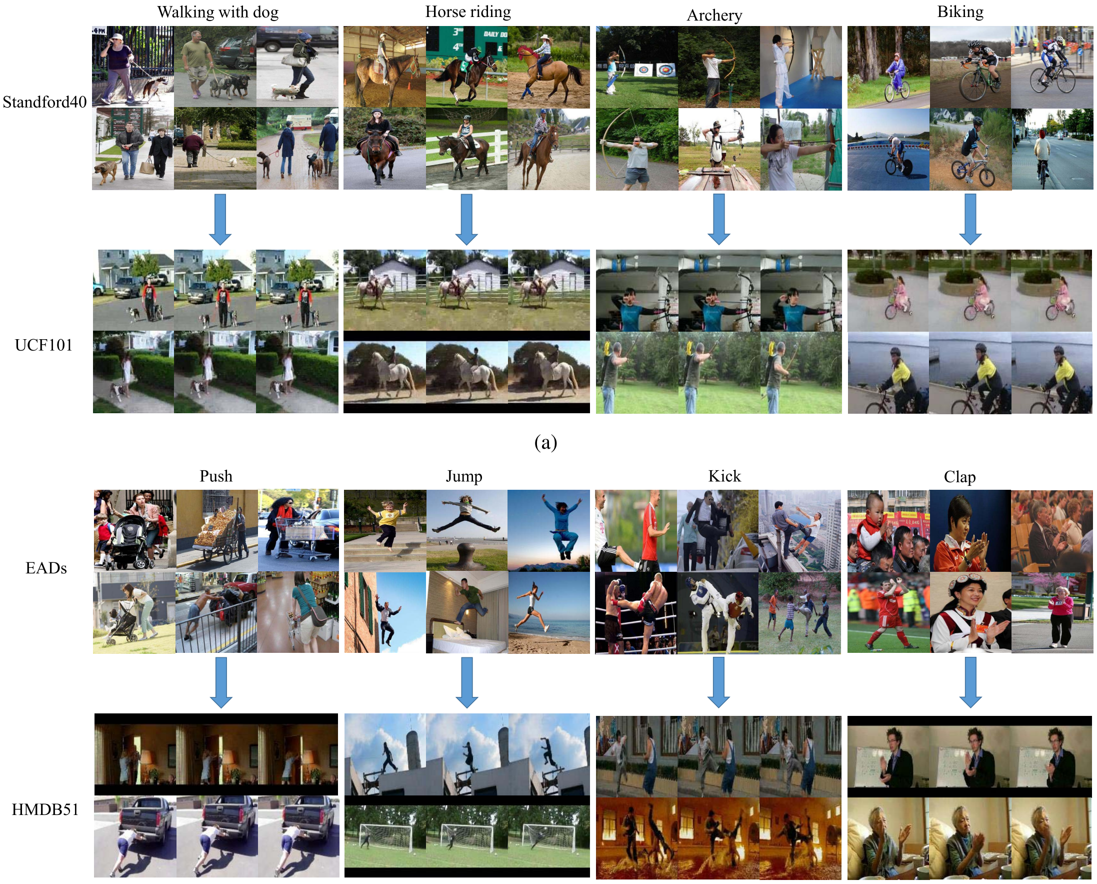
A. Datasets

The experiments are conducted on two video benchmarks, i.e., UCF101 (U) [56] and HMDB51 (H) [57], to evaluate the performance of the proposed method. We report the mean of classification accuracies for all methods. For the UCF101 as the target video domain, the source images come from the Stanford40 (S) dataset [58]. For the HMDB51 as the target video domain, the source image domain consists of Standford40 dataset and HII dataset [59], denoted by EADs (E) dataset. So there are two image-to-video adaptation tasks: S→U and E→H.

The **UCF101** dataset is an action database of videos collected from YouTube. It has 13000 videos with 101 action categories. There are mainly five typical kinds of actions, including sports, playing musical instruments, body-motion only, human-object interaction and human-human interaction. This dataset is challenging for action recognition since most of the videos are recorded in realistic scenes with large variations in illumination, cluttered background, object appearance, motion style, viewpoint and camera movement.

