Deep Multi-Modality Adversarial Networks for Unsupervised Domain Adaptation

Xinhong Ma, Tianzhu Zhang, *Member, IEEE,* Changsheng Xu, *Fellow, IEEE*

***Abstract*—Unsupervised Domain Adaptation (UDA) aims to transfer domain knowledge from existing well-defined tasks to new ones where labels are unavailable. In the real world applications, domain discrepancy is usually uncontrollable especially for multi-modality data. Therefore, it is significantly motivated to deal with multi-modality domain adaptation task. As labels are unavailable in target domain, how to learn semantic multi-modality representations and how to successfully adapt the classifier from source to target domain, remain open challenges in multi-modality domain adaptation task. To deal with these issues, we propose a Multi-Modality Adversarial Network (MMAN), which applies stacked attention to learn semantic multi-modality representations and reduces domain discrepancy via adversarial training. Unlike the previous domain adaptation methods which cannot make full use of source domain categories information, multi-channel constraint is employed to capture fine-grained categories knowledge that could enhance the discrimination of target samples and boost target performance on single-modality and multi-modality domain adaptation problem. We apply the proposed MMAN to two applications including cross domain object recognition and cross domain social event recognition. The extensive experimental evaluations demonstrate the effectiveness of the proposed model for unsupervised domain adaptation.**

***Index Terms*—Unsupervised Domain Adaptation, Triplet Loss, Stacked Attention, Multi-Modality, Social Event Recognition.**

I. Introduction

Recently, deep networks have significantly improved many state-of-the-art algorithms for diverse machine learning problems and applications [1–5]. Note that, its impressive performance is guaranteed only when massive labeled training data available in training process. However, the cost of annotating labeled data is often an obstacle for applying deep networks. Furthermore, for problems lacking labeled data, it is possible to train deep models in similar domains with enough training data, but shift between train and test data distribution may lead to poor performance. To address above issues, some researchers have attempted to explore unlabeled data, referred as semi-supervised learning and transfer learning. It is different that semi-supervised learning focuses on training on labeled and unlabeled data in the same domain, while transfer learning explores data or models from unlabeled target domain and labeled source domain. Domain Adaptation (DA) is a particular case of transfer learning (TL) that leverages labeled data in one or more related source domains and learns a classifier for unseen or unlabeled data in a target domain.

Currently, most existing domain adaptation methods only focus on single-modality knowledge transfer, that is, only one type of data (images or text) is considered in the training stage. However, different modalities of data can provide abundant and complementary content knowledge for domain adaptation. Therefore, we tackle a new task, *i.e.*, multi-modality domain adaptation. Compared with the conventional single-modality domain adaptation problem, there remain two challenges in multi-modality domain adaptation:

- ***Heterogeneity Gap*** mainly refers to the semantic difference between data in different modalities. The existence of semantic difference is due to many aspects, such as heterogeneity feature space of each modality, data format, data processing, data content and so on. These factors decide that it is quite difficult to directly measure the similarity and semantic association between multi-modality data.
- ***Domain Gap***, also known as "domain shift", refers to the difference in data distributions between two domains, which is common in real-life applications. For images, domain gap comes from consequences of changing conditions, *i.e.*, background, location, pose changes, and the domain gap might be larger, if the source and target domains contain images of different types, such as photos, NIR images, paintings or sketches.

We take video analysis as an example task to further explain "Heterogeneity Gap" and "Domain Gap". The audio sequence can be deemed as a one-dimensional signal and the video sequence can be treated as a three-dimensional signal which contains both spatial and temporal information. As for "Heterogeneity Gap", the difficulties come from two aspects. On the one hand, we need to adopt different methods to preprocess and analyze audio and video signals. On the other hand, it is quite challenging to acquire and associate semantic information of different signals, and then generate semantic features. In addition, videos come from multiple domains, such as different social medias (Flickr, Google News, YouTube, and Twitter), which thus arises the "Domain Gap" problem. That is, even if data from different domains are annotated with the

Manuscript received *** **, 2018; revised *** **, 2018; accepted *** **, 2019. This work is supported in part by the National Natural Science Foundation of China under Grants 61432019, 61572498, 61532009, 61728210, 61721004, 61751211, 61772244, 61472379, 61720106006 and U1705262, and the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJ-SSW-JSC039, the Beijing Natural Science Foundation 4172062, and Youth Innovation Promotion Association CAS 2018166.

The authors are with the National laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; the University of Chinese Academy of Sciences, Beijing 100049, China and also with Peng Cheng Laboratory, Shenzhen, China. (e-mail: xinhong.ma@nlpr.ia.ac.cn; tzzhang10@gmail.com; csxu@nlpr.ia.ac.cn).

same category labels, video frames are still quite different in the aspect of content, style, and so on. Therefore, models trained on labeled source domain cannot generalize well on target domain.

For solving the above challenges respectively, several works have been done in multi-modality representation learning and single-modality domain adaptation. To deal with "Heterogeneity Gap", remarkable developments have achieved in multi-modality representation learning and many other computer vision tasks, for instance, multi-modality retrieval [6, 7], visual question answering (VQA) [8, 9], image and video caption [1, 10] and so on. Among them, similar pipeline techniques are used. First, individual representations for each modality are extracted from the independent feature extractors. And then, a common feature space between different modalities is learned for cross-modality similarity measure. Features in the common space are usually generated following principles such as maximizing cross-model pairwise correlation [11], maximizing classification accuracy in the common space [12], etc. These common feature space methods can be further divided into two major subcategories: one is linear projections methods, the other is the DNN-based methods. Representative works in Linear projections methods are Canonical Correlation Analysis (CCA) [11] and its variants, which focus on learning linear projections to embed multi-modality data into the common feature space. However, linear projections cannot capture the complex cross-modality correlation with high non-linearity. With the rapid development of deep learning, the DNN-based methods [13–16] have currently become an active research line, which takes DNNs as basic models to extract features for each modality and then merge multi-modality information to learn the fused representations. However, these methods ignore the semantic correlation between multi-modality data. For example, given a pair of multi-modality data, *i.e.*, an image and the corresponding description, the entities (subjects, objects and verbs in a sentence) of description are semantically related to a set of regions in the image. However, not all image regions can reflect the semantic meanings of the description. Motivated by this observation, a stacked attention mechanism is applied to model this semantic correlation of multi-modality data and learn the fused multi-modality representation.

To reduce "Domain Gap", a variety of single-modality domain adaptation approaches have been proposed [17–19]. Among existing methods, Maximum Mean Discrepancy (MMD) [20], which regards mean difference between two distributions as domain discrepancy, is one of the most widely used strategies to measure distribution difference between source and target domains [21, 22]. Later on, numerous domain adaptation approaches have been proposed by designing a revised class-wise MMD, such as, class-wise MMD [21, 23], multi-kernel MMD [24, 25]. Recently, numerous deep adversarial adaptation methods [26–29] have been proposed, which is analogous to generative adversarial networks [30]. A domain classifier is trained to tell whether the sample comes from source domain or target domain. The feature extractor is trained to minimize the classification loss and maximize the domain confusion loss. Domain-invariant yet discriminative features are seemingly obtainable through the principled lens of adversarial training. However, previous methods cannot consider fine-grained categories information, which means they cannot constrain "relative distance" between samples of different categories and capture subtle discrimination between samples to learn more discriminative representations. Therefore, even with strong transfer ability, the generalization performance on target domain task is not very well. Motivated by the above observation, we construct multi-channel constraints to capture fine-grained categories information in feature space, and transfer the knowledge to target data, which can further improve the performance on the target task.

Considering the above factors, we propose a Multi-Modality Adversarial Network (MMAN) to address unsupervised domain adaptation. As shown in Figure 1, the proposed MMAN consists of multi-modality feature extractor, label predictor, domain classifier, and multi-channel constraint implemented by triplet loss. During the training stage, the unlabeled target data and triplets generated from source data are extracted features via multi-modality feature extractor. Then, all obtained source and target representations are regrouped for different purposes. Specifically, source representations are input into triplet loss layer and label predictor so that fine-grained categories information is explored to benefit classification task. Both source and target representations are fed into domain classifier to reduce domain gap via adversarial training. After training, the target data can loop through the multi-modality feature extractor and label predictor to get target task label.

Our contributions in this work are threefold:

- We propose an end-to-end general framework denoted as the Multi-Modality Adversarial Network (MMAN) for both unsupervised multi-modality domain adaptation and single-modality domain adaptation.
- The MMAN can leverage the stacked attention mechanism and a multi-channel constraint in the adversarial domain adaptation framework to explore the fine-grained categories information and generate semantic multi-modality representations.
- The proposed MMAN algorithm performs favorably against the state-of-the-art methods on three domain adaptation benchmark datasets: *Office-31*, *Caltech-Office*, *ImageCLEF-DA*. Furthermore, experiments on a multi-domain and multi-modality social event dataset collected by ourselves, can further demonstrate the effectiveness of our model.

