Deep Multimodal Transfer Learning for Cross-Modal Retrieval

Liangli Zhen, Peng Hu, Xi Peng, Rick Siow Mong Goh, and Joey Tianyi Zhou

***Abstract*— Cross-modal retrieval (CMR) enables flexible retrieval experience across different modalities (e.g., texts versus images), which maximally benefits us from the abundance of multimedia data. Existing deep CMR approaches commonly require a large amount of labeled data for training to achieve high performance. However, it is time-consuming and expensive to annotate the multimedia data manually. Thus, how to transfer valuable knowledge from existing annotated data to new data, especially from the known categories to new categories, becomes attractive for real-world applications. To achieve this end, we propose a deep multimodal transfer learning (DMTL) approach to transfer the knowledge from the previously labeled categories (source domain) to improve the retrieval performance on the unlabeled new categories (target domain). Specifically, we employ a joint learning paradigm to transfer knowledge by assigning a pseudolabel to each target sample. During training, the pseudolabel is iteratively updated and passed through our model in a self-supervised manner. At the same time, to reduce the domain discrepancy of different modalities, we construct multiple modality-specific neural networks to learn a shared semantic space for different modalities by enforcing the compactness of homoinstance samples and the scatters of heteroinstance samples. Our method is remarkably different from most of the existing transfer learning approaches. To be specific, previous works usually assume that the source domain and the target domain have the same label set. In contrast, our method considers a more challenging multimodal learning situation where the label sets of the two domains are different or even disjoint. Experimental studies on four widely used benchmarks validate the effectiveness of the proposed method in multimodal transfer learning and demonstrate its superior performance in CMR compared with 11 state-of-the-art methods.**

***Index Terms*— Cross-modal retrieval (CMR), domain adaptation, multimodal learning, multimodal transfer learning.**

Manuscript received December 2, 2019; revised May 23, 2020 and August 1, 2020; accepted September 24, 2020. Date of publication October 22, 2020; date of current version February 4, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U19A2081, Grant U19A2078, Grant 61806135, and Grant 61625204; in part by the Fundamental Research Funds for the Central Universities under Grant YJ201949; and in part by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme under Project A18A1b0045 and Project A1892b0026. *(Liangli Zhen and Peng Hu contributed equally to this work.) (Corresponding author: Joey Tianyi Zhou.)*

Liangli Zhen, Rick Siow Mong Goh, and Joey Tianyi Zhou are with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: llzhen@outlook.com; gohsm@ihpc.a-star.edu.sg; joey.tianyi.zhou@gmail.com).

Peng Hu is with the College of Computer Science, Sichuan University, Chengdu 610065, China, and also with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: penghu.ml@gmail.com).

Xi Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: pengx.gm@gmail.com).

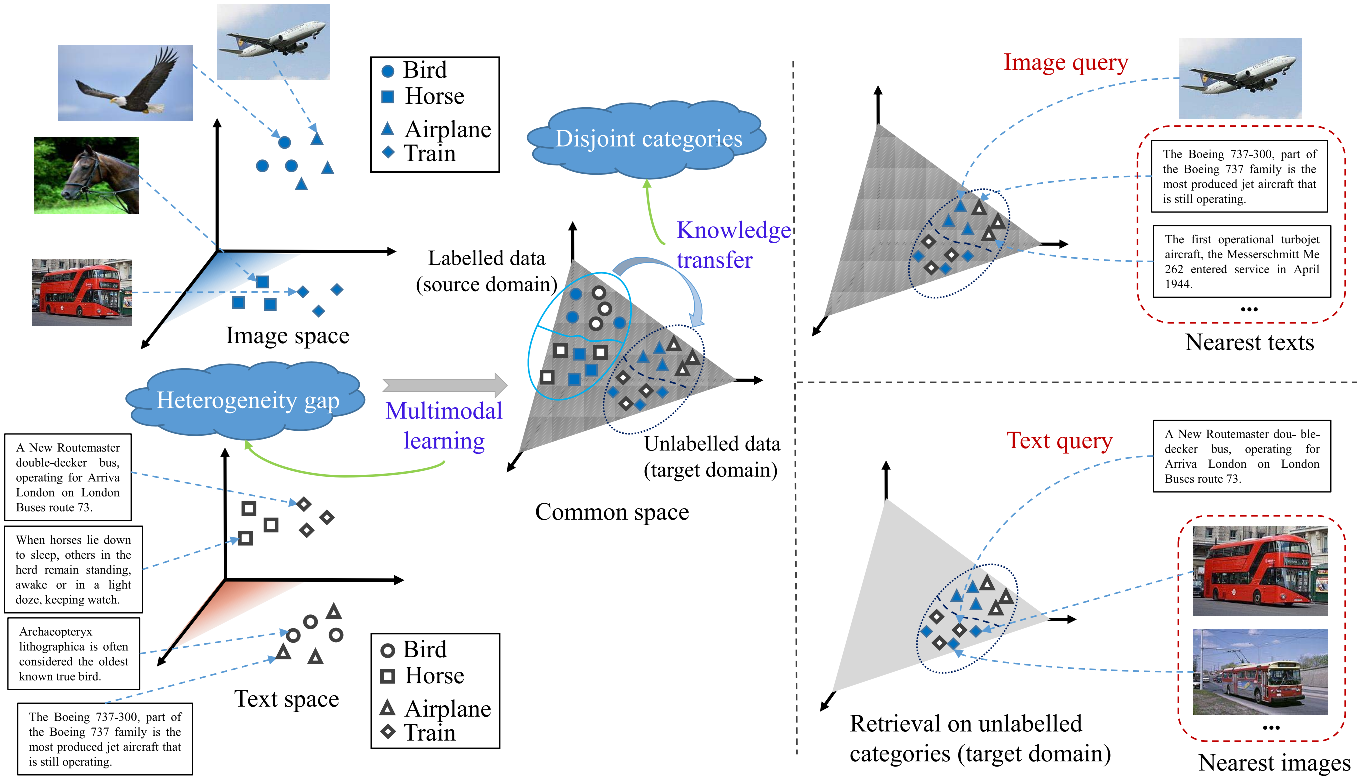
Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3029181

I. Introduction

Over the past decades, various types of media data, such as audios, texts, images, and videos, have shown explosive growth on Internet, and different types of data are usually used for describing the same event or topic [1]. For example, a post on Facebook may consist of paired texts and images. Cross-modal retrieval (CMR) provides an efficient way to search semantically relevant results of different modalities for a given query of any modality, enabling the users to earn more information about the concerned event/topic. It thus has attracted increasing interests in both academia and industry. One major challenge in CMR is that the distributions and representations of different media types are inconsistent, known as the heterogeneity gap [1], which makes measuring the distance between the samples from different modalities hard [2]. To bridge the heterogeneity gap, numerous multimodal analysis approaches [3], [4] have been developed to learn modality-specific transformations to map different modalities into a common space [5]. The early attempts are lying on exploring correlations from the heterogeneous data. In recent years, supervised methods [6]–[8], which utilize the label information to learn discriminative features for CMR, have been proposed and achieved much higher retrieval accuracy than unsupervised methods. Supervised approaches commonly require a large amount of labeled data for training to achieve high performance. However, we can get access to the abundant target data but with no labels, and it would be expensive to obtain the annotations for them, especially with various and dynamically increasing categories in the real-world applications.

To deal with such a problem, we investigate the possibility of transferring valuable knowledge from existing labeled categories (source domain) to unlabeled new categories (target domain) for boosting the CMR accuracy. As shown in Fig. 1, it is intractable to approach such a multimodal transfer learning case through conventional transfer learning methods due to the following two challenges: 1) how to bridge the heterogeneity gap between different modalities and 2) how to transfer semantic knowledge between two domains whose label sets are disjoint. To overcome the abovementioned two challenges, we propose a deep multimodal transfer learning (DMTL) approach to narrow the heterogeneity gap of different modalities and transfer the knowledge learned from the labeled categories to the unlabeled categories. More specifically, we develop a novel joint learning paradigm to transfer knowledge by synthesizing and assigning a pseudolabel to the samples from the new categories.