The **HMDB51** dataset has around 7,000 video clips with manual annotations covering 51 action categories. These clips are extracted from various sources, including commercial movies and public datasets (e.g., YouTube and Google videos). This dataset provides adequate variety, where the action videos can be mainly divided into five types, ranging from the general body movements such as push and kick to the fine-grained facial expressions like laugh and smile. In comparison with the UCF101 dataset, HMDB51 dataset involves more cluttered background and larger intra-class variations, for the reason that it presents a fine multifariousness of surroundings, situations and light conditions.

The **Stanford40** dataset contains images of humans performing 40 diverse daily actions. All the images are obtained from Google, Bing, and Flickr. Each action category has 180 to 300 images with large variations in human pose, appearance and background clutter.



(a)

(b)

Fig. 2. Several examples of source images and target video frames on the tasks of S→U (a) and E→H (b). For each task, the upper part shows the images from the source domain, and the lower part shows the video frames from the target domain. (a) S→U. (b) E→H.

The **EADs** dataset consists of Stanford40 and HII datasets. The HII dataset has a total of 1972 images with 10 action classes, and each class contains at least 150 images.

For the **S→U** task, 12 common action categories are chosen between the UCF101 and Stanford40 datasets, including "brushing teeth", "cleaning floor", "climbing", "cutting vegetables", "playing guitar", "playing violin", "biking", "horse riding", "rowing", "shooting", "walking with dog" and "writing on a board". The source domain is comprised of all the labeled images from the Stanford40 dataset. The unlabeled videos from the UCF101 dataset construct the target domain. The UCF101 dataset (target domain) is split into two parts: one for training and the other for test.

For the **E→H** task, all the labeled images from the EDAs dataset construct the source domain. The target domain consists of unlabeled videos from the HMDB51 dataset. There are 13 shared action categories between these two datasets, including "clap", "climb", "drink", "hug", "jump", "kick", "kiss", "pour", "push", "run", "smoke", "talk" and "wave". Similar to the UCF101 dataset, the HMDB51 dataset (target domain) is also split into training and test parts. Some examples of images and videos on these four datasets are shown in Figure 2.

The evaluation protocol is the same for the two image-to-video adaptation tasks. For the target domains (i.e., UCF101 and HMDB51 datasets), since they both provide three splits of training and test sets, the average classification accuracy over these three splits is reported for evaluation. It is worth emphasizing that any labels of the target data are not used during the training phase.

B. Setup

1) Features: We split each video into 16-frame clips without overlap and extract a 512D feature vector of each video clip from the pool5 layer of 3D CoveNets [47] which is trained on the Sports-1M dataset [60]. The clip features from the target training videos construct the training data of target domain. For each video clip, we randomly sample one frame and all the frames from all the video clips compose the video frame domain. We utilize the JAN [5] method based on the ResNet [61] to learn the transferable image-frame feature between source images and target video frames, which comes from the *pool5* layer of JAN with the dimension of 2048.

2) Implementation details:

a) Network architecture: To build the two generators G_T and G_S in Sym-GANs, we use three-layered feed-forward neural networks activated by relu function: $2048 \rightarrow 1024 \rightarrow 1024 \rightarrow 512$ for G_T and $512 \rightarrow 1024 \rightarrow 2048 \rightarrow 2048$ for G_S. For the two discriminators D_S and D_T, we both use two fully connected layers ($2560 \rightarrow 640 \rightarrow 1$) activated by relu function, except for the last layer. For the classifier f, we use four-layered feed-forward neural networks (2560 → 1280 → 640 → 256 → the number of categories), activated by relu function, except for the last layer.

b) Training detail: We employ the Adam solver [62] with a batch size of 128. All the networks are trained from scratch with the learning rate of 0.00008. Since the two GANs in Sym-GANs are symmetric, we set $\lambda_1 = \lambda_2$ for the CORAL losses to simplify the parameter selection. Since the adversarial and CORAL losses have different orders of magnitude, we set $\lambda 1 = \lambda 2 = 100$ to somehow balance them.