The remainder of this paper is organized as follows. In Section II, we review the related work. Section III introduces the formulation of our network. In Section IV, the experimental results and analysis are given on *Office-31*, *Caltech-Office*, *ImageDLEF-DA* and a social event dataset collected by ourself. Our conclusion and future work are presented in Section V.

II. Related work

In this section, we briefly review some methods which are most related to our work including single-modality domain adaptation and multi-modality representation learning.

A. Single-modality Domain Adaptation

According to literature survey [17], existing single-modality domain adaptation methods can be roughly organized into two categories: shallow domain adaptation methods and deep domain adaptation methods.

Shallow domain adaptation can be further divided into four subcategories: instance reweighting [31–33], feature augmentation [34–36], feature space alignment [35–37] and feature transformation methods [33, 38]. The main idea of instance reweighing methods is to weight each instance by the ratio of likelihoods, *i.e.*, independent posterior probabilities achieved by a domain classifier [39] or Kullback-Leibler divergence between densities [31, 40]. Unlike instance reweighting methods, feature augmentation embed raw features into d-dimensional linear subspaces via theories of Geodesic Flow Sampling (GFS) [34, 36] and Geodesic Flow Kernel (GFK) [35, 36] that treat data as points on the Grassman manifold so that local geometry structure can be explored for two domains. Different from augmenting features, feature space alignment methods attempt to align the source features with the target ones. For example, Subspace Alignment (SA) [41] learns an alignment by minimizing Bregman divergence between the subspaces. The linear Correlation Alignment (CORAL) [37] reduces domain shift with second-order statistics of the source and target distributions. Finally, feature transformation methods, *e.g.*, Transfer Component Analysis (TCA) [33], are proposed to find a projection into a latent space where discrepancy between the source and target distributions are decreased.

Recently, with the rapid development of deep learning, deep domain adaptation (deepDA) methods have been a hot research topic in domain adaptation (DA) community. The first deepDA method is the Stacked Denoising Autoencoders (SDA) [42], which relys on denosing autoencoders to learn common features. Later on, some methods [24, 25, 43] are inspired by the siamese deep architecture where two branches of networks are used to model domain distributions respectively. Usually, the general discrepancy constrains (in general MMD) are defined between activation layers in the two branches. Recently, numerous adversarial adaptation methods have been proposed [26, 29, 44–47], which is analogous to generative adversarial networks [30]. Thereinto, the Domain Adversarial Neural Networks (DANN) [26] creatively integrates a gradient reversal layer into the deep network, which can insure the learned features domain-invariant and discriminative for the main learning task. To improve the DANN architecture, we propose the MMAN model that captures source domain fine-grained categories information by means of multi-channel constraints. And then the gradient reversal layer is adopted to transfer the knowledge to the target task for the purpose of enhancing the discriminative of target features.

B. Multi-modality Representation Learning

In multimedia, multi-modality representation learning is the basic technique in many applications, for example, multi-modality retrieval [6, 7], visual question answering (VQA) [8, 9], image and video caption [1, 10] and so on. Among them, Canonical Correlation Analysis (CCA) is one of the most important statistical methods to investigate relationships among multi-modality data [48–50]. Moreover, [50] combines CCA and multi-class logistic regression to learn semantic representation for each image. Later on, there are some deep learning methods proposed for multi-modality representation learning [51–53]. A quite popular framework consists of several neural layers and a common hidden layer. Each modality data is firstly fed into the individual layers and then projects into a joint feature space by the common hidden layer [54]. The joint representation will be further processed for prediction. In addition, some methods focus on learning multi-modality representation in an unsupervised way. [42] introduces and motivates a new training principle for unsupervised common feature representations learning from both source and target domains so that the learned representations are robust to input patterns of partial damage. More recently, multi-modality representations fusion is extensive studied in addressing visual question answering task [8, 55]. Thereinto, bilinear models are popularly used because they encode full second-order interactions. However, above methods only care about the correlation between two modalities. The semantic information is ignored. Inspired by the above observations, stacked attention feature extractor is applied in MMAN model to learn transferable features encoded in semantic content information of multi-modality data.

III. METHODOLOGY

A. Formulation and Notation

We follow definitions and notations in [18, 19]. A domain D is composed of a d-dimensional feature space $\mathcal{X}$ with distribution $P(\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n\} \in \mathcal{X}$. Given a domain D, a task T is to predict class label by learning a predictive function $f(\cdot)$ from feature vectors and label pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$ in source data set, *i.e.*, $D_S = \{(\mathbf{x}_{S1}, \mathbf{y}_{S1}), \cdots, (\mathbf{x}_{Sn}, \mathbf{y}_{Sn})\}$. Specifically, $\mathbf{x}_{Si} \in \mathcal{X}_S$ is the i-th data instance of D_S and $\mathbf{y}_{Si} \in \mathcal{Y}_S$ is the corresponding class label. Analogously, D_T is defined as the target domain data set, *i.e.*, $D_T = \{(\mathbf{x}_{T1}), \cdots, (\mathbf{x}_{Tn})\}$ where $\mathbf{x}_{Ti} \in \mathcal{X}_T$ is the i-th data instance of D_T and $\mathbf{y}_{Ti} \in \mathcal{Y}_T$ is the corresponding class label. Further, T_S and T_T are respectively represent source task, target task. $f_S(\cdot)$ and $f_T(\cdot)$ respectively stand for source and the target predictive function.

Based on the above definition, domain adaptation is to explore information from D_S and D_T ($D_S \neq D_T$) so that the target predictive function $f_T(\cdot)$ could be improved. Both source and target distributions are assumed unknown, similar but different. That is to say, the target distributions $P(\mathbf{X}_T)$ can be "shifted" from $P(\mathbf{X}_S)$ by some *domain shift*. The training set is made up of samples from the two domains. Remarkably, we can only get access to the class labels of source instances while target labels are unknown during training, but we want to predict such unknown labels of target instances during testing. What's more, we denote binary value (domain label) d_i for the i-th instance. For example, if x_i from the source distribution ($\mathbf{x}_i \sim P(\mathbf{X}_S)$), $d_i = 0$. Otherwise, $d_i = 1$ if $\mathbf{x}_i \sim P(\mathbf{X}_T)$.

B. Multi-Modality Adversarial Network

Figure 1 shows an overview of our model that shares weights among three pathways and predicts category label