A New Routemaster double-decker bus, operating for Arriva London on London Buses route 73.

When horses lie down to sleep, others in the herd remain standing, awake or in a light doze, keeping watch.

Archaeopteryx lithographica is often considered the oldest known true bird.

The Boeing 737-300, part of the Boeing 737 family is the most produced jet aircraft that is still operating.

The Boeing 737-300, part of the Boeing 737 family is the most produced jet aircraft that is still operating.

The first operational turbojet aircraft, the Messerschmitt Me 262 entered service in April 1944.

A New Routemaster double-decker bus, operating for Arriva London on London Buses route 73.

Fig. 1. Two challenges of transferring knowledge from the known categories to new categories in multimodal learning. In the figure, different shapes represent different categories; the filled and unfilled shapes denote the image modality and the text modality, respectively. The source domain contains the labeled categories of bird and horse, whereas the target domain contains the unlabeled categories of airplane and train as a showcase.

During training, the pseudolabel is iteratively updated and feedbacks to our model in a self-supervised manner, thus boosting the retrieval performance. Furthermore, to bridge the heterogeneity gap, we construct multiple modality-specific neural networks to learn a shared semantic space, which could preserve the latent structure of different modalities by enforcing compactness of homoinstance samples and scatters of heteroinstance samples. The proposed method transfers the knowledge from every source modality driving the learning of all modality-specific networks forward. In other words, our approach can use all available modalities of the source domain, and it is flexible to the number of available modalities in the source domain.

Note that the problem setting of transfer learning in this work is different from semisupervised learning (SSL) [9]. SSL aims to exploit unlabeled data to improve the performance of a supervised learner. Generally, the categories of unlabeled data are the same as the categories of labeled data in SSL, such as semisupervised cross-modal learning approaches [10]–[12], while our proposed transfer learning method aims to leverage labeled data to improve the performance of unsupervised learning, where the categories of the unlabeled data (target domain) differ from the categories of labeled data (source domain). Furthermore, our method is remarkably different from the existing transfer learning approaches from the following aspects. On the one hand, unlike the conventional transfer learning methods that aim to explore the structure of the unlabeled data set that contains the same categories with the labeled data set, our method considers a more challenging situation where the label sets of the two domains are different or even disjoint. On the other hand, our method differs from existing pseudolabel-based methods. For instance, Ding *et al.* [13], Grandvalet and Bengio [14], and Zhong *et al.* [15] performed classification by minimizing the entropy of the output probabilistic vectors (i.e., pseudolabels) on the unlabeled data, and the cotraining method [16] leverages the consistency of the pseudolabels from two classifiers on unlabeled samples. Our method minimizes the difference between the pseudolabels on the unlabeled data with respect to two continuous iterations. In this manner, the knowledge extracted from the source domain is gradually transferred to the target domain so that the learned features could be discriminative and robust for the task at hand.

The main contributions and the novelty of this work could be summarized as follows.

1) We proposed a DMTL approach, which transfers the valuable knowledge from existing labeled data to unlabeled data in a joint learning manner so that the semantic relationships between the samples of the new unlabeled categories are discovered and exploited. To the best of our knowledge, this could be the first work which explores the joint learning paradigm of assigning pseudolabels to the samples of the new categories for knowledge transfer in the multimodal learning scenario.

2) We design modality-specific networks that aim to learn a conditional distribution space to match the desired distribution space induced by all the available data samples. Such a distribution approximation process could enforce the learned transformations be capable of narrowing the heterogeneity gap between different modalities.

II. Related Work

This section mainly discusses the related work in two lines and highlights the difference of our method by comparing it with existing ones.

A. Multimodal Learning

Multimodal learning aims to learn a set of modality-specific transformations that map the samples from different modalities into a shared space. As a result, the semantic similarity between the samples of different modalities could be measured in the shared space. With such a goal, a variety of approaches [17]–[21] have been proposed during the past decade, which could be divided into three groups: 1) unsupervised methods; 2) supervised methods; and 3) semisupervised methods. For the unsupervised methods, only the co-occurrence information is used to learn transformations for different modalities. One representative approach is canonical correlation analysis (CCA) [22], which determines the linear projections by maximizing the pairwise correlations between two sets of heterogeneous data samples. CCA is further extended to learn nonlinear transformations for different modalities, such as deep canonical correlation analysis (DCCA) [23] and deep canonically correlated autoencoders (DCCAEs) [24].

Different from unsupervised methods, supervised methods exploit label information to learn transformations for different modalities. In brief, they enforce to map the intraclass samples as close as possible, while the interclass samples to be far apart in the shared space. Following this manner, Kan *et al.* [25] proposed a multiview discriminant analysis (MvDA) method to learn the linear transformations by using a Fisher criterion. Zhen *et al.* [6] introduced deep supervised CMR (DSCMR) to extract discriminative features by minimizing the discrimination losses in the label space and the common representation space simultaneously. Xu *et al.* [26] adopted a hybrid matching approach to perform cross-modal attention for local region-word alignment and multilabel prediction for global semantic consistency. Shen *et al.* [27] proposed a dubbed subspace relation learning strategy to exploit relation information of labels in semantic space to make similar data from different modalities closer in the common Hamming subspace.

Besides, Zhang *et al.* [12] proposed a so-called generalized semisupervised structured subspace learning (GSS-SL) method to learn the transformations in a semisupervised way. All these methods learn the transformations from the source domain and directly apply the learned model to the target domain, thus suffering from the domain shift problem caused by the different data distributions of the source domain and the target domain.

This article proposes a multimodal learning strategy that transfers valuable knowledge from the labeled data to unlabeled data in a joint learning manner, where the source domain and the target domain have disjoint label sets.

B. Transfer Learning

Transfer learning aims to improve the learning of the target predictive function in the target domain by using the knowledge in the source domain [28]. In recent, transfer learning has shown effectiveness to deep learning, whose performance heavily relies on a large-scale well-annotated data set. Generally, transfer learning could be categorized into three groups: 1) unsupervised transfer learning [29]; 2) inductive transfer learning [30]; and 3) transductive transfer learning [31]. The common characteristic of the first two settings is that the target task is different from the source task. Their difference is that the inductive transfer learning uses some labeled samples from the target domain to induce an objective predictive model for the target domain. In contrast, unsupervised transfer learning assumes that there are no labeled data in both source and target domains during training. Different from these two settings, this article proposes a transductive transfer learning method, where the source and the target task are the same, but the source and target domains are different. Moreover, we have no labeled data in the target domain, but the labeled data in the source domain are available.

A large number of transductive transfer learning methods have been developed in the past decade. For example, Ding *et al.* [32] developed a graph adaptive knowledge transfer model, which jointly optimizes the target labels and domain-free features in a unified framework. Long *et al.* [33] proposed a residual transfer network approach, which simultaneously learns adaptive classifiers and transferable features. Tzeng *et al.* [34] outlined a generalized framework for adversarial transfer learning and developed the adversarial discriminative domain adaptation (ADDA) approach. These approaches all assume that the label sets of the source and target domains are the same. Zhang *et al.* [35] relaxed this assumption by treating the target label set as a subset of the source label set. However, this assumption is still easily violated. For instance, labeled animals from different data sets are easily accessible. We still face challenges when we want to conduct CMR on the samples of animals in the wild since some native species do not exist in the labeled data sets. Nevertheless, they are all focused on single-modal application scenarios, where the source and target domains share the same single modality (such as an image data set to another image data set).