3) Baseline methods:

a) Homogeneous Domain Adaptation: In order to validate the effectiveness of capturing temporal motion information for image-to-video adaptation in our method, we compare our method with the existing homogeneous domain adaptation methods where each target video is represented by the static image feature without considering the temporal relationship between sequential. The homogeneous domain adaptation methods include traditional shallow and deep methods, where the source and target data are represented by the same type of feature. For traditional shallow methods, source image is represented by the feature extracted from the *pool5* layer of ResNet. Each target video is represented by the mean of the ResNet features of all the frames. For deep methods, we take source images and target frames as input to train the networks. At test time, the output scores (from the last *fc* layer) of all the frames within a video are further averaged to determine the class label of the video. The homogeneous domain adaptation methods are listed below.

- **SVM** [63] trained on the labeled source data is used as a baseline without domain adaptation.
- **GFK** [64] proposes a geodesic flow kernel to leverage low-dimensional feature structures.
- **JDA** [20] aims to jointly adapt both marginal and conditional distributions in a principled dimension reduction procedure, and generate new feature representation with good robustness to substantial distribution difference.
- **ARTL** [65] learns an adaptive classifier by modeling the distribution adaptation and label propagation in a unified framework based on the regularization theory and the structural risk minimization principle.
- **TJM** [18] designs a principled dimensionality reduction method to simultaneously perform feature matching and reweighting instances across domains for domain adaptation.
- **TKL** [21] introduces a kernel-based transfer learning method that learns a invariant kernel to different domains by straightly aligning the distributions of source and target domains in the reproducing kernel Hilbert space.
- **CORAL** [34] diminishes the domain shift by matching the feature distributions of the source and target domains with the respect of their second-order statistics.
- **LRSR** [22] transforms both the source and target data to a shared feature space, in which source samples can be effectively combined to represent each sample in the target domain.
- **BDA** [19] adaptively weights the importance of both marginal and conditional distribution adaptations.
- **ATI** [66] proposes an iterative method for domain adaptation which iteratively label the target samples and compute a source-to-target mapping by minimizing the distances of the assignments.
- **MEDA** [67] uses the principle of structural risk minimization to learn a domain-invariant classifier in Grassmann manifold, and simultaneously aligns the distributions of different domains in a dynamic manner to quantitatively calculate the relative importance of marginal and conditional distributions.
- **ResNet** [61] is performed on the labeled source data as a baseline without domain adaptation.
- **DAN** [30] introduces multiple kernel learning to match the mean embeddings of the source and target data distributions for reducing the domain discrepancy in higher task-specific layers of deep neural networks.
- **RTN** [31] jointly learns adaptive classifiers and transferable features with the assumption that the main difference between the source and target classifiers is formulated a residual function.
- **DANN** [38] enjoins the hidden layers of deep neural networks to learn feature representations that can not be distinguished between the source and target domains.
- **JAN** [5] reduces the domain shift by aligning the joint distributions of multiple domain-specific layers which is implemented by jointly maximizing mean discrepancies of feature distributions of these layers
- **DAL** [28] introduces a new domain adaptation layer to reduce the domain discrepancy by aligning source and target distributions to a reference one.
- **WGRL** [68] employs the Wasserstein distance to measure the domain discrepancy between the source and target data which is calculated by a neural network, and minimizes this distance in an adversarial fashion to learn the feature representations with superior transferability.

b) Heterogeneous Domain Adaptation: To evaluate the superiority of learning augmented feature using symmetric GANs on image-to-video adaptation, our method is also compared with several existing heterogeneous domain adaptation methods where the source and target data are represented by different types of features. Specifically, the source images are represented by image-frame features extracted by the finetuned JAN model and the traget videos are represented by C3D features. The compared heterogeneous domain adaptation methods are listed as follows:

- **KCCA** [23] applies kernel method to canonical correlation analysis for heterogeneous domain adaptation.
- **HEMAP** [46] transforms both source and target data into a common subspace using a spectral embedding method and incorporates a sample selection algorithm to select related source samples for further improving the adaptation performance.
- **DAMA** [25] proposes a manifold alignment based approach to construct mappings for linking feature spaces of different domains.