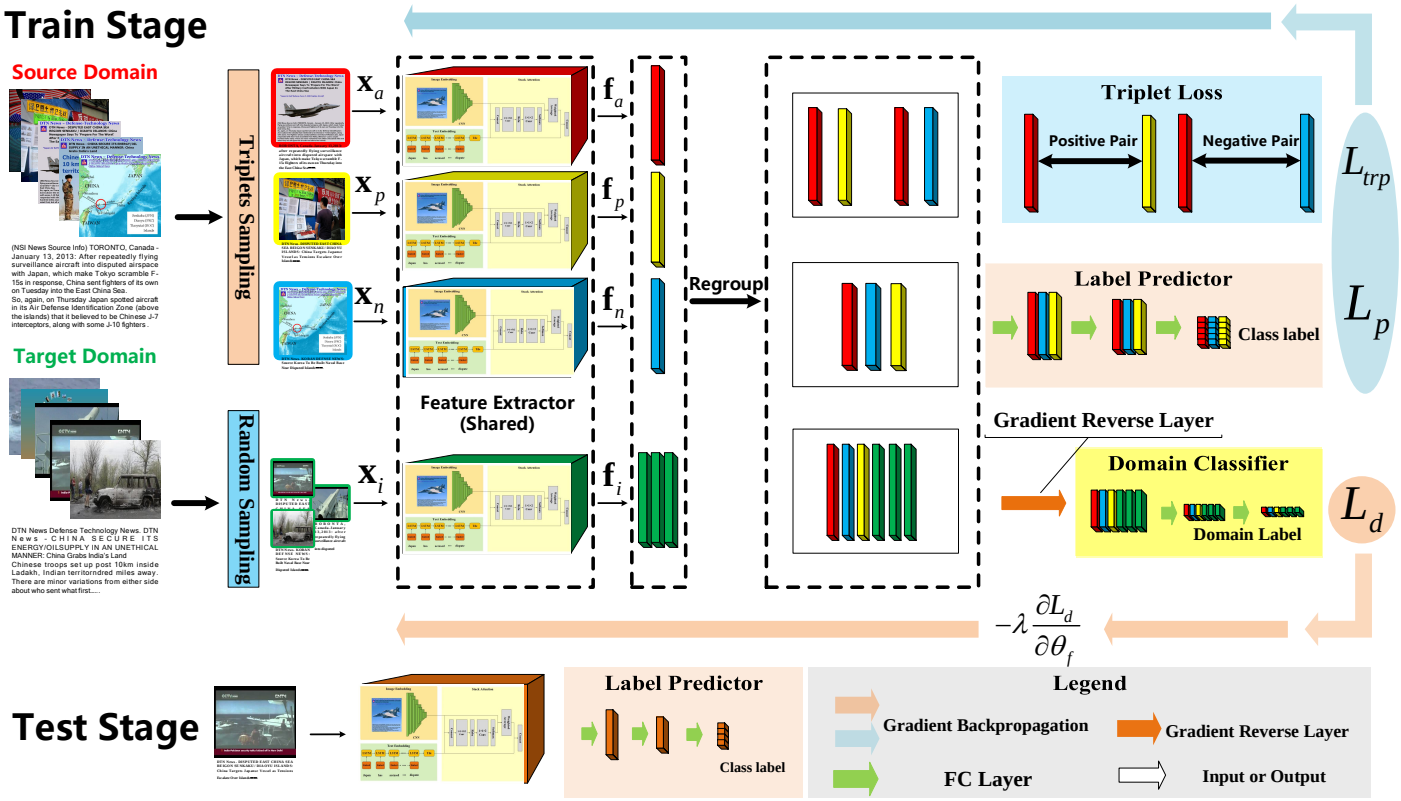


Fig. 1: The architecture of Multi-Modality Adversarial Network (MMAN). During forward propagation, triplets sampling strategy act on source domain data to generate triplets $\mathbf{x}_a, \mathbf{x}_p, \mathbf{x}_n$. In the meanwhile, the equivalent number of instances $\mathbf{x}_i$ are sampled randomly from the target domain. All sampled data are fed into the weight-shared feature extractor to extract multi-modality representations $\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n, \mathbf{f}_i$, which would be regrouped for different purposes. Specifically, source features $\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n$ is not only fed into label predictor for inferring class labels but also constrained by triplet loss to capture fine-grained categories information. In addition, both source and target features $\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n, \mathbf{f}_i$ are used to reduce domain discrepancy via adversarial training implemented by the gradient reverse layer and domain classifier. When back-propagation, the gradients of label predictor loss L_p and triplet loss L_{trp} can directly be passed to the proceeding layers, while gradient of domain classifier loss L_d would be multiplied by $-\lambda$. Once the model is trained, the target data can be directly fed into the feature extractor and label predictor to get label prediction.

$\mathbf{y} \in \mathcal{Y}$ and domain label $d \in \{0,1\}$ for each input $\mathbf{x}$. We decompose such mapping into four parts, namely, feature extractor, label predictor, domain classifier and triplet loss layer. We assume that the input $\mathbf{x}$ includes images and the corresponding text descriptions. During training, we firstly use triplet sampling mechanism to sample triplets in the source domain while random sampling is used in the target domain. Then, the data is fed into four weight-shared stacked attention based feature extractor to extract multi-modality representations. For the purpose of obtaining domain-invariant features, both source and target representations are transported into the gradient reversal layer and domain classifier while only annotated features from source domain input into triplet loss layer and label predictor. Once the model is trained, the target data can be directly fed into the feature extractor and label predictor to get label prediction. The details of the proposed MMAN are given as follows.

Stacked Attention Feature Extractor. In order to capture the semantic correlations between multi-modality data, *i.e.*, images and texts, we apply the stacked attention feature extractor to learn multi-modality representations. The stacked attention is quite effective in visual question answering [56, 57] but has not been applied for tackling multi-modality domain adaptation problem. The architecture of feature extractor is shown in Figure 2, which contains three major components: (1) Image Model, which uses a CNN to extract high level image representations, *e.g.*, one vector for each region of the image; (2) Text Model, which uses a LSTM to extract a semantic vector of the text descriptions and (3)a Stacked Attention Model, which associates the image regions that are relevant to the descriptions for final classification task. The pipeline of generating semantic multi-modality representation can be found in the caption of Figure 2.

Image Model: The image model uses a pre-trained convolution neural network (CNN) model (ResNet [58] or AlexNet [4]) to extract representation $\tilde{\phi}$ of the input image I. Instead of using features from the last fully connected layer, we choose the features $\tilde{\phi}$ from the input of the last pooling layer, which retains spatial information of the original images.

$$\tilde{\phi} = CNN(I). \tag{1}$$

$\tilde{\phi}$ is a three-dimensional tensor with $h \times w \times m$ dimensions.

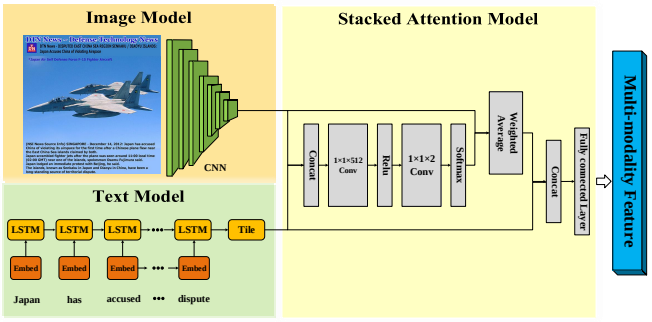


Fig. 2: The Multi-Modality Feature Extractor contains three major components: image model, text model and stacked attention model. The image model uses a CNN to extract high level image representations. The text model uses a LSTM to extract a semantic vector of the corresponding description of the image. The stacked attention model first concatenates text description vector with the image feature matrix over the depth dimension, which is passed through attention layers to compute attention distributions. The high-level attention layer gives as semantic attention distribution focusing on the regions that are more relevant to the text description. Finally, we combine the image features from the attention layer with the text description vector, and pass the features through a fully connected layer to output the final multi-modality representation.

$h \times w$ is the number of regions in the image and m is the dimension of the feature vector for each region. We furthermore perform l_2 normalization on the depth (last) dimension of image features to obtain the image feature ϕ. The feature vector ϕ_l (l-th feature vector of ϕ in depth dimension) corresponds to the image region indexed by l.

Text Model: We tokenize and encode a piece of given event description D_L into word embedding $E_q = \{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_K\}$, where $\mathbf{e}_i \in \mathbb{R}^D$, D_L is the length of the distributed word representation, and K represents words number in the event description. The embeddings are then fed into a long short-term memory (LSTM) [59].

$$\mathbf{s} = \text{LSTM}(E_q). \tag{2}$$

The event description features are the final state $\mathbf{s}$ of the LSTM. The length of the vector $\mathbf{s}$ is d.

Stacked Attention Model: When given the image feature matrix $\phi \in \mathbb{R}^{h \times w \times m}$ and the description feature vector $\mathbf{s} \in \mathbb{R}^d$, the stacked attention model generates semantic multi-modality representations via the multi-attention reasoning. In many cases, the semantic meanings in the text description are only related to some regions in an image. Therefore, using one global image feature vector to associate with text vector semantically could cause suboptimal results, due to the noises introduced from regions that are irrelevant to the potential semantic meanings of text description. Instead, reasoning via the attention layer, the stacked attention model is able to generate multi-modality representations encoded in semantic information between images and texts.

Given the image feature matrix ϕ and the text description vector $\mathbf{s}$, we concatenate text description vector with the image feature matrix over the depth dimension and pass through attention layers F_c to compute attention distributions, as shown in Eq. 3.

$$\alpha_{c,l} \propto \exp F_c(\mathbf{s}, \phi_l) \tag{3}$$

where $\alpha_{c,l} \in \mathbb{R}^m$ is an m dimensional vector, which corresponds to the attention probability of each image region. In addition, $F = [F_1, F_2, \cdots, F_c]$ is modeled with two convolution layers. The first layer is a 1×1 dimensional and depth 512 convolution layer followed by the ReLU nonlinearity. The output feature is passed through another depth C convolution layer followed by the softmax over spatial dimensions. In addition, the F_is share parameters in the first layer. The attention weights $\alpha_{c,l}$ are normalized separately for each glimpse $c = 1, 2, ..., C$, as shown in Eq. 4.

$$\sum_{l}^{L} \alpha_{c,l} = 1 \tag{4}$$

Based on the attention distribution, we can calculate the weighted sum of the image vectors as shown in Eq. 5.

$$\phi' = \sum_{l} \alpha_{c,l} \phi_l \tag{5}$$

Here, each image feature glimpse ϕ' is the weighted average of image features ϕ over all the spatial locations $l = 1, 2, ..., L$. After that, we combine the image glimpses with the text description vector that would be passed through a fully connected layer to obtain the final multi-modality representations $\mathbf{f}$ as in Eq. 6.

$$\mathbf{f} = fc(\phi' \oplus \mathbf{s}) \tag{6}$$

where $\oplus$ operation means directly concatenating text feature vector with image feature glimpses over the depth dimension. $fc(\cdot)$ represents a fully connected layer with ReLU nonlinearity. Finally, we obtain the D-dimension multi-modality feature vector $\mathbf{f} \in \mathbb{R}^D$ mapped by multi-modality feature extractor G_f, *i.e.*, $\mathbf{f} = G_f(\mathbf{x}; \theta_f)$, whose parameters are denoted as θ_f.

Adversarial Training. To learn task-specific feature, the representation $\mathbf{f}$ generated by multi-modality feature extractor is mapped to label $\mathbf{y}$ by a mapping function G_y (label predictor). In the meantime, the same feature $\mathbf{f}$ is transformed to domain label by G_d (*domain classifier*). θ_y and θ_d respectively represent the parameters of the above two mapping functions.