In recent years, there are some attempts to knowledge transfer in a multimodal learning scenario. For example, modal-adversarial hybrid transfer network (MHTN) [36] is proposed to transfer knowledge from single-modal source domain to cross-modal target domain. Furthermore, some multiview domain adaption methods [37], [38] are proposed for classification tasks. In [37], all the data are employed to impose consistencies among multiple views, and the labeled data from the source domain is exploited to construct a large margin classifier. In [38], the maximum mean discrepancy regularizer is conducted to minimize the view disagreement under the CCA framework. However, these methods that require the data in the target domain are labeled or have the same categories of data in the source domain and the target domain.

In this work, we consider the multimodal transfer learning in the situation where the data in the target domain are unlabeled and have different categories with the labeled categories in the source domain. Such a transfer paradigm is essential to real-world CMR applications, which is also challenging because of the heterogeneity gap between different modalities and the semantic difference between the labeled categories in the source domain and the unlabeled categories in the target domain.

Recently, to address this challenge, a few zero-shot learning-based cross-modal learning methods have been proposed. Xu *et al.* [39] designed a self-supervised module to leverage the word vectors of both seen categories and unseen labels as guidance to enable the knowledge transfer and utilized the adversarial learning scheme to minimize the discrepancy among different modalities. Chi and Peng [40] proposed a dual adversarial network (DADN), which learns the transformations that could preserve the structure of the data set and strengthen relations of different categories. They have achieved promising performance, but they need the word vectors of both seen and unseen labels in the training process, while our method does not have such a constraint.

III. Our Proposed Method

In this section, we first introduce the CMR problem considered in this work. Then, we present our proposed method to transfer knowledge from the labeled categories to the unlabeled categories for boosting CMR performance.

A. Problem Description

Without loss of generality, we focus on CMR for bimodal data, e.g., for image and text data. We assume having $c_{\mathcal{S}}$ labeled categories (in source domain) with m instances of image-text pairs, denoted as $\mathcal{S} = \{(\mathbf{s}_i^\alpha, \mathbf{s}_i^\beta)\}_{i=1}^m$ and $c_{\mathcal{X}}$ unlabeled new categories (in target domain) with n instances of image-text pairs, denoted as $\mathcal{X} = \{(\mathbf{x}_j^\alpha, \mathbf{x}_j^\beta)\}_{j=1}^n$. Here, $\mathbf{s}_i^\alpha$ and $\mathbf{s}_i^\beta$ are the input image and text samples of the ith instance of the labeled data, and $\mathbf{x}_j^\alpha$ and $\mathbf{x}_j^\beta$ are the input image sample and text samples of the jth instance of the unlabeled data. Each pair of labeled samples in the source domain $(\mathbf{s}_i^\alpha, \mathbf{s}_i^\beta)$ has been assigned a semantic label vector $\mathbf{y}_i = [y_{1i}, y_{2i}, \ldots, y_{c_{\mathcal{S}}i}] \in \mathbb{R}^{c_{\mathcal{S}}}$. If the ith instance belongs to the kth category, then $y_{ki} = 1$; otherwise, $y_{ki} = 0$. However, the samples in the target domain are provided without labels, and the target categories are assumed to be disjoint with the labeled categories in the source domain, i.e., there is no overlap categories between $\mathcal{S}$ and $\mathcal{X}$. This kind of multimodal transfer learning problems is challenging and commonly exists in real-world applications.

The samples from different modalities cannot be directly compared since they are in different representation spaces and typically have different statistical properties. Multimodal learning aims to learn two transformations for the two modalities, i.e., h^α for the image modality and h^β for the text modality, to map the samples of different modalities into a common representation space. Therefore, the similarity of the samples from different modalities can be measured with commonly used metrics, e.g., the cosine similarity. Furthermore, in such a setting, the unlabeled new categories in the target domain are disjoint with the labeled known categories in the source domain. The underlying data distributions of the labeled categories and unlabeled new categories are different. If we learn the transformations from the source domain, then directly using them to the target domain without transfer learning would cause an unknown bias, i.e., so-called category gap. Thus, the learning approaches have to overcome such a category gap as well.

B. Framework of DMTL

The general framework of the proposed method is shown in Fig. 2. One could see that our model includes two coupled networks h^α and h^β, which projects the image and text modalities into a shared semantic space. More specifically, a convolutional neural network (CNN) is used to generate the original high-level semantic representation of the image modality. After that, several fully connected layers with nonlinear activation functions are connected to obtain the common representation for each image. To learn the common representation from the text modality, we employ a natural language processing network (NLP Net) to generate the high-level representation, which is further passed through several fully connected layers similarly. Each network is learned to map the input samples to approximate a conditional true matching distribution in the shared space. At the same time, the category information about the labeled data set in the source domain is exploited to learn the semantic relationships of the unlabeled samples in the target domain. More specifically, the common representations of the source and the target domain are both feedforward to a linear classifier for labels prediction. Different from the standard semisupervised scenario, where the unlabeled data set has the same categories as the labeled data set, our multimodal transfer learning has no overlap categories between the source domain and the target domain.

To the end, we develop a joint learning strategy to assign a soft pseudolabel for each sample in the target domain, and such a pseudo-label would feedback the information to guide the learning of the modality-specific transformations during the training process. In this manner, the heterogeneity gap as well as the category gap between different modalities is narrowed. Consequently, the relevant samples of different data types in the data set can be returned for one query of any data type for the unlabeled new categories in CMR.

C. Objective Function

The goal of DMTL is to transfer the knowledge from labeled data in the source domain to unlabeled data in the target domain, thus enjoying the discriminative and modality-invariant features learned for the CMR in the target domain. It transfers the knowledge from the labeled

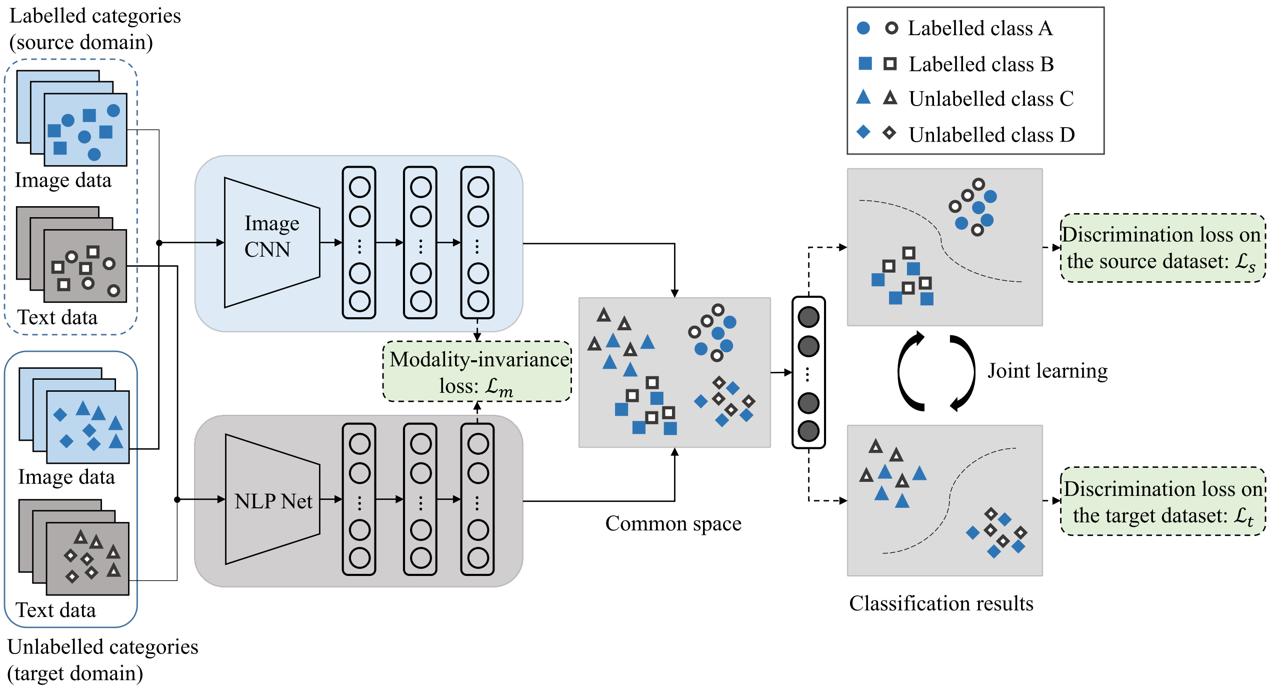


Fig. 2. General framework of our proposed DMTL method. It consists of two modality-specific subnetworks and a joint learning module. The samples from different modalities are mapped into a common shared semantic space such that the similarity of the samples from different modalities can be measured directly.

categories to the disjoint unlabeled categories. At the same time, it bridges the heterogeneity gap of different modalities. To achieve this goal, we propose to minimize the following objective function:

$$\mathcal{L} = \mathcal{L}_m + \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_t \tag{1}$$

where $\mathcal{L}_m$ is the modality-invariance loss, $\mathcal{L}_s$ and $\mathcal{L}_t$ are the discrimination losses for the labeled source data set and the unlabeled target data set, respectively, and λ_1 and λ_2 are the tradeoff parameters that control the contributions of these three terms. In the following, we will detail these three terms.