TABLE I

COMPARISON OF CLASSIFICATION ACCURACY (%) BETWEEN OUR METHOD AND THE HOMOGENEOUS DOMAIN ADAPTATION METHODS

Method	S→U	E→H
SVM [63]	83.6	35.9
GFK [64]	80.7	29.1
JDA [20]	86.7	33.8
ARTL [65]	89.5	39.9
TJM [18]	90.5	31.2
TKL [21]	90.9	38.9
CORAL [34]	91.5	39.3
LRSR [22]	89.3	38.0
ATI [66]	90.4	32.8
BDA [19]	83.8	31.0
MEDA [67]	94.3	43.1
ResNet [61]	81.4	38.5
DAN [30]	84.2	39.5
RTN [31]	83.8	40.2
DANN [38]	85.9	38.4
JAN [5]	91.4	40.9
DAL [28]	97.6	45.5
WGRL [68]	91.3	40.4
Ours	**97.7**	**55.0**

- **HFA** [24] first augments the heterogeneous features and then finds the two projection matrices to deal with the augmented features.
- **CDLS** [26] resorts to learning typical cross-domain landmarks to generate a good feature space for transferring knowledge between heterogeneous domains.
- **HiGAN** [55] combines a low-level conditional GAN and a high-level conditional GAN to learn a domain-invariant feature representation between source images and target videos.

Since the methods of KCCA, HEMAP, DAMA, HFA and CDLS should require some labeled data in the target domain, we assign labels to 80 target videos per category for training. Note that our method can be easily extended to the semi-supervised heterogeneous domain adaptation for fair comparison, where we take labeled target videos into account when training classifier. The HiGAN can work in both unsupervised and semi-supervised scenarios, so we compare our method with it in both unsupervised and semi-supervised settings.

C. Results

1) Comparison With Homogeneous Domain Adaptation Methods: Table I shows the comparison results between homogeneous domain adaptation methods and our method. The upper part reports the results of traditional shallow methods, the middle part reports the results of deep methods and the last row is the result of our method. From Table I, we can have the following observations. (1) Compared with those homogeneous methods which extract static frame features for video representation, our method achieves much better classification performance, clearly demonstrating the benefit of the proposed feature augmentation adaptation strategy by simultaneously capturing the temporal motion information between sequential frames and static appearance features within individual frames. (2) On the S→U task, almost all the traditional shallow methods and deep methods perform better than the baseline SVM and ResNet, respectively, which indicates that leveraging images can improve video classification performance to some extend. While on the E→H task, half of traditional shallow methods substantially underperform the baseline SVM, which might result from the large difference between source images and target video frames, leading to negative transfer [69]. On the other hand, deep methods outperform the baseline ResNet, which verifies that deep neural networks work better at addressing the negative transfer issue. (3) It is interesting to observe that traditional shallow domain adaptation methods achieve comparable results with deep methods on the easy task of S→U and perform worse than deep methods on the difficult task of E→H. This further implies that deep models are excellent in handling more challenging domain adaptation, benefited from the end-to-end learning of domain-invariant feature and classifier.

TABLE II

COMPARISON OF CLASSIFICATION ACCURACY (%) BETWEEN OUR METHOD AND THE HETEROGENEOUS DOMAIN ADAPTATION METHODS

Method	S→U	E→H
Target only	94.0	65.7
KCCA [23]	92.0	66.0
HEMAP [46]	92.7	65.4
DAMA [25]	93.5	67.9
HFA [24]	93.9	69.9
CDLS [26]	94.9	65.4
Ours	**99.7**	**89.1**

TABLE III

COMPARISON OF CLASSIFICATION ACCURACY (%) BETWEEN OUR METHOD AND THE HIGAN

Method	S→U	E→H
HiGAN[55] (unsupervised)	95.4	44.6
Ours (unsupervised)	**97.7**	**55.0**
HiGAN[55] (semi-supervised)	98.0	74.0
Ours (semi-supervised)	**99.7**	**89.1**

2) Comparison With Heterogeneous Domain Adaptation Methods: Table II demonstrates the classification accuracies of heterogeneous domain adaptation methods. The first row reports the results of classifiers only trained on the target domain. Apparently, our approach outperforms all the comparison methods on both tasks. Note that the HFA method is more related to our method, which also augments features for heterogeneous domain adaptation. The higher classification accuracies achieved by our method than HFA clearly confirms that the deep learning based Sym-GANs can learn domain-invariant augmented feature with more powerful transferability and distinguishability for heterogeneous image-to-video adaptation, producing a significant boost in performance. In Table III, we compare our method with the HiGAN method which can work in both unsupervised and semi-supervised scenarios. The encouraging results highlight the key importance of symmetric architecture as well as feature augmentation for effective heterogeneous domain adaptation.