During the training stage, annotated source instances are used to minimize the label prediction loss so that the feature extractor and the label predictor are jointly optimized to insure discrimination of the features $\mathbf{f}$. Meanwhile, the features $\mathbf{f}$ should be domain invariant so that the label predictor trained on the annotated source instances can precisely make predictions on target instances. That is, the distribution $P(\mathbf{f}_S) = \{G_f(\mathbf{x}; \theta_f) | \mathbf{x} \in P(\mathbf{X}_S)\}$ and $P(\mathbf{f}_T) = \{G_f(\mathbf{x}; \theta_f) | \mathbf{x} \in P(\mathbf{X}_T)\}$ should be made as similar as possible. However, the source and target distributions and high-dimensional $\mathbf{f}$ are constantly changing as learning progresses. Therefore, it is significantly important to precisely measure the source and target distributions.

One possible method estimating the dissimilarity is to observe the loss values of domain classifier G_d. If the output

values of the domain classifiers are close to 0.5, the learned features of two domain are quite similar. In other words, the domain classifier cannot well distinguish source features with target features, which indicates that the learned features are domain-invariant. This observation leads to the following adversarial training paradigm. At training, we maximize the loss of the domain classifier to optimize the parameters θ_f of the feature mapping, which can make the two domain distributions similar. At the same time, optimizing the parameters θ_d of the domain classifier to keep domain-specific information by means of minimize the loss of the domain classifier. Such an idea can be accomplished by gradient reversal layer (GRL) proposed in [60].

Gradient Reversal Layer (GRL). In order to obtain the domain-invariant representation, we can insert the Gradient Reversal Layer (GRL) into the model to process adversarial training. The inserted position of GRL in the architecture is shown in Figure 1. There are no other parameters in the gradient reverse layer except the meta-parameter λ. During the forward process, the GRL does not do any operations on the passing features. When back-propagating, the gradient propagated to the GRL would be multiplied by $-\lambda$, *i.e.*, $\frac{\partial L_d}{\partial \theta_f}$ is effectively replaced with $-\lambda\frac{\partial L_d}{\partial \theta_f}$ during back-propagation. After that, the processed gradient is passed to the preceding layer.

Mathematically, the gradient reversal layer $R_\lambda(\mathbf{x})$ can be formulated as two equations describing above opreations:

$$R_\lambda(\mathbf{x}) = \mathbf{x}, \tag{7}$$

$$\frac{dR_\lambda}{d\mathbf{x}} = -\lambda\mathbf{I}, \tag{8}$$

where $\mathbf{I}$ is an identity matrix, and λ is not updated by back-propagation.

Multi-channel Constraints. For the purpose of capturing the fine-grained categories information to guide the model to learn discriminative features on target data, we apply multi-channel constraints to our model. Our multi-channel constraint is achieved by a triplet loss [61, 62] which is normally trained on a series of triplets $\{\mathbf{f}_a, \mathbf{f}_p, \mathbf{f}_n\}$ in source domain, where $\mathbf{f}_a$ and $\mathbf{f}_p$ are features extracted from samples with the same category, and $\mathbf{f}_n$ is from a different category. The triplet loss can be formulated as follows.

$$L_{trp}(\theta_f) = \sum_{a,p,n}^{M} \max\left(\|\mathbf{f}_a - \mathbf{f}_p\|_2^2 - \|\mathbf{f}_a - \mathbf{f}_n\|_2^2 + \alpha_{trp}, 0\right) \tag{9}$$

where $\mathbf{f}_a = G_f(\mathbf{x}_a;\theta_f)$, $\mathbf{f}_p = G_f(\mathbf{x}_p;\theta_f)$, $\mathbf{f}_n = G_f(\mathbf{x}_n;\theta_f)$. M is the number of sampled triplets. The Euclidean distance is adopted to measure the feature similarity in Eq. 9, The triplet loss can achieve the following goals: (1) The distance of samples with the same label can be very close in the feature space. (2) The distance of samples with different labels can be as large as possible. (3) Further, the triplet loss can ensure the distance between $\mathbf{f}_a$ and $\mathbf{f}_n$ is greater than the distance between $\mathbf{f}_a$ and $\mathbf{f}_p$. In addition, the margin α_{trp} constrains the minimal distances between positive pairs and negative pairs. That is, when minimizing Eq. 9, the learned representation can satisfy $\|\mathbf{f}_a - \mathbf{f}_p\|_2^2 \to 0$ and $\|\mathbf{f}_a - \mathbf{f}_n\|_2^2 > \|\mathbf{f}_a - \mathbf{f}_p\|_2^2 + \alpha_{trp}$.

Compared with the traditional classification loss, *e.g.*, the softmax, the triplet loss can learn the subtle discrimination between samples. It is to add the measurement of the inputs difference to learn more discriminative representations. At the same time, the GRL and domain classifier can transfer the fine-grained categories information to the target task to improve performance.

In addition, The accuracy and convergence speed of the triplet loss approach heavily depend on the triplet sampling method as discussed in [63]. For simplicity, we use the off-line random sampling mechanism to generate triplets, which can achieve impressive performance and stable convergence as proved by our experiments in IV-A5.

Objective Function. More formally, we consider the following functions:

$$\begin{aligned} L(\theta_f,\theta_y,\theta_d) &= \sum_{i=1\cdots N}\sum_{d=0} L_y\left(G_y\left(G_f\left(\mathbf{x}_i;\theta_f\right);\theta_y\right),y_i\right) \\ &+ \sum_{i=1\cdots N} L_d(G_d(G_f(\mathbf{x}_i;\theta_f);\theta_d),d_i) + \gamma L_{trp}(\theta_f) \\ &= \sum_{i=1\cdots N}\sum_{d=0} L_y^i(\theta_f,\theta_y) + \sum_{i=1\cdots N} L_d^i(\theta_f,\theta_d) + \gamma L_{trp}(\theta_f) \end{aligned} \tag{10}$$

Here, $L_y(\cdot,\cdot)$ is the loss for label prediction, $L_d(\cdot,\cdot)$ is the loss for the domain classification, while L_y^i and L_d^i denote the corresponding loss function evaluated at the i-th training instance. G_f,G_y,G_d respectively represent feature extractor, label predictor and domain classifier with parameters $\theta_f,\theta_y,\theta_d$. N is batch size. d is the domain label of samples. λ and γ are hyper-parameters. Based on the above idea, we are seeking the parameters $\hat{\theta}_f$, $\hat{\theta}_y$, $\hat{\theta}_d$ by optimizing the Eq. 11-12:

$$(\hat{\theta}_f,\hat{\theta}_y) = \arg\min_{\theta_f,\theta_y} L(\theta_f,\theta_y,\hat{\theta}_d) \tag{11}$$

$$\hat{\theta}_d = \arg\max_{\theta_d} L(\hat{\theta}_f,\hat{\theta}_y,\theta_d) \tag{12}$$

However, the standard stochastic gradient solvers (SGD) cannot be directly adapted to search the saddle point in Eq. 11-12. Thanks to gradient reversal layer, it is no longer a problem. According to Eq. 7, we can define the final objective function of $\tilde{L}(\theta_f,\theta_y,\theta_d)$ that is being optimized by SGD:

$$\begin{aligned} \tilde{L}(\theta_f,\theta_y,\theta_d) &= \sum_{i=1\cdots N}\sum_{d=0} L_y\left(G_y\left(G_f\left(\mathbf{x}_i;\theta_f\right);\theta_y\right),y_i\right) \\ &+ \sum_{i=1\cdots N} L_d(G_d(R_\lambda(G_f(\mathbf{x}_i;\theta_f));\theta_d),d_i) + \gamma L_{trp}(\theta_f) \end{aligned} \tag{13}$$

IV. Experiments

This section reports the experimental validation of our method. To demonstrate the versatility of our approach, we perform experiments on two different tasks: object recognition and social event recognition. The object recognition task is for single modality unsupervised domain adaptation while the social event recognition is for multi-modality domain adaptation. Experimental results demonstrate the effectiveness of our model for unsupervised domain adaptation under single modality and multi-modality settings.

A. Object recognition for Single-modality Domain Adaptation

1) ***Datasets****:* Three benchmark datasets, namely *Office-31*, *Office-Caltech* and *ImageCLEF-DA*, are used to evaluate the object recognition task for unsupervised domain adaptation.

Office-31 includes 4,110 images in 31 categories collected from three different domains: Amazon (A), Webcam (W) and DSLR (D). To achieve an unbiased evaluation, we compare all methods on all six transfer tasks A→W, D→W, W→D, A→D, D→A and W→A.

Office-Caltech is built by selecting 10 common categories shared by Office-31 and Caltech-256 (C). We can create 12 transfer tasks: A→W, D→W, W→D, A→D, D→A, W→A, A→C, W→C, D→C, C→A, C→W, and C→D.

ImageCLEF-DA is a benchmark dataset for *ImageCLEF 2014*[1] domain adaptation challenge, organized by selecting 12 common shared categories for the following three Public datasets: *Caltech-256* (**C**), *ImageNet ILSVRC 2012* (**I**), and *Pascal VOC 2012* (**P**). Here, each dataset is treated as a domain. There are 50 images in each category and total 600 images in each domain. All domain combinations are used to create six transfer tasks: I→P, P→I, I→C, C→I, C→P and P→C.