It is well known that the label information is greatly helpful to the learning of discriminative features. To exploit the label information of the labeled source data set, we propose to minimize the discrimination loss function

$$\mathcal{L}_s = \frac{1}{m}\sum_{i=1}^{m}\Big(\|\mathbf{P}h^{\alpha}(\mathbf{s}_i^{\alpha}) - \mathbf{y}_i\|_2 + \|\mathbf{P}h^{\beta}(\mathbf{s}_i^{\beta}) - \mathbf{y}_i\|_2\Big) \tag{2}$$

where $\mathbf{P}$ is the weight matrix of the linear classifier and h^{α} and h^{β} are the transformation functions for the image and text samples, respectively.

To explore the unlabeled target data set, we assume that there are pseudolabels $\mathbf{z}_j^{\alpha}$ and $\mathbf{z}_j^{\beta}$ (initialized randomly) for the image and text samples in the unlabeled target data set. The information of the pseudolabels is passed through the networks to learn discriminative features for unlabeled categories by minimizing the discrimination loss for the target data set

$$\mathcal{L}_t = \frac{1}{n}\sum_{j=1}^{n}\Big(\|\mathbf{P}h^{\alpha}(\mathbf{x}_j^{\alpha}) - \mathbf{z}_j^{\alpha}\|_2 + \|\mathbf{P}h^{\beta}(\mathbf{x}_j^{\beta}) - \mathbf{z}_j^{\beta}\|_2\Big) \tag{3}$$

where $\mathbf{P}$, h^{α}, and h^{β} are the same as the definitions in (2).

During training, the pseudolabels of the samples in the target data set will be updated at each iteration by

$$\begin{aligned}\mathbf{z}_j^{\alpha} &= \mathbf{P}h^{\alpha}(\mathbf{x}_j^{\alpha})\\ \mathbf{z}_j^{\beta} &= \mathbf{P}h^{\beta}(\mathbf{x}_j^{\beta}).\end{aligned} \tag{4}$$

It is worth noting that some works [13], [14] adopt the entropy minimization strategy on the predicted pseudolabels to improve the learning of discriminative features for the unlabeled data in the target domain. However, such a strategy is unsuitable for our setting, which includes different label sets between the source and target domains instead of having the same categories constraint for the existing methods. Another difference between our DMTL and the existing methods is that our pseudolabels are the flexible similarities/dissimilarities from the source domain to the target domain instead of the exactly predicted class labels in these methods. In other words, we use the categories of the source domain to represent/transfer the semantics of the target domain with the real values, i.e., pseudolabels, instead of predicting the exact label. The degrees of dissimilarity and similarity are both important for the semantic representation, and negative and positive values are used to measure the amount of the dissimilarity and similarity, respectively. The similarity representations (pseudolabels) are calculated by the linear classifier that is jointly trained on both the source and target domains. Thus, the knowledge of the source domain can be transferred into the target domain, and the semantic information can also be transmitted in the target domain. Our proposed strategy learns the discriminative features for the target data set by minimizing the difference of the pseudolabels between two continuous iterations. It makes the learning process more smooth and stable in terms of the prediction results for the unlabeled data, thus improving the knowledge transfer between the source and target domains. Also, to bridge the heterogeneity gap between different modalities, we design two coupled networks to learn modality-specific transformations. For each image sample $\mathbf{d}_i^\alpha \in \mathcal{S}^\alpha \bigcup \mathcal{X}^\alpha$, we define a conditional distribution over all text samples $\mathbf{d}_1^\beta, \mathbf{d}_2^\beta, \ldots, \mathbf{d}_{m+n}^\beta$, and $\mathbf{d}_j^\beta \in \mathcal{S}^\beta \bigcup \mathcal{X}^\beta$ such that

$$p(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha) = \frac{e^{-\|h^\alpha(\mathbf{d}_i^\alpha)-h^\beta(\mathbf{d}_j^\beta)\|_2}}{\sum_{k=1}^{m+n} e^{-\|h^\alpha(\mathbf{d}_i^\alpha)-h^\beta(\mathbf{d}_k^\beta)\|_2}} \tag{5}$$

where h^α and h^β are the transformation functions that are the same as the definitions in (2).

To achieve an ideal matching distribution, we assume that only the samples from the same pair are mapped to a single point and infinitely far away from the other samples, i.e.,

$$q(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The coupled networks are designed to enforce the transformation functions h^α and h^β that can make $p(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha)$ be as close as possible to $q(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha)$, thus preserving the modality-invariance of different modalities on both the source and the target data sets. To match these two distributions, we minimize the following KL divergence function from image to text:

$$\mathcal{J}_{i2t} = \frac{1}{m+n} \sum_{i=1}^{m+n} \sum_{j=1}^{m+n} q(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha) \log \frac{q(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha)}{p(\mathbf{d}_j^\beta|\mathbf{d}_i^\alpha) + \sigma} \tag{7}$$

where σ is a small number to avoid numerical problem and typically set it as 10^{-6}.

Similarly, the KL divergence function from text to image can be computed by exchange $\mathbf{d}_i^\alpha$ and $\mathbf{d}_j^\beta$ in (5)–(7) as

$$\mathcal{J}_{t2i} = \frac{1}{m+n} \sum_{i=1}^{m+n} \sum_{j=1}^{m+n} q(\mathbf{d}_i^\alpha|\mathbf{d}_j^\beta) \log \frac{q(\mathbf{d}_i^\alpha|\mathbf{d}_j^\beta)}{p(\mathbf{d}_i^\alpha|\mathbf{d}_j^\beta) + \sigma} \tag{8}$$

where σ is the same as the definition in (7).

At last, we obtain the modality-invariance loss by combining the losses in (7) and (8) as

$$\mathcal{L}_m = \mathcal{J}_{i2t} + \mathcal{J}_{t2i}. \tag{9}$$

Note that the heterogeneity gap between two different modalities is narrowed both on the source and the target data set by minimizing $\mathcal{L}_m$. The proposed method transfers not only the knowledge about different semantic categories but also the correlation relationship of different modalities.

D. Optimization

The objective function of DMTL in (1) involves two subproblems. For the first subproblem, we use a stochastic gradient descent optimization algorithm to update the weights by fixing the pseudolabels of the new category samples. After that, we exploit the learned parameters to update the pseudolabels $\mathbf{z}_j^\alpha$ and $\mathbf{z}_j^\beta$ in (4). These two subproblems are optimized alternatively.