D. Ablation Study

To analyze the proposed approach in depth, ablation study is conducted for empirically evaluating the importance of

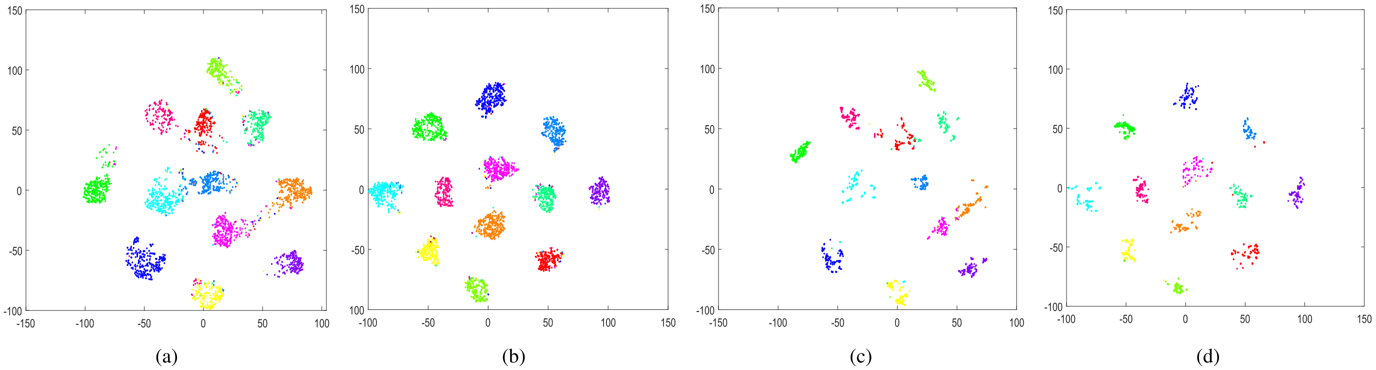


Fig. 3. Feature visualization: t-SNE of Ours-original feature representations on the source domain (a) and the target domain (c); t-SNE of our feature representations on the source domain (b) and the target domain (d). Different colors denote 12 different action categories. (a) Ours-orginal: *Source*=**S**. (b) Ours: *Source*=**S**. (c) Ours-orginal: *Target*=**U**. (d) Ours: *Target*=**U**.

TABLE IV

COMPARISON OF CLASSIFICATION ACCURACY (%) BETWEEN OUR METHOD AND THE SINGLE GAN

Method	S→U	E→H
Only G_T	87.3	25.3
Only G_S	92.0	50.2
Ours	**97.7**	**55.0**

TABLE V

CLASSIFICATION ACCURACY (%) OF DIFFERENT LOSS FUNCTIONS

Method	S→U	E→H
w/o Adversarial Loss	90.1	25.0
w/o CORAL Loss	96.3	42.8
Replace CORAL with L1	92.5	40.4
Replace CORAL with L2	93.9	39.4
Ours (CORAL)	**97.7**	**55.0**

each individual component. To prove the effectiveness of symmetrical structure in Sym-GANs, we perform an ablation study and report the results on both tasks in Table IV. Here we compare our method with other two single-GAN variants: only G_T learns the mapping from the image-frame feature to the video feature and only G_S learns the mapping from the video feature to the image-frame feature. From Table IV, we conclude that G_T and G_S are complementary, as their fusion significantly improves the recognition performance on both tasks.

In Table V, we analyze the effectiveness of different loss functions. We compare our method with other four variations: without adversarial loss, without CORAL loss, replacing CORAL loss with L1 loss and L2 loss, respectively. It is obvious that the recognition results will substantially degrade when removing the adversarial loss or the CORAL loss, indicating that both these two losses are critical to the overall performance. Compared with the two distance losses (i.e., L1 and L2), the CORAL loss achieves the highest accuracy, which demonstrates the effectiveness of aligning the distributions of the generated and real features by exploring their second-order statistics in our method.