There are more categories in *Office-31* than those in *Office-Caltech*, which makes *Office-31* more difficult for domain adaptation. However, *Office-Caltech* provides more transfer tasks to enable an unbiased observation. Besides, the three domains in *ImageCLEF-DA* are of equal size, which makes it a good complement to *Office-31* for more controllable experiments. The results on these three benchmark datasets can illustrate the effectiveness of our proposed framework for unsupervised domain adaptation.

2) ***Baseline****:* We compare our model with both conventional and deep domain adaptation methods including Transfer Component Analysis (**TCA**) [33], Geodesic Flow Kernel (**GFK**) [35], Deep Convolutional Neural Network (**DCNN** including **AlexNet** [4] and **ResNet** [58]), Deep Domain Confusion (**DDC**) [44], Deep Adaptation Network (**DAN**) [25], Reverse Gradient (**RevGrad**) [60], Residual Transfer Networks (**RTN**) [64], Joint Adaptation Networks (**JAN**) [43], Duplex Generative Adversarial Network (**DuGAN**) [65], Wasserstein Distance Guided Representation Learning (**WDGRL**) [66], Graph Adaptive Knowledge Transfer (**GAKT**) [67], Weighted Maximum Mean Discrepancy (**WDAN**) [23], Joint Geometrical and Statistical Alignment (**JGSA**) [68] and Probabilistic Unsupervised Domain Adaptation (**PUnDA**) [69]. Some details are as follows:

- **TCA:** It is a traditional shallow transfer learning method which aims to apply MMD-regularized kernel PCA to embed features into a high-dimensional space to preserve the shared attributes between two domains.
- **GFK:** Inspired by manifold learning, GFK focuses on using infinite number of intermediate subspaces to preserve the shared local structure between two domains.
- **DCNN:** The classic deep network architectures (AlexNet or ResNet) are used to extract features on source data to train a classifier that will be directly applied to predict labels on the target data.
- **DDC:** In DDC, a linear-kernel MMD is firstly inserted into deep networks to maximize domain invariance.
- **DAN:** Representations of all task-specific layers in DAN network, are embedded into a Reproducing Kernel Hilbert Space (RKHS). Then, an optimal multi-kernel selection method is used to match mean embedding of two domain distributions in the Reproducing Kernel Hilbert Space.
- **RevGrad:** The RevGrad is designed to obtain deep features that are discriminative for the main learning task and invariant with respect to the shift between domains. It shows that the adaptation behavior can be achieved in almost any feed-forward model by inserting few standard layers and a gradient reversal layer via adversarial training paradigm.
- **RTN:** The basic idea is based on the assumption that the source classifier and target classifier differ by a residual function. The Adaptation between classifiers can be achieved by explicitly learning the residual function with inserting multiple layers into the deep network. And not only that, the features of multiple layers are fused with tensor product and embedded into a Reproducing Kernel Hilbert Spaces (RKHS) to match distributions for feature adaptation.
- **JAN:** The method learns the transformation network by coordinating multiple domain-specific layers across multiple domains based on the JMMD (Joint Maximum Mean Difference) criteria. The adversarial training strategies are employed to maximize the performance of JMMD and make it easier to distinguish the distribution of source and target domains.
- **DuGAN:** Following the similar idea of GAN, this work proposes a novel GAN architecture with duplex adversarial discriminators. The generator is pitted against duplex discriminators to ensure the reality of domain transformation so that the latent representation domain invariant and the categories information can be preserved.
- **WDGRL:** It utilizes a neural network, denoted by the domain critic, to estimate empirical Wasserstein distance between the source and target samples and optimizes the feature extractor network to minimize the estimated Wasserstein distance in an adversarial manner. In this way, the model can guarantee the learned representations should be discriminative in prediction.
- **GAKT:** It models to jointly optimize target labels and domain-free features in a unified framework. Specifically, the semi-supervised knowledge adaptation and label propagation on target data are coupled to benefit each other.
- **WDAN:** It shows that MMD cannot account for class weight bias and result in degraded domain adaptation performance. Specifically, they introduce class-specific auxiliary weights into the original MMD to exploit the class prior probability on source and target domains.
- **JGSA:** They propose a unified framework that reduces the shift between domains both statistically and geometrically. Specifically, it learns two coupled projections that

[1]http://www.imageclef.org/2014/adaptation

TABLE I: Classification accuracy (%) on OFFICE-31 for unsupervised domain adaptation. From left to right: transfer methods, references, six transfer tasks and average classification accuracy. Methods in top half of the table use AlexNet as feature extractor while the remain methods employ ResNet. Notation: Amazon (A), Webcam (W) and Dslr (D). A→W represents the transfer task from Amazon to Webcam.

Transfer Methods	Reference	A→W	D→W	W→D	A→D	D→A	W→A	Avg
TCA	IEEE Trans. Neural Networks 2011	61.0±0.0	93.2±0.0	95.2±0.0	60.8±0.0	51.6±0.0	50.9±0.0	68.8
GFK	CVPR2012	60.4±0.0	95.6±0.0	95.0±0.0	60.6±0.0	52.4±0.0	48.1±0.0	68.7
AlexNet	NIPS2012	61.6±0.5	95.4±0.3	99.0±0.2	63.8±0.5	51.1±0.6	49.8±0.0	70.1
DDC	ICCV2015	61.8±0.4	95.0±0.5	98.5±0.4	64.4±0.3	52.1±0.8	52.2±0.4	70.6
RevGrad	ICML2015	73.0±0.5	96.4±0.3	99.2±0.3	56.6±0.1	53.6±0.1	48.7±0.5	71.3
DAN	ICML2015	68.5±0.5	96.0±0.3	99.0±0.3	67.0±0.4	54.0±0.5	53.1±0.5	72.9
RTN	NIPS2016	73.3±0.3	96.8±0.2	99.6±0.1	71.0±0.2	50.5±0.3	51.0±0.1	73.7
JAN	ICML2017	74.9±0.3	96.6±0.2	99.5±0.2	71.8±0.3	58.3±0.3	55.0±0.4	76.0
DuGAN	CVPR2018	73.2±0.2	-	-	74.1±0.6	**61.5±0.5**	**59.1±0.5**	-
MMAN	Ours	**78.5±0.5**	**97.0±0.2**	**99.6±0.1**	**75.3±0.3**	54.2±0.3	54.5±0.4	**76.5**
ResNet	CVPR2016	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
TCA	IEEE Trans. Neural Networks 2011	72.7±0.0	96.7±0.0	99.6±0.0	74.1±0.0	61.7±0.0	60.9±0.0	77.6
GFK	CVPR2012	72.8±0.0	95.0±0.0	98.2±0.0	74.5±0.0	63.4±0.0	61.0±0.0	77.5
DDC	ICCV2015	75.6±0.2	96.0±0.2	98.2±0.1	76.5±0.3	62.2±0.4	61.5±0.5	78.3
DAN	ICML2015	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
RTN	NIPS2016	84.5±0.2	96.8±0.1	99.4±0.1	77.5±0.3	66.2±0.2	64.8±0.3	81.6
RevGrad	ICML2015	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
JAN	ICML2017	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
MMAN	Ours	**85.8**±0.3	**97.4±0.4**	**100.0±0.0**	**85.8±0.3**	**70.3±0.2**	**71.2±0.1**	**85.1**

TABLE II: Classification accuracy (%) on ImageCLEF-DA for unsupervised domain adaptation. From left to right: transfer methods, references, six transfer tasks and average classification accuracy. All methods use AlexNet as feature extractor. Notation: Caltech-256 (C), ImageNet ILSVRC 2012 (I) and Pascal VOC 2012 (P). C→P represents the transfer task from Caltech to Pascal VOC.

Transfer Methods	Reference	I→P	P→I	I→C	C→I	C→P	P→C	Avg
AlexNet	NIPS2012	66.2±0.2	70.0±0.2	84.3±0.2	71.3±0.4	59.3±0.5	84.5±0.3	73.9
DAN	ICML2015	67.3±0.2	80.5±0.3	87.7±0.3	76.0±0.3	61.6±0.3	88.4±0.2	76.9
RTN	NIPS2016	67.4±0.3	81.3±0.3	89.5±0.4	78.0±0.2	62.0±0.2	89.1±0.1	77.9
JAN	ICML2017	67.2±0.5	**82.8±0.4**	**91.3±0.5**	80.0±0.5	63.5±0.4	91.0±0.4	79.3
MMAN	Ours	**68.7±0.2**	82.2±0.2	90.4±0.2	**80.6±0.1**	**64.7±0.1**	**91.5±0.2**	**79.7**

TABLE III: Classification accuracy (%) on OFFICE-10+Caltech-10 for unsupervised domain adaptation. From left to right: transfer methods, references, twelve transfer tasks and average classification accuracy. Methods in top half of the table use AlexNet as feature extractor while the remain methods employ VGG. Notation: Amazon (A), Webcam (W), Dslr (D) and Caltech-256 (C). A→W represents the transfer task from Amazon to Webcam.