To solve the first subproblem, we assume that $\mathbf{z}_j^\alpha$ and $\mathbf{z}_j^\beta$ are assigned and calculate the gradients of the objective function $\mathcal{L}$ in(1) with respect to the parameters of the two coupled networks Θ_α and Θ_β as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Theta^\alpha} &= \frac{\partial \mathcal{L}_m}{\partial \Theta^\alpha} + \lambda_1 \frac{\partial \mathcal{L}_s}{\partial \Theta^\alpha} + \lambda_2 \frac{\partial \mathcal{L}_t}{\partial \Theta^\alpha} \\ \frac{\partial \mathcal{L}}{\partial \Theta^\beta} &= \frac{\partial \mathcal{L}_m}{\partial \Theta^\beta} + \lambda_1 \frac{\partial \mathcal{L}_s}{\partial \Theta^\beta} + \lambda_2 \frac{\partial \mathcal{L}_t}{\partial \Theta^\beta}. \end{aligned} \tag{10}$$

Regarding to Θ^α, we have

$$\frac{\partial \mathcal{L}_s}{\partial \Theta^\alpha} = \frac{1}{m} \sum_{i=1}^{m} \frac{\mathbf{P}^T \mathbf{P} h^\alpha(\mathbf{s}_i^\alpha) - \mathbf{P}^T \mathbf{y}_i}{\|\mathbf{P} h^\alpha(\mathbf{s}_i^\alpha) - \mathbf{y}_i\|_2} \frac{\partial h^\alpha(\mathbf{s}_i^\alpha)}{\partial \Theta^\alpha} \tag{11}$$

$$\frac{\partial \mathcal{L}_t}{\partial \Theta^\alpha} = \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{P}^T \mathbf{P} h^\alpha(\mathbf{x}_j^\alpha) - \mathbf{P}^T \mathbf{z}_j^\alpha}{\|\mathbf{P} h^\alpha(\mathbf{x}_j^\alpha) - \mathbf{z}_j^\alpha\|_2} \frac{\partial h^\alpha(\mathbf{x}_j^\alpha)}{\partial \Theta^\alpha} \tag{12}$$

and

$$\begin{aligned} \frac{\partial \mathcal{L}_m}{\partial \Theta^\alpha} &= \frac{\partial \left(\frac{1}{m+n} \sum_{j=1}^{m+n} \log \frac{1}{p(\mathbf{d}_j^\beta|\mathbf{d}_j^\alpha) + \sigma} \right)}{\partial \Theta^\alpha} \\ &= -\frac{1}{m+n} \sum_{j=1}^{m+n} \frac{1}{p\left(\mathbf{d}_j^\beta|\mathbf{d}_j^\alpha\right) + \sigma} \frac{\partial p\left(\mathbf{d}_j^\beta|\mathbf{d}_j^\alpha\right)}{\partial \Theta^\alpha}. \end{aligned} \tag{13}$$

Denoting $\|h^\alpha(\mathbf{d}_j^\alpha) - h^\beta(\mathbf{d}_k^\beta)\|_2$ in (5) as ξ_k for $k \in \{1, 2, \ldots, m+n\}$, we obtain that

$$\begin{aligned} &\frac{\partial p(\mathbf{d}_j^\beta|\mathbf{d}_j^\alpha)}{\partial \Theta^\alpha} \\ &\quad = \frac{(\sum_{k=1}^{m+n} e^{-\xi_k}) \frac{1}{\xi_j} e^{-\xi_j} (h^\beta(\mathbf{d}_j^\beta) - h^\alpha(\mathbf{d}_j^\alpha)) \frac{\partial h^\alpha(\mathbf{d}_j^\alpha)}{\partial \Theta^\alpha}}{(\sum_{k=1}^{m+n} e^{-\xi_k})^2} \\ &\quad\quad + \frac{e^{-\xi_j} \sum_{k=1}^{m+n} \frac{1}{\xi_k} e^{-\xi_k} (h^\beta(\mathbf{d}_k^\beta) - h^\alpha(\mathbf{d}_j^\alpha)) \frac{\partial h^\alpha(\mathbf{d}_j^\alpha)}{\partial \Theta^\alpha}}{(\sum_{k=1}^{m+n} e^{-\xi_k})^2}. \end{aligned} \tag{14}$$

Regarding to Θ^β, we have

$$\frac{\partial \mathcal{L}_s}{\partial \Theta^\beta} = \frac{1}{m} \sum_{i=1}^{m} \frac{\mathbf{P}^T \mathbf{P} h^\beta\left(\mathbf{s}_i^\beta\right) - \mathbf{P}^T \mathbf{y}_i}{\left\|\mathbf{P} h^\beta\left(\mathbf{s}_i^\beta\right) - \mathbf{y}_i\right\|_2} \frac{\partial h^\beta\left(\mathbf{s}_i^\beta\right)}{\partial \Theta^\beta} \tag{15}$$

$$\frac{\partial \mathcal{L}_t}{\partial \Theta^\beta} = \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{P}^T \mathbf{P} h^\beta\left(\mathbf{x}_j^\beta\right) - \mathbf{P}^T \mathbf{z}_j^\beta}{\left\|\mathbf{P} h^\beta\left(\mathbf{x}_j^\beta\right) - \mathbf{z}_j^\beta\right\|_2} \frac{\partial h^\beta\left(\mathbf{x}_j^\beta\right)}{\partial \Theta^\beta} \tag{16}$$

and

$$\begin{aligned} \frac{\partial \mathcal{L}_m}{\partial \Theta^\beta} &= \frac{\partial \left(\frac{1}{m+n} \sum_{j=1}^{m+n} \log \frac{1}{p(\mathbf{d}_j^\alpha|\mathbf{d}_j^\beta) + \sigma} \right)}{\partial \Theta^\beta} \\ &= -\frac{1}{m+n} \sum_{j=1}^{m+n} \frac{1}{p(\mathbf{d}_j^\alpha|\mathbf{d}_j^\beta) + \sigma} \frac{\partial p\left(\mathbf{d}_j^\alpha|\mathbf{d}_j^\beta\right)}{\partial \Theta^\beta}. \end{aligned} \tag{17}$$

Algorithm 1 Optimization Procedure of Our DMTL

```
Input: The labelled source dataset $\mathcal{S}$ and the corresponding semantic labels $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$, the unlabelled target dataset $\mathcal{X}$, the dimensionality of the shared semantic space $d$, the batch size $m_b$, the learning rate $\tau$ and the parameter $\lambda$.
Output: The optimal coupled networks for the transformation functions $h^{\alpha}$ and $h^{\beta}$.
Randomly initialise the weight parameters of the two coupled networks $\Theta_{\alpha}$ and $\Theta_{\beta}$, the weight matrix of the linear classifier $\mathbf{P}$, and the pseudo-labels $\mathbf{z}_j^{\alpha}$ and $\mathbf{z}_j^{\beta}$ ($j = 1, 2, \dots, n$).
while not converge do
    for $\ell = 1, 2, \dots, \lfloor \frac{m+n}{m_b} \rfloor\}$ do
        Randomly sample $m_b$ image-text pair samples from $\mathcal{S} \bigcup \mathcal{X}$ to construct a mini-batch.
        Calculate the objective function in Equation (1) based on the mini-batch.
        Update the parameters $\Theta^{\alpha}$ and $\Theta^{\beta}$ by minimising $\mathcal{L}$ in Equation (1) with descending their stochastic gradient:
            $\Theta^{\alpha} \leftarrow \Theta^{\alpha} - \tau \frac{\partial \mathcal{L}}{\partial \Theta^{\alpha}}$;
            $\Theta^{\beta} \leftarrow \Theta^{\beta} - \tau \frac{\partial \mathcal{L}}{\partial \Theta^{\beta}}$.
        Update the pseudo-labels $\mathbf{z}_j^{\alpha}$ and $\mathbf{z}_j^{\beta}$ via Equation (4).
    end for
end while
```