In Table VI, we compare our method with another feature augmentation method, called Ours-original. In this method, the augmented features for the source and target domains are represented by $\hat{\mathbf{x}}_s = [\mathbf{x}_s; G_T(\mathbf{x}_s)]$ and $\hat{\mathbf{x}}_t = [G_S(\mathbf{x}_t); \mathbf{x}_t]$, respectively. In our method, the augmented source and target features are represented by $\hat{\mathbf{x}}_s = [G_S(G_T(\mathbf{x}_s)); G_T(\mathbf{x}_s)]$ and $\hat{\mathbf{x}}_t = [G_S(\mathbf{x}_t); G_T(G_S(\mathbf{x}_t))]$, respectively. It is interesting to notice that our method outperforms the Ours-original, proving that the generated features $G_S(G_T(\mathbf{x}_s))$ and $G_T(G_S(\mathbf{x}_t))$ are more discriminative than their corresponding original features $\mathbf{x}_s$ and $\mathbf{x}_t$, respectively.

TABLE VI

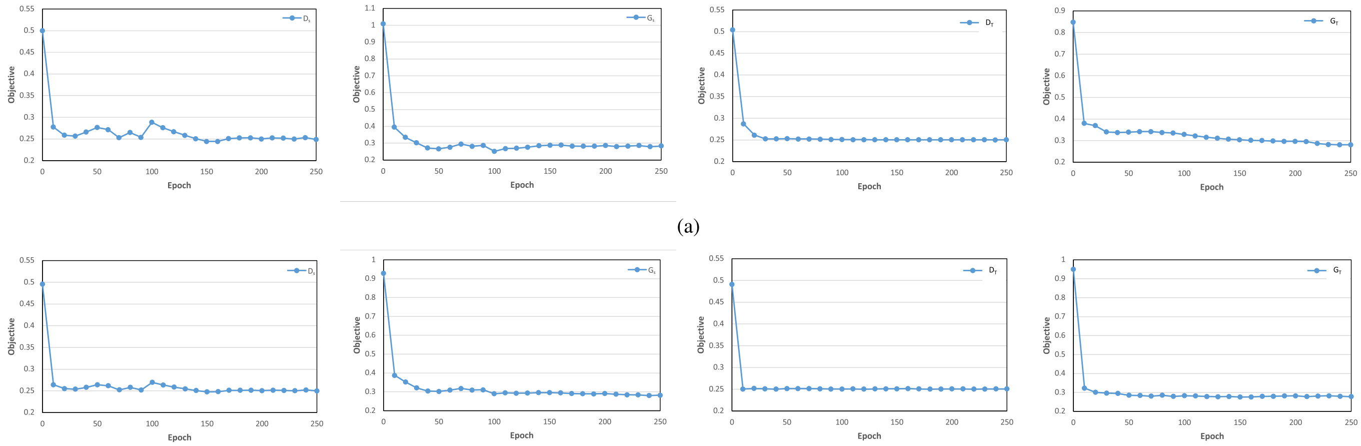
COMPARISON OF CLASSIFICATION ACCURACY (%) BETWEEN THE ORIGINAL FEATURES AND THE GENERATED NEW FEATURES

Method	S→U	E→H
Ours-orginal	93.5	50.9
Ours	**97.7**	**55.0**

To further demonstrate the transferability and distinguishability of augmented features (i.e., $\hat{\mathbf{x}}_s$ and $\hat{\mathbf{x}}_t$), we visualize in Figure 3(a)-3(d) the t-SNE embeddings [70] of augmented features of source images and target videos learned by the Ours-original and our method, respectively, on the $S \rightarrow U$ task. Figure 3(a) and 3(c) demonstrate the source and target features of Ours-original, respectively. Similarly, Figure 3(b) and 3(d) show the source and target features learned by our method, respectively. Compared with the augmented features learned by the Ours-original, the features of the our method become more clear to be categorized, which suggests that the augmented features of our method are more discriminative than that of the Ours-original, since the generators G_S and G_T are jointly learned with the classifier with the supervision information from the labeled source data. Besides, the source and target domains are aligned better in our method, which shows the superior transferability of the augmented features learned by our method.

E. Convergence Performance

Figure 4 shows the convergence performance of training discriminators of D_s and D_t as well as generators of G_s and G_t in Sym-GANs. It can be observed that Sym-GANs



(a)

(b)

Fig. 4. Convergence performance of training Sym-GANs (i.e., D_S, G_S, D_T and G_T) on tasks of S→U (a) and E→H (b). (a) Convergence performance on S→U. (b) Convergence performance on E→H.

can reach a steady performance on both datasets and gradually converge, which indicates the training advantage of Sym-GANs in cross-domain tasks.