Transfer Methods	Reference	A→W	D→W	W→D	A→D	D→A	W→A	A→C	W→C	D→C	C→A	C→W	C→D	Avg
GFK	CVPR2012	89.5	97.0	98.1	86.0	89.8	88.5	76.2	77.1	77.9	90.7	78.0	77.1	85.5
AlexNet	NIPS2012	79.5	97.7	**100.0**	87.4	87.1	83.8	83.0	73.0	79.0	91.9	83.7	87.1	86.1
DDC	ICCV2015	83.1	98.1	**100.0**	88.4	89.0	84.9	83.5	73.4	79.2	91.9	85.4	88.8	87.1
RevGrad	ICML2015	90.8	98.3	98.7	89.2	90.6	93.8	85.7	86.9	83.7	92.8	88.1	87.9	88.9
DAN	ICML2015	91.8	98.5	**100.0**	91.7	90.0	92.1	84.1	81.2	80.3	92.0	90.6	89.3	90.1
RTN	NIPS2016	95.2	99.2	**100.0**	95.5	93.8	92.5	88.1	86.6	84.6	93.7	94.2	93.4	93.4
WDGRL	AAAI2018	89.47	97.89	**100.0**	93.68	91.69	93.67	86.99	**89.43**	**90.24**	93.54	91.58	94.74	92.74
MMAN	Ours	**96.6**	99.3	**100.0**	**97.5**	**94.3**	**94.2**	**88.7**	87.0	87.9	93.7	**98.3**	**98.1**	**94.6**
WDAN	CVPR2017	92.26	99.28	**100.00**	92.87	91.87	92.87	86.93	84.12	83.92	93.11	93.67	93.48	92.03
GAKT	ECCV2018	90.18	**100.00**	**100.00**	95.48	93.98	93.84	88.46	88.84	86.82	**95.12**	95.36	96.42	93.71
JGSA	CVPR2017	84.75	98.64	**100.00**	85.35	92.28	91.44	85.04	84.68	85.75	91.75	85.08	92.36	89.76
PUnDA	ICCV2017	82.86	98.24	99.16	85.86	89.24	89.06	86.64	83.28	83.48	93.12	86.76	90.98	89.06

project the source domain and target domain data into low dimensional subspaces where the geometrical shift and distribution shift are reduced simultaneously.

- **PUnDA:** It simultaneously minimizes the domain disparity while maximizing the discriminative power of classifiers. In addition, a novel regularized Variational Bayes (VB) algorithm is also developed for efficient estimation of the model parameters.

*3) **Setting up**:* Training deep networks from scratch on small datasets results in poor performance. Therefore, an effective technique used in practice is to fine-tune networks trained on a related task with large labeled data. Following other domain adaptation work [43, 60], we implement all deep methods based on the **Caffe** framework and fine-tune from Caffe-provided models of AlexNet and ResNet.

For MMAN model, we use standard back-propagation to fine-tune the **Feature Extractor** (only CNN model is used in image transfer tasks), and train bottleneck **Label Predictor** and **Domain Classifier**. Since these bottleneck layers are trained from scratch, we set their learning rate to be 10 times that of the other layers. We use mini-batch stochastic gradient decent (SGD) with momentum of 0.9 and the learning rate

is 0.001, which is adjusted during SGD using the following formula: $\eta = \frac{\eta_0}{(1+\alpha p)^\beta}$, where p is the training process linearly changing from 0 to 1, $\eta_0 = 0.01$, $\alpha = 10$ and $\beta = 0.75$. In addition, we set $\lambda = 0.1$, $\gamma = 0.01$.

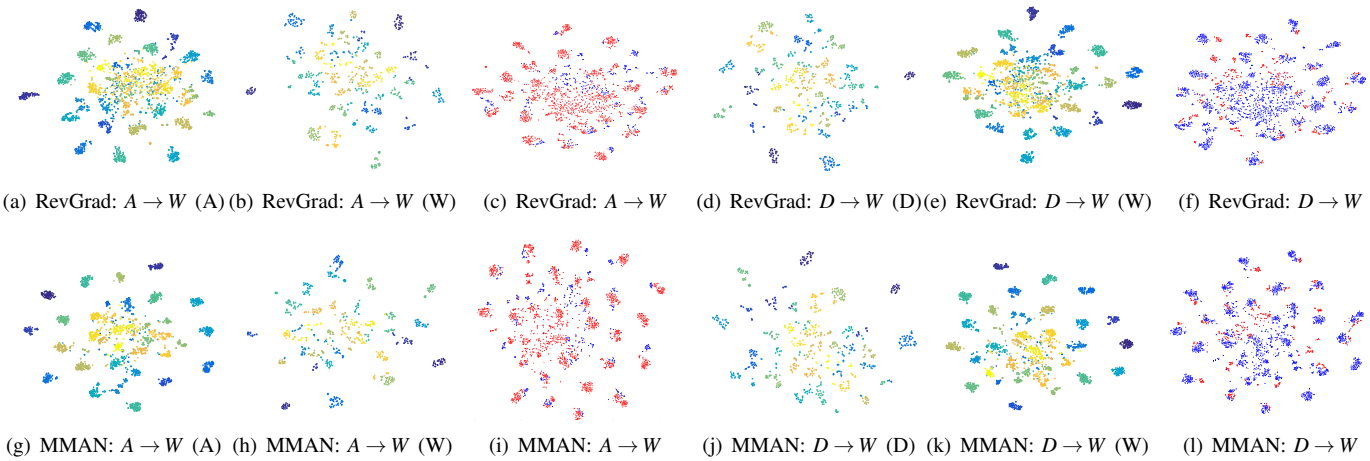
*4) **Results**:* In our experiments, we follow the standard unsupervised protocol using the entire labeled data in the source domain and unlabeled data in the target domain. The classification accuracy results on the six transfer tasks of *office-31* are shown in TABLE I. TABLE II and TABLE III illustrate the results on *ImageCLEF-DA* and *Office-Caltech*, respectively. As fair comparison with identical evaluation setting, all baseline results are directly cited from their published papers. From the results, we can observe that the proposed MMAN model outperforms these representative unsupervised domain adaptation methods on most transfer tasks, especially in some hard transfer tasks like $\mathbf{A} \rightarrow \mathbf{W}$ and $\mathbf{C} \rightarrow \mathbf{W}$. The performance of these two hard tasks can reflect the transfer robustness of algorithms, because the source and target domain are quite different in the two tasks, which makes it a challenge for safe transfer. Not only that, with regards to some easy tasks, *e.g.*, $\mathbf{D} \rightarrow \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{D}$ where domains are similar, the classification accuracy almost achieve 100%. These results can demonstrate that MMAN can learn adaptive classifiers and domain-invariant features for unsupervised domain adaptation.

Office-31: Based on the results of classification accuracy in TABLE I, we can make the following observations. (1) some shallow transfer learning methods perform better with more transferable deep features extracted by ResNet. This confirms that deep networks can learn abstract feature representations, which is able to reduce, but not remove the domain discrepancy [70]. (2) Deep transfer learning methods outperform both **DCNN** methods (AlexNet and ResNet) and traditional shallow transfer learning methods. This validates that domain discrepancy can be further reduced by inserting domain adaptation constraints into deep networks (DDC, DAN, RevGrad, RTN, and JAN). (3) The MMAN outperforms most baseline methods by large margin including the state-of-the-art model JAN. The reason is that the triplet loss is a depiction of the relative distance of data samples in data space, which reflects more fine-grained categories information of data samples. In addition, GRL and domain classifier will transfer the fine-grained categories information to the target task, which can improve the performance of the target domain task. Although the MMAN only models the marginal distributions based on independent feature layers (one layer for MMAN and RevGrad, and multi-layer for DAN and RTN), not the joint distribution, fine-grained feature marginal distribution can be learned by the triplet loss. (4) From the results of MMAN, AlexNet and RevGrad, we can also show the effect of different components of MMAN, namely, domain classifier, gradient reverse layer and multi-channel constraint. From AlexNet to RevGrad, the average performance is increased by 1.2%, which illustrates that domain classifier and gradient reversal layer can reduce domain discrepancy. From RevGrad to M-MAN, the multi-channel constraint is applied to capture fine-grained categories information so that the average accuracy is increased by 5.2%. That is, the learned representations of MMAN are more discriminative than those of RevGrad. (5) However, the MMAN does not exceed DuGAN on task $D \rightarrow A$ and $W \rightarrow A$. Two reasons, *i.e.*, size of train dataset and overfitting, can emphasize the experiment phenomenon. DuGAN is a generative adversarial network based methods that can generate images to offset the negative effect of training on small dataset. In addition, compared with target domain dataset, as the size of source domain dataset is quite small in task $D \rightarrow A$ and $W \rightarrow A$, the multi-channel constraint may cause overfitting on source data, which leads to low generalization on target task.