Denoting $\|h^{\beta}(\mathbf{d}_j^{\beta}) - h^{\alpha}(\mathbf{d}_k^{\alpha})\|_2$ as ζ_k for $k \in \{1, 2, \dots, m+n\}$, we obtain that

$$\begin{aligned} &\frac{\partial p(\mathbf{d}_j^{\alpha}|\mathbf{d}_j^{\beta})}{\partial \Theta^{\beta}} \\ &\quad = \frac{(\sum_{k=1}^{m+n} e^{-\zeta_k}) \frac{1}{\zeta_j} e^{-\zeta_j} (h^{\alpha}(\mathbf{d}_j^{\alpha}) - h^{\beta}(\mathbf{d}_j^{\beta})) \frac{\partial h^{\beta}(\mathbf{d}_j^{\beta})}{\partial \Theta^{\beta}}}{(\sum_{k=1}^{m+n} e^{-\zeta_k})^2} \\ &\qquad + \frac{e^{-\zeta_j} \sum_{k=1}^{m+n} \frac{1}{\zeta_k} e^{-\zeta_k} (h^{\alpha}(\mathbf{d}_k^{\alpha}) - h^{\beta}(\mathbf{d}_j^{\beta})) \frac{\partial h^{\beta}(\mathbf{d}_j^{\beta})}{\partial \Theta^{\beta}}}{(\sum_{k=1}^{m+n} e^{-\zeta_k})^2}. \end{aligned} \tag{18}$$

Then, Θ^{α} and Θ^{β} can be updated by using the gradient descent algorithm as follows until the termination condition been reached:

$$\begin{aligned} \Theta^{\alpha} &\leftarrow \Theta^{\alpha} - \tau \frac{\partial \mathcal{L}}{\partial \Theta^{\alpha}} \\ \Theta^{\beta} &\leftarrow \Theta^{\beta} - \tau \frac{\partial \mathcal{L}}{\partial \Theta^{\beta}} \end{aligned} \tag{19}$$

where τ is the learning rate.

Once we have optimized the model parameters, the pseudolabels for the unlabeled data in the target domain are updated via (4). Then, the updated pseudolabels are further used to optimize the deep model parameters Θ^{α} and Θ^{β}.

The optimization procedure of our method is summarized in Algorithm 1. The maximal number of training epochs T is taken as the termination condition in this work.

IV. Experimental Study

To verify the effectiveness of our proposed method, we conduct experiments on four widely used benchmark data sets: the Pascal Sentences data set [41], the Wikipedia data set [42], the NUS-WIDE data set [43], and the XMediaNet data set [2]. In the experiments, we first compare the proposed DMTL method with state-of-the-art methods to evaluate its performance. Then, we provide further analysis of DMTL. It includes the convergence investigation, the visualization of the learned representations in the common space, and the impacts of different components in (1).

TABLE I

Statistical Results of the Four Benchmark Data Sets Used in Our Experiments, Where d_i and d_t Are the Dimensionalities of the Image and Text Features, Respectively

Dataset	# train data	# test data	# classes	d_i	d_t
Pascal Sentences	800	200	20	4,096	300
Wikipedia	2,173	693	10	4,096	300
NUS-WIDE	42,942	28,661	10	4,096	300
XMediaNet	32,000	8,000	200	4,096	300

A. Data Sets and Features

Pascal Sentences [41] is selected from the 2008 PASCAL development kit, which contains 1000 image-text pairs. Each pair has one image and its description of five sentences. The image-text pairs are classified into 20 semantic classes, such as person, bird, and train. We split the data set into two sets: a training set of 800 image-text pairs and a test set of 200 image-text pairs.

Wikipedia [42] is constructed from Wikipedia's "featured articles," which contains a total of 2866 image-text pairs of ten semantic categories. Each image-text pair only belongs to one category. In each pair, it contains a single image and a text of (several paragraphs) description about this image. The categories are of high-level semantics, such as art, history, and sports. The data set is randomly split as a training set of 2173 pairs and a test set of 693 pairs by following [42].

NUS-WIDE [43] is originally a real-world web image data set. It contains about 270 000 images with their tags from 81 semantic categories. The data set is highly imbalanced in different categories. In the experiments, we only select ten largest categories by following [44] and [45], and the set of the tags of an image is viewed as the text description of the target image. As a result, there are 71 643 image-text pairs, where 42 942 image-text pairs are for training and 28 661 image-text pairs are for testing.

XMediaNet [2] is a large-scale data set with five modalities, which contains 100 000 instances of text, image, audio, video, and 3-D model. These instances are grouped into 200 semantic categories: 153 artifact species and 47 animal species. In this work, we use image-text data in XMediaNet for CMR experiments by following [40]. There are 40 000 image-text pairs, which are divided into two parts: a training set of 32 000 image-text pairs and a test set of 8000 image-text pairs.

In the experiments, to evaluate the performance of the transfer learning from the existing labeled categories to unlabeled new categories, i.e., the transductive setting [46], [47], we follow the data set partition and feature exaction strategies from [40]. Furthermore, we adopt a 19-layer VGGNet [48] to learn the representations of the image samples and obtain a 4096-D representation vector outputted by the fc7 layer of the network for each image. For representing text samples, we adopt the Doc2Vec model [49] to extract a 300-D representation vector for each text. The statistical results of the four data sets are summarized in Table I. It is notable that all the compared methods adopt the same image and text features as the features used in our method. For all the four data sets, inspired by [40] and [50], we further randomly split the training data set and the testing data set into a subset of labeled categories (source data set) and a subset of unlabeled new categories (target data set), respectively, and each subset includes 50% categories. The testing of the CMR is performed on the unlabeled new categories (target data set).

B. Evaluation Metric

In our experiments, we evaluate the performance of the compared methods for two different CMR tasks: 1) retrieving text samples using image queries (image-query-text) and 2) retrieving images using text queries (text-query-image). In the testing stage, we map the multimedia data into the common space using the multimodal learning methods. Then, for each image (or text) query, it returns the nearest neighbors from the text (or image) retrieval database by measuring the similarity between the common representation vectors of the query and the samples in the retrieval database with the cosine distance.

We adopt two evaluation metrics to evaluate the retrieval performance: the mean average precision (mAP) score [51] and the precision-recall (PR) curve [52]. The mAP score is the mean value of average precision (AP) values over all queries in the query set, and AP can be calculated as

$$\mathrm{AP} = \frac{1}{R}\sum_{r=1}^{N} P(r)\sigma(r) \tag{20}$$

where N is the number of samples in the retrieval database, R is the number of relevant items, $P(r)$ denotes the precision of the top r retrieved items, and $\sigma(r) = 1$ if the rth retrieved item is relevant to the query (i.e., the rth retrieved item belongs to the category of the query) and $\sigma(r) = 0$ otherwise. The mAP metric jointly considers the ranking information and precision, which is a widely used performance evaluation criterion in the research of CMR [1], [5]. The higher the mAP score, the better the performance.

The PR curve shows the tradeoff between precision and recall for different thresholds. In information retrieval, precision is a measure of result relevancy, whereas recall is a measure of how many truly relevant results are returned. A larger area under the PR curve represents both higher recall and higher precision, which indicates better performance.

C. Experimental Settings

In this work, we connect three fully connected layers with rectified linear unit (ReLU) [53] active function on the top of each high-level feature extractor (i.e., VGGNet and Doc2Vec) to output the representations in the common space. The numbers of the hidden units for the three layers are 4096, 4096, and 512, respectively. The entire model is trained in an end-to-end manner on an NVIDIA GTX 1080 Ti GPU in PyTorch.[1] For training, we employ the Adam optimizer [54] with a learning rate of 10^{-4} and a batch size of 100 and set the maximal number of epochs as 50.