V. Conclusion

This paper mainly solves the problem of heterogeneous domain adaptation from images to videos for video recognition, where the image and video are represented by different types of features. We have proposed a symmetric adversarial learning approach to learn domain-invariant augmented feature for heterogeneous image-to-video adaptation. To this end, two generative adversarial networks have been built to model the bidirectional mappings between source images and target videos, which also generates the augmented features of both image and video features. These augmented features can preserve both static appearance and temporal motion information with superior transferable, distinguishable and descriptive abilities. Moreover, a joint optimization algorithm has been presented to train the Symmetric GANs and the classifier simultaneously. Extensive experiments on the challenging UCF101 and HMDB51 video datasets validate the effectiveness of the proposed method on exploiting images for video recognition.

References

[1] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, “Attention transfer from Web images for video recognition,” in *Proc. ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1–9.

[2] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff, “Do less and achieve more: Training CNNs for action recognition utilizing action images from the Web,” *Pattern Recognit.*, vol. 68, pp. 334–345, Aug. 2017.

[3] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei, “You lead, we exceed: Labor-free video concept learning by jointly exploiting Web videos and images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 923–932.

[4] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[5] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 2208–2217.

[6] Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, “Transfer tagging from image to video,” in *Proc. 19th ACM Int. Conf. Multimedia*, Nov. 2011, pp. 1137–1140.

[7] L. Duan, D. Xu, and S.-F. Chang, “Exploiting Web images for event recognition in consumer videos: A multiple source domain adaptation approach,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1338–1345.

[8] H. Wang, X. Wu, and Y. Jia, “Video annotation via image groups from the Web,” *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1282–1291, Aug. 2014.

[9] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, “Temporal localization of fine-grained actions in videos by domain transfer from Web images,” in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 371–380.

[10] C. Gan, C. Sun, L. Duan, and B. Gong, “Webly-supervised video recognition by mutually voting for relevant web images and Web video frames,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 849–866.

[11] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, “Show, adapt and tell: Adversarial training of cross-domain image captioner,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 521–530.

[12] J. Hoffman *et al.*, “CYCADA: Cycle-consistent adversarial domain adaptation,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1994–2003.

[13] L. Duan, I. W. Tsang, and D. Xu, “Domain transfer multiple kernel learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, Mar. 2012.

[14] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, “Visual event recognition in videos by learning from Web data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1667–1680, Sep. 2012.

[15] H. Song, X. Wu, W. Yu, and Y. Jia, “Extracting key segments of videos for event detection by learning from Web sources,” *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1088–1100, May 2018.

[16] X. Wu, H. Wang, C. Liu, and Y. Jia, “Cross-view action recognition over heterogeneous feature spaces,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4096–4108, Nov. 2015.

[17] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, “Domain adaptive faster R-CNN for object detection in the wild,” 2018, arXiv: 1803.03243. [Online]. Available: https://arxiv.org/abs/1803.03243

[18] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer joint matching for unsupervised domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1410–1417.

[19] J. Wang, Y. Chen, S. Hao, W. Feng, and Z. Shen, “Balanced distribution adaptation for transfer learning,” in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1129–1134.

[20] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2200–2207.

[21] M. Long, J. Wang, J. Sun, and P. S. Yu, “Domain invariant transfer kernel learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1519–1532, Jun. 2014.

[22] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, “Discriminative transfer subspace learning via low-rank and sparse representation,” *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 850–863, Feb. 2016.

[23] S. Akaho, “A kernel method for canonical correlation analysis,” 2006, arXiv: 0609071. [Online]. Available: https://arxiv.org/abs/cs/0609071

[24] L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for heterogeneous domain adaptation," 2012, arXiv: 1206.4660. [Online]. Available: https://arxiv.org/abs/1206.4660

[25] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jun. 2011, vol. 22, no. 1, p. 1541.

[26] Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5081–5090.

[27] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1785–1792.

[28] F. M. Cariucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò, "Auto-DIAL: Automatic domain alignment layers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5077–5085.