In addition, we can attain a more in-depth understanding of feature transferability. (1) In terms of accuracy on recognition experiments, ResNet-based methods outperform AlexNet-based methods by large margins. This validates that very deep convolutional networks, e.g. VggNet [71], GoogLeNet [72], and ResNet, are able to learn not only discriminative representations for main task but also more transferable or domain-invariant representations for domain adaptation. (2) The M-MAN model significantly outperforms ResNet-based methods, illustrating that deep networks can only reduce, but not remove the domain discrepancy.

ImageCLEF-DA: With more balanced transfer tasks in *ImageCLEF-DA*, we can evaluate whether the domain size influences the transfer performance. The classification accuracy results based on AlexNet are shown in TABLE II. MMAN models outperform the baseline methods on most transfer tasks, but by less improvement. This means that domain sizes may cause shift. However, MMAN model can still deal with domain shift caused by domain sizes. In TABLE II, the performance of MMAN on task $P \rightarrow I$ and $I \rightarrow C$ is not better than RTN. The reasons are as following: In JAN, the representations from multiple fully connected layers are used to estimate the domain discrepancy while the MMAN only uses the feature from the last fully-connected layer. However, MMAN can still outperform all baselines on the most transfer tasks.

Office-Caltech: The dataset can provide more transfer tasks to test the performance of our MMAN model. From the results in TABLE III, we can observe that MMAN outperforms the baseline methods on most transfer tasks. Reasons are similar to the results on Office-31. However, some other interesting conclusions can be drawn. (1) It has been proved that very deep convolutional networks is able to not only learn better representations for general vision tasks, but also learn more transferable representations for domain adaptation. It is noticed that WDAN, GAKT, JGSA and PUnDA use the VGG as their base feature extractor, while the MMAN use the AlexNet for feature extraction. Surprisingly, the MMAN can still surpass other shallow and deep methods (WDAN), which can verify the great power of the multi-channel constraint. (2) Compared with GAKT which utilizes the semi-supervised knowledge of target domain, MMAN can suppress GAKT via only capturing the source domain categories information, which illustrates the effectiveness of the proposed multi-channel constraint. (3) In TABLE III, we can observe that MMAN underperforms WDGRL on task $W \rightarrow C$ and $D \rightarrow C$. Small labeled training dataset and the overfitting on source data can explain the



(a) RevGrad: $A \rightarrow W$ (A) (b) RevGrad: $A \rightarrow W$ (W) (c) RevGrad: $A \rightarrow W$ (d) RevGrad: $D \rightarrow W$ (D) (e) RevGrad: $D \rightarrow W$ (W) (f) RevGrad: $D \rightarrow W$

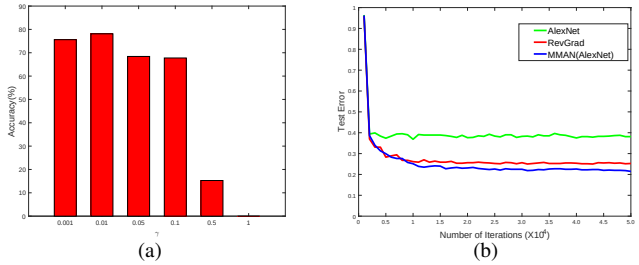
(g) MMAN: $A \rightarrow W$ (A) (h) MMAN: $A \rightarrow W$ (W) (i) MMAN: $A \rightarrow W$ (j) MMAN: $D \rightarrow W$ (D) (k) MMAN: $D \rightarrow W$ (W) (l) MMAN: $D \rightarrow W$

Fig. 3: The representation visualization over transfer tasks $A \rightarrow W$ and $D \rightarrow W$ in office-31. Here, we demonstrate the effectiveness of ourmethod through the learned representation visualization using the t-distributed stochastic neighbor embedding (t-SNE) [73]. Different colors represents different categorizes except red and blue. Red points are source samples and blue ones are target samples. (a)-(f) are trained with baseline methods, *i.e.*, RevGrad [60]. (g)-(l) are trained by MMAN. For each model, we provide three visualization results for each transfer task. For example, as for RevGrad in task $A \rightarrow W$, (a)-(c) respectively show visualization of source features, target features and hybrid features (source and target features). Compared to RevGrad, our methods successfully fuse the source and target domain features. In addition, features in the same class are mapped closer and features with different classes are dispersed, which shows the features of MMAN are more discriminative.

experiment phenomenon. In addition, WDGRL proposes an improved adversarial loss to reduce domain discrepancy while MMAN uses the original adversarial loss, which may lead to performance drop. Even so, MMAN can achieve the best average performance among the mentioned methods.

*5) **Further Remarks**:* In this section, we will further discuss some properties of the proposed model.

Feature Visualization: As shown in Figure 3 (a)-(l), we visualize the features generated by the **Feature Extractor** in task $\mathbf{A} \rightarrow \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{A}$ learned by RevGrad and MMAN using t-SNE embedding [73], respectively. As for results on RevGrad in Figure 3 (a)-(f), features are successfully fused but it also exhibits a serious problem: features generated are near class boundary. For example, features of class A in target domain could be easily mapped to the intermediate space between class A and class B, which is obviously a damage to classification tasks. In contrast, the representations learned by our method as shown in Figure 3 (g)-(l) are more discriminative and domain-invariant. Specifically, features in the same class are mapped closer. In particular, features with different classes are dispersed, making the features more discriminative. The well-behaved learned features illustrate the effectiveness of the proposed multi-channel constraint which mines the fine grained categories information. In addition, the feature distributions of source and target obtained from the MMAN are much similar than those generated by RevGrad, which emphasizes MMAN really can learn a common feature space shared by the source and target. These observations suggest that the adaptation of MMAN is a powerful approach to unsupervised domain adaptation.



(a) (b)

Fig. 4: (a) The classification accuracy results with different γ. (b) The test error of AlexNet, RevGrad and MMAN during training.

Parameter Setting and Sensitivity: We check the sensitivity of the MMAN parameter γ, *i.e.*, the triplet loss weight parameter of MMAN. Figure 4(a) shows the variation tendency of accuracy with different γ on task $\mathbf{A} \rightarrow \mathbf{W}$. The accuracy firstly increases and then decreases as γ varies. The change of accuracy illustrates that the performance of MMAN is a little sensitive to the hyper-parameter γ. The reason is that the proposed multi-channel constraint acted on labeled source data, has a great power to matching data distribution. However, the labeled source images are limited for training deep models, which may cause overfitting on the source domain and low generalization on the target domain.

Convergence Performance: As the MMAN involves adversarial training and triplets sampling methods, which may cause the unstable training process [63], it is necessary to test the convergence performance of MMAN. The convergence performance of MMAN is shown in Figure 4(b). The results have shown the error rate of different methods on task $\mathbf{A} \rightarrow \mathbf{W}$, which supports the following conclusions: (1) Adversarial training and triplets sampling methods do not trigger the unstable training. In addition, the offline triplets sampling does not cause low training efficiency in our experimental settings. On the contrary, the MMAN has similar convergence speed as RevGrad. (2) Although the MMAN does not speed up

TABLE IV: Details of social event dataset including social event name, duration, and documents numbers of each event.

Event ID	Event Name	Start Time	End Time	Google (G)		Flicker (F)	
				Images	Text	Images	Texts
1	Senkaku Islands Dispute	2008.06	2012.12	3743	2945	6617	6617
2	Occupy Wall Street	2011.09	2012.09	5601	3108	7151	7151
3	United States Presidential Election	2009.10	2013.01	5169	3446	7352	7352
4	War in Afghanistan	2001.10	2012.08	5373	2915	7172	7172
5	North Korea nuclear program	2000.01	2012.04	3969	2640	8635	8635
6	Greek protests	2011.05	2012.04	3900	2630	7385	7385
7	Mars Reconnaissance Orbiter	2005.04	2012.08	3901	2600	7188	7188
8	Syrian civil war	2011.01	2013.01	4899	3266	7426	7426

the convergence, it has significantly improved accuracy in the whole procedure of convergence thanks to the multi-channel constraint.

B. Cross Domain Event Recognition

*1) **Dataset**:* For social event recognition task, a popular dataset called social event detection (SED) dataset [74] has been proposed. However, there remains several problems so that we cannot apply SED dataset to our experiments: (1) The dataset contains only social media descriptions without any popular social events. (2) In addition, it does not contain multi-modality cross-domain information. For the above reasons, the existing SED datasets cannot meet with our requirements, so we have to collect a new dataset for our evaluation. To analyze social event data with multi-modality and multi-domain properties, we mainly focus on 8 complex and public social events that has occurred in the past few years. And we follow the procedure of [75–77] to collect the dataset by ourselves from Google News and Flcikr. The collected 8 social events cover a wide range of topics including politics, economics, military, society, and so on. For each social event, there are about 2000 to 9000 documents including texts and its corresponding images. The details of the dataset are displayed in TABLE IV.

*2) **Baseline**:* In order to illustrate the effectiveness of different components of the proposed model, ablation study is carried out on the task Google → Flicker and Flicker → Google. As shown in TABLE V, we study three variants of MMAN, i.e. training without gradient reverse layer and domain classifier, training without multi-channel constraints, and training without stacked attention based feature extractor (directly combine the features generated by CNN and LSTM as final multi-modality representation).