TABLE II

PERFORMANCE COMPARISON IN TERMS OF THE (MEAN ± STD) MAP SCORES ON PASCAL SENTENCES OVER TEN TIMES OF MONTE CARLO SIMULATIONS. † REFERS TO THE RESULT REPORTED IN [40]

Training Set	Method	image→text	text→image	Average
Source	MvDA-VC [25]	0.215±0.014	0.213±0.014	0.214±0.011
	MvDA [55]	0.189±0.014	0.193±0.014	0.191±0.013
	MvDA-VC [25]	0.255±0.029	0.252±0.029	0.254±0.027
	ACMR [5]	0.275±0.034	0.258±0.021	0.267±0.027
	DANZCR [45]	$0.334^{\dagger}$	$0.338^{\dagger}$	$0.336^{\dagger}$
	DADN [40]	$0.359^{\dagger}$	$0.353^{\dagger}$	$0.356^{\dagger}$
	Ours	0.359±0.023	**0.363±0.034**	**0.361±0.027**
Source + Target	MCCA [56]	0.549±0.040	0.560±0.040	0.555±0.034
	PLS [57]	0.580±0.042	0.533±0.042	0.557±0.035
	DCCA [23]	0.570±0.037	0.555±0.037	0.563±0.033
	DCCAE [24]	0.576±0.038	0.567±0.038	0.571±0.034
	GSS-SL [12]	0.343±0.028	0.343±0.028	0.343±0.022
	Ours	**0.632±0.041**	**0.637±0.053**	**0.634±0.047**

D. Comparison With State-of-the-Art Methods

To verify the effectiveness of our proposed method, we compare DMTL with 11 state-of-the-art methods in the experiments, including six supervised learning-based methods, namely GMLDA [17], MvDA [55], MvDA-VC [25], ACMR [5], DANZCR [45], and DADN [40], four unsupervised learning-based methods, namely MCCA [56], PLS [57], DCCA [23] and DCCAE [24], and an SSL method GSS-SL [12]. The experiment on each data set has two settings: 1) only use the labeled data samples (source data set) in the training set to train the model and apply the model on the unlabeled categories of the test set for testing and 2) use both the labeled data samples (source data set) and the unlabeled data samples (target data set) in the training set to train the model and apply the model on the unlabeled categories of the test set for testing. Since the data set is randomly into a source data set of 50% categories and a target data set of the other 50% categories, we report the mean the standard deviation (std) of mAP scores over ten times of Monte Carlo simulations.

1) Results on Pascal Sentences: The mAP scores of our DMTL and the compared methods on Pascal Sentences [41] are reported in Table II, from which we have the following four observations.

1) Our method obtains the highest mAP scores under both of the two experimental settings. Specifically, it outperforms the peer methods with an improvement of 0.5% and 6.3% average mAP scores on the two settings, respectively.
2) The unsupervised methods and our method, which use both the labeled data (source domain) and the unlabeled data (target domain) for training, significantly outperform other supervised methods, which only use the labeled data (source domain) for training.
3) The performance of our method on the second setting is much better than that on the first setting. Specifically, it has an improvement of 27.3% in terms of average mAP scores.
4) GSS-SL, a semisupervised method, is inferior to other methods under the second experimental setting, even it uses the label information of source data set. The potential reason is that the semantic gap between the labeled categories (in the source domain) and the unlabeled categories (in the target domain) is quite large.

[1] Pytorch Open Source Toolkit at https://github.com/pytorch/pytorch

TABLE III

PERFORMANCE COMPARISON IN TERMS OF THE (MEAN ± STD) MAP SCORES ON WIKIPEDIA OVER TEN TIMES OF MONTE CARLO SIMULATIONS. † REFERS TO THE RESULT REPORTED IN [40]

Training Set	Method	image→text	text→image	Average
Source	GMLDA [17]	0.255±0.010	0.241±0.010	0.248±0.011
	MvDA [55]	0.263±0.014	0.249±0.014	0.256±0.014
	MvDA-VC [25]	0.258±0.014	0.249±0.014	0.254±0.013
	ACMR [5]	0.287±0.033	0.297±0.029	0.292±0.030
	DANZCR [45]	0.297†	0.287†	0.292†
	DADN [40]	0.305†	0.291†	0.298†
	Ours	**0.306±0.039**	**0.297±0.024**	**0.301±0.031**
Source + Target	MCCA [56]	0.344±0.055	0.415±0.055	0.380±0.046
	PLS [57]	0.492±0.063	0.474±0.063	0.483±0.061
	DCCA [23]	0.451±0.048	0.488±0.048	0.469±0.047
	DCCAE [24]	0.454±0.037	0.497±0.037	0.476±0.036
	GSS-SL [12]	0.250±0.017	0.248±0.017	0.249±0.015
	Ours	**0.531±0.076**	**0.574±0.078**	**0.552±0.076**

TABLE IV

PERFORMANCE COMPARISON IN TERMS OF THE (MEAN ± STD) MAP SCORES ON NUS-WIDE OVER TEN TIMES OF MONTE CARLO SIMULATIONS. † REFERS TO THE RESULT REPORTED IN [40]

Training Set	Method	image→text	text→image	Average
Source	GMLDA [17]	0.432±0.074	0.441±0.074	0.436±0.069
	MvDA [55]	0.408±0.062	0.412±0.062	0.410±0.062
	MvDA-VC [25]	0.417±0.064	0.416±0.064	0.416±0.064
	ACMR [5]	0.406±0.062	0.418±0.062	0.412±0.062
	DANZCR [45]	0.416†	0.469†	0.443†
	DADN [40]	0.423†	0.472†	0.448†
	Ours	**0.572±0.082**	**0.576±0.080**	**0.574±0.081**
Source + Target	MCCA [56]	0.636±0.101	**0.648±0.101**	0.642±0.100
	PLS [57]	0.629±0.091	0.636±0.091	0.633±0.089
	DCCA [23]	0.635±0.114	0.642±0.114	0.638±0.112
	DCCAE [24]	0.631±0.122	0.638±0.122	0.634±0.120
	GSS-SL [12]	0.398±0.063	0.391±0.063	0.394±0.062
	Ours	**0.656±0.091**	0.634±0.095	**0.645±0.092**

2) Results on Wikipedia: The mAP scores of our DMTL and the compared methods on Wikipedia [42] are reported in Table III, from which we can see that the following holds.

1) Our method outperforms other methods under both of the two experimental settings. Specifically, it achieves an improvement of 0.3% and 6.9% average mAP scores over the second best method on the two settings, respectively.
2) The mAP score of our method on the image→text retrieval is higher than that on the text→image retrieval, especially the case in the second experimental setting.
3) The mAP scores of the tested methods on Wikipedia are lower than the results on Pascal Sentences.

The potential reasons are that some images are not closely related to their corresponding text description and their assigned categories, which leads to the extracted image feature vectors cannot reflect the semantic properties appropriately, and the categories of Wikipedia are of high-level semantics, such as art, history, and sports. Some of the samples may belong to multiple categories but are assigned to one category in the labeling process, which makes the search much more difficult than that on Pascal Sentences, which contains the categories of lower level semantics, e.g., person, bird, and horse.

TABLE V

PERFORMANCE COMPARISON IN TERMS OF THE (MEAN ± STD) MAP SCORES ON XMEDIANET OVER TEN TIMES OF MONTE CARLO SIMULATIONS. † REFERS TO THE RESULT REPORTED IN [40]

Training Set	Method	image→text	text→image	Average
Source	GMLDA [17]	0.042±0.003	0.048±0.003	0.045±0.002
	MvDA [55]	0.051±0.003	0.058±0.003	0.054±0.002
	MvDA-VC [25]	0.054±0.003	0.063±0.003	0.058±0.003
	ACMR [5]	0.037±0.004	0.056±0.005	0.046±0.004
	DANZCR [45]	0.106†	0.117†	0.112†
	DADN [40]	0.112†	0.130†	0.121†
	Ours	0.089±0.005	0.122±0.006	0.106±0.005
Source + Target	MCCA [56]	0.115±0.005	0.163±0.005	0.139±0.005
	PLS [57]	0.088±0.003	0.103±0.003	0.096±0.003
	DCCA [23]	0.030±0.002	0.032±0.002	0.031±0.003
	DCCAE [24]	0.030±0.003	0.031±0.003	0.031±0.003
	GSS-SL [12]	0.031±0.002	0.027±0.002	0.029±0.002
	Ours	**0.736±0.020**	**0.700±0.013**	**0.718±0.014**

3) Results on NUS-WIDE: The mAP scores of our DMTL and the compared methods on NUS-WIDE [43] are shown in Table IV. From the results, we find that our method still outperforms the peer methods under two settings. The proposed transfer learning strategy can effectively transfer the knowledge from the labeled categories to the unlabeled categories, even they have disjoint label sets, i.e., a large semantic gap in the two domains. The NUS-WIDE data set is a large-scale data set, which provides a large number of training samples for the tested methods. Therefore, they all obtain much higher mAP scores on NUS-WIDE than the results on Pascal Sentences and Wikipedia.