[29] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, arXiv: 1412.3474. [Online]. Available: https://arxiv.org/abs/1412.3474

[30] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, May 2015, pp. 97–105.

[31] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 136–144.

[32] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1180–1189.

[33] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 443–450.

[34] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 13th AAAI Conf. Artif. Intell. (AAAI)*, Mar. 2016, pp. 2058–2065.

[35] W.-Y. Chen, T.-M. H. Hsu, Y.-H. H. Tsai, Y.-C. F. Wang, and M.-S. Chen, "Transfer neural trees for heterogeneous domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 399–414.

[36] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[37] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.

[38] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, Jan. 2016.

[39] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," 2018, arXiv: 1803.09210. [Online]. Available: https://arxiv.org/abs/1803.09210

[40] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 770–778.

[41] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 95–104.

[42] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, arXiv: 1611.02200. [Online]. Available: https://arxiv.org/abs/1611.02200

[43] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 469–477.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[45] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: Symmetric bi-directional adaptive GAN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8099–8108.

[46] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2010, pp. 1049–1054.

[47] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[48] X. Huang, M.-Y. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," 2018, arXiv: 1804.04732. [Online]. Available: https://arxiv.org/abs/1804.04732

[49] J. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 465–476.

[50] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 700–708.

[51] P. Luc, C. Couprie, S. Chintala, and J., "Semantic segmentation using adversarial networks," 2016, arXiv: 1611.08408. [Online]. Available: https://arxiv.org/abs/1611.08408

[52] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.

[53] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1951–1959.

[54] X. Mao, Q. Li, H. Xie, R. Y. Lau, and Z. Wang, "Least squares generative adversarial networks," 2016, arXiv: 1611.04076. [Online]. Available: https://arxiv.org/abs/1611.04076

[55] F. Yu, X. Wu, Y. Sun, and L. Duan, "Exploiting images for video recognition with hierarchical generative adversarial networks," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 1107–1113.

[56] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, arXiv: 1212.0402. [Online]. Available: https://arxiv.org/abs/1212.0402

[57] H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre, "HMDB51: A large video database for human motion recognition," in *High Performance Computing in Science and Engineering*. Berlin, Germany: Springer, 2013, pp. 571–582.

[58] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1331–1338.

[59] G. Tanisik, C. Zalluhoglu, and N. Ikizler-Cinbis, "Facial descriptors for human interaction recognition in still images," *Pattern Recognit. Lett.*, vol. 73, pp. 44–51, Apr. 2016.

[60] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1725–1732.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv: 1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

[63] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.

[64] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2066–2073.

[65] M. Long, J. Wang, G. Ding, S. J. Pan, and P. S. Yu, "Adaptation regularization: A general framework for transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1076–1089, May 2014.

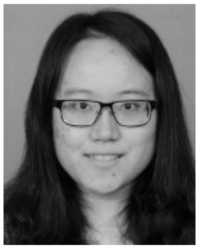
[66] P. P. Busto and J. Gall, "Open set domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 754–763.

[67] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proc. ACM Multimedia Conf. Multimedia Conf.*, Oct. 2018, pp. 402–410.

[68] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. 32nd Conf. Artif. Intell. (AAAI)*, 2018, pp. 4058–4065.

[69] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[70] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.



Feiwu Yu received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2016, where she is currently pursuing the M.S. degree with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science. Her research interests include action recognition and transfer learning.

Xinxiao Wu (M'09) received the B.A. degree in computer science from the Nanjing University of Information Science and Technology in 2005, and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. She was a Post-Doctoral Research Fellow at Nanyang Technological University, Singapore, from 2010 to 2011. She is currently an Associate Professor with the School of Computer Science, Beijing Institute of Technology. Her current research interests include machine learning, computer vision, and video analysis and understanding.

Jialu Chen received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2018, where she is currently pursuing the master's degree with the School of Computer Science. Her research interests include domain adaptation and deep learning.

Lixin Duan received the B.Eng. degree from the University of Science and Technology of China in 2008, and the Ph.D. degree from Nanyang Technological University in 2012. He is currently a Full Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include machine learning algorithms (especially in transfer learning and domain adaptation) and their applications in computer vision. He was a recipient of the Microsoft Research Asia Fellowship in 2009 and the Best Student Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition 2010.