*3) **Setting Up**:* The input images are scaled while p-reserving aspect ratio and centre cropped to 299×299 dimensions. Image features are extracted from pretrained 152 layer ResNet [58] model. We take the last layer before the average pooling layer (of size 14×14×2048) and perform l_2 normalization on the depth (last) dimension. The input text description is tokenized and embedded to a 300 dimensional vector. The state size of LSTM layer is set to 1024. The depth C of the second convolution layer in stacked attention model is 2 and the size of the final fully connected layer is 1024. We use dropout 0.5 on input features of all layers including the LSTM, convolutions, and the fully connected layers.

We optimize this model with Adam optimizer [78] with batch size 128. We use exponential decay to gradually decrease the learning rate according to the following equation.

$$l_{step} = 0.5^{\frac{step}{decay\ steps}} l_0$$

The initial learning rate is set to $l_0 = 0.001$, and the decay step is set to $50K$. We set $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

TABLE V: Classification accuracy (%) on social event dataset for unsupervised domain adaptation. From left to right: transfer methods, two transfer tasks and average classification accuracy. All methods use ResNet as feature extractor. Notation: Google (G) and Flicker (F). G→F represents the transfer task from Google to Flicker.

Transfer Task	G → F	F → G	Avg
MMAN	**40.85**	**20.33**	**30.59**
No GRL, Domain Classifier	$32.38_{\downarrow 8.47}$	$14.36_{\downarrow 5.97}$	$23.37_{\downarrow 7.22}$
No Multi-channel Constraint	$35.68_{\downarrow 5.17}$	$17.57_{\downarrow 2.76}$	$26.63_{\downarrow 3.96}$
No Stacked Attention	$39.73_{\downarrow 1.12}$	$19.84_{\downarrow 0.49}$	$29.79_{\downarrow 0.62}$

*4) **Results**:* As shown in TABLE V, we report the classification accuracy on social event dataset for unsupervised domain adaptation. Based on the results, we can find that the performance of MMAN and its variants are quite low. We can attribute it to two reasons: (1) The event dataset is directly collected from the Internet so that the content information of samples is quite complex. (2) In terms of sample numbers and quality, the same event data collected from Google news and Flicker is different, which can dramatically cause a huge domain shift between the two domains. All above factors make the event dataset quite challenging for social event recognition task. However, based on the results, we can still observe the effectiveness of different components of our model: (1) Compared with different variants with MMAN, we can observe that the performance decline of model without GRL and domain classifier is the largest among three variants of MMAN. It indicates that the gradient reverse layer and domain classifier can play important roles in reducing domain discrepancy, especially domain gap caused by multi-modality data. (2) Without the multi-channel constrain, there appears obvious performance drop, *i.e.*, 5.17% on Google→Flicker and 2.76% on Flicker→Google. The experiment phenomenon illustrates that multi-channel constraint can capture fine-grained categories information and guides the network to learn more discriminative representations. (3) Note that the accuracy improvement of the MMAN is not significant in comparison with

the model without the stacked attention, we can still conclude that with the help of the stacked attention, multi-modality data feature extractor can generate semantic representation that benefits recognition task on the target domain.

V. Conclusion

This paper presents a multi-channel constraint based multi-modality adversarial network for unsupervised domain adaptation. Our network is able to not only learn transferable multi-modality features by stacked attention and transfer framework, but also exploit source domain fine-grained categories information that could enhance the discrimination of target samples and boost target performance on single-modality and multi-modality domain adaptation problem. Our approach successfully improves the performance on single-modality benchmark datasets and multi-modality social event dataset collected by ourselves for unsupervised domain adaptation. Our future work would focus on the basic theory of multi-modality transfer learning and applications on cross-domain social event recognition in multimedia.

References

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[6] P. Daras, S. Manolopoulou, and A. Axenopoulos, "Search and retrieval of rich media objects supporting multiple multimodal queries," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 734–746, June 2012.

[7] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multiordered discriminative structured subspace learning," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1220–1233, June 2017.

[8] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *ICCV*, 2017, pp. 2631–2639.

[9] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *ECCV*, 2016, pp. 451–466.

[10] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, and X. Xue, "Weakly supervised dense video captioning," in *CVPR*, 2017, pp. 5159–5167.

[11] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010, pp. 251–260.

[12] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2014.

[13] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *ACM MM*, 2014, pp. 7–16.

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011, pp. 689–696.

[15] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks." in *IJCAI*, 2016, pp. 3846–3853.

[16] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Worshop at ICML*, vol. 79, 2012.

[17] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

[18] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[19] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[20] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *NIPS*, 2007, pp. 513–520.

[21] C.-A. Hou, Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, "Unsupervised domain adaptation with label and structural consistency," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5552–5562, 2016.

[22] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Transactions on Cybernetics*, 2018.

[23] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *CVPR*, 2017, pp. 945–954.

[24] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[25] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015, pp. 97–105.

[26] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[27] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016, pp. 469–477.

[28] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017, pp. 1857–1865.

[29] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 2962–2971.
[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.
[31] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *NIPS*, 2008, pp. 1433–1440.
[32] J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, "Covariate shift by kernel mean matching," *Dataset Shift in Machine Learning*, pp. 131–160.
[33] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.
[34] R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2288–2302, 2014.
[35] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*, 2012, pp. 2066–2073.
[36] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *ICML*, 2013, pp. 222–230.
[37] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016, pp. 2058–2065.
[38] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *CVPR*, 2013, pp. 407–414.
[39] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *ICML*, 2004, p. 114.
[40] T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," in *NIPS*, 2009, pp. 809–816.
[41] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013, pp. 2960–2967.
[42] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008, pp. 1096–1103.
[43] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *ICML*, 2017, pp. 2208–2217.
[44] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *ICCV*, 2015, pp. 4068–4076.
[45] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *ICML*, 2018, pp. 5419–5428.
[46] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.
[47] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. Frank Wang, "Detach and adapt: Learning cross-domain disentangled deep representation," in *CVPR*, 2018, pp. 8867–8876.
[48] H. Theil and C.-F. Chung, "Relations between two sets of variates: The bits of information provided by each variate in each set," *Statistics & probability letters*, vol. 6, no. 3, pp. 137–139, 1988.
[49] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
[50] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010, pp. 251–260.
[51] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, Tech. Rep., 1986.
[52] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[53] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *NIPS*, 2007, pp. 153–160.
[54] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, "Vqa: Visual question answering," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 4–31, 2017.
[55] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *EMNLP*, 2016, pp. 457–468.
[56] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," *arXiv preprint arXiv:1704.03162*, 2017.
[57] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016, pp. 21–29.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[60] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.
[61] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
[62] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.
[63] C. Y. Wu, R. Manmatha, A. J. Smola, and P. K, "Sampling matters in deep embedding learning," in *ICCV*,

2017, pp. 2859–2867.
[64] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *NIPS*, 2016, pp. 136–144.
[65] L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *CVPR*, 2018, pp. 1498–1507.
[66] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," *AAAI*, 2018.
[67] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *ECCV*, 2018, pp. 37–52.
[68] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *CVPR*, 2017, pp. 5150–5158.
[69] B. Gholami, O. Rudovic, and V. Pavlovic, "Punda: Probabilistic unsupervised domain adaptation for knowledge transfer across visual categories," in *ICCV*, 2017, pp. 3601–3610.
[70] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *NIPS*, 2014, pp. 3320–3328.
[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[72] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.
[73] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[74] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, and S. Geva, "Social event detection at mediaeval 2013: Challenges, datasets, and evaluation," in *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona*, 2013.
[75] S. Qian, T. Zhang, and C. Xu, "Cross-domain collaborative learning via discriminative nonparametric bayesian model," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2086–2099, 2018.
[76] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 233–246, 2016.
[77] S. Qian, T. Zhang, and C. Xu, "Online multimodal multiexpert learning for social event tracking," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2733–2748, 2018.
[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Xinhong Ma received the bachelor's degree in Automation from Beijing Institute of Technology, Beijing, China, in 2017. He is currently pursuing the master degree at the Multimedia Computing Group, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include multimedia analysis and computer vision, especially deep learning, multimedia computing and transfer learning.

Tianzhu Zhang (M'11) received the bachelor's degree in communications and information technology from Beijing Institute of Technology, Beijing, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011. Currently, he is an Associate Professor at the Institute of Automation, Chinese Academy of Sciences. His current research interests include computer vision and multimedia, especially action recognition, object classification, object tracking, and social event analysis.

Changsheng Xu (M'97–SM'99–F'14) is a Professor in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and Executive Director of China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. He has hold 30 granted/pending patents and published over 200 refereed research papers in these areas. Dr. Xu is an Associate Editor of IEEE Trans. on Multimedia, ACM Trans. on Multimedia Computing, Communications and Applications and ACM/Springer Multimedia Systems Journal. He received the Best Associate Editor Award of ACM Trans. on Multimedia Computing, Communications and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He served as Program Chair of ACM Multimedia 2009. He has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops. He is IEEE Fellow, IAPR Fellow and ACM Distinguished Scientist.