4) Results on XMediaNet: The results on the XMediaNet data set [2] are reported in Table V. They are consistent with the results on the other three data sets that our method performs better than all the peer methods with a large margin, especially under the second setting. Our method achieves an improvement of 57.9% average mAP scores over the second best method when both the labeled source data and the unlabeled target data are available for training. This illustrates the effectiveness of our proposed multimodal transfer learning method. Also, it improves the average mAP score from 10.6% (by using only the source data) to 71.8% (by using both the labeled source data and unlabeled target data), which means that the unlabeled data of the target domain is significant for our method to achieve high performance on CMR. At last, we can see that the mAP scores of the other CMR on this data set are much lower than the other three data sets. The potential reason is the number of categories is up to 200, which makes the recognition of different categories more challenging. However, our methods transfer the knowledge from a large number of labeled data and conquer the domain-shit problem, thus obtaining a much higher mAP score.

Overall, from the above analysis, we find that our DMTL outperforms other peer methods under the two settings, which is consistent with the results of the PR curves on the four data sets from one simulation, as shown in Fig. 3. From the results, we can see that our method outperforms all other methods under all thresholds of the decision function on the four data sets. In addition, our method can achieve a recall of 80% with the corresponding precision of > 75% on NUS-WIDE. It validates the effectiveness of our DMTL. The supervised methods result in relatively low mAP scores. The potential reason is that they only learn on the labeled categories, but retrieval on the unlabeled categories. The labeled known categories and the unlabeled new categories have different data distribution and disjoint label sets. The supervised methods lack the knowledge about the unlabeled new categories and suffer from the domain shift problem. This is also verified by

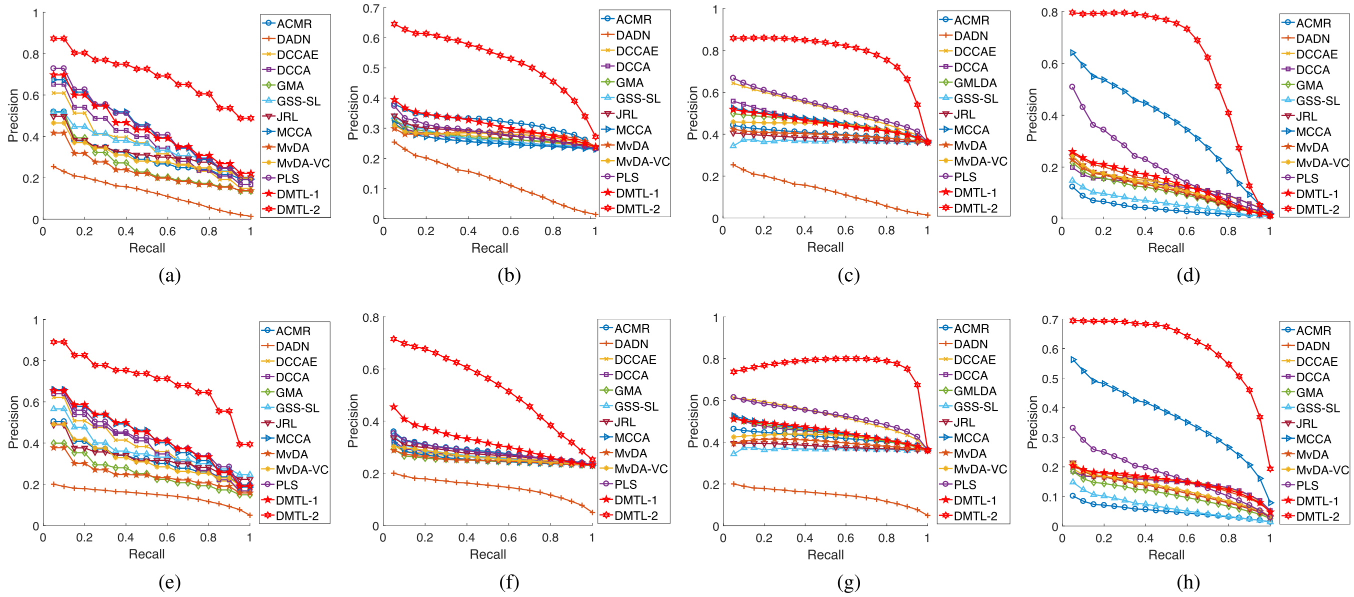


Fig. 3. PR curves of the image-query-texts (image→text) and text-query-images (text→image) on Wikipedia, Pascal Sentences, NUS-WIDE, and XMediaNet. DMTL-1 and DMTL-2 denote our proposed method under the first setting and the second setting, respectively. (a) Pascal Sentences (image→text). (b) Wikepedia (image→text). (c) NUS-WIDE (image→text). (d) XMediaNet (image→text). (e) Pascal Sentences (text→image). (f) Wikepedia (text→image). (g) NUS-WIDE (text→image). (h) XMediaNet (text→image).

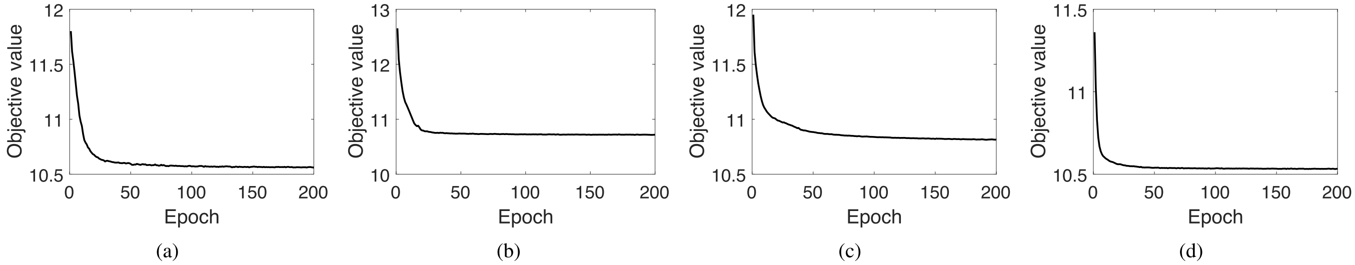


Fig. 4. Value of the objective function of DMTL versus the number of training epochs on the four benchmark data sets. (a) Pascal Sentences. (b) Wikipedia. (c) NUS-WIDE. (d) XMediaNet.

the results of our method on the two different experimental settings, where it uses the source data only in the first setting and uses both the source data and the target data in the second setting. The potential reason why the semisupervised method of GSS-SL obtains such a low mAP score is also due to that the source domain and the target domain have different categories. The classifier may enforce the whole network to learn more discriminative features for the source domain instead of that for the target domain. Specifically, GSS-SL constructs a label graph constraint to ensure the intrinsic geometric structures of different feature spaces consistent with that of label space. Besides, GSS-SL proposes to classify each of the unlabeled samples into one of the labeled categories, i.e., it adopts a hard-label strategy. However, the relationships of the categories in the target domain may significantly different from that of the categories in the source domain, leading GSS-SL to a domain-shit problem as well.

E. Further Analysis of DMTL

In this section, we investigate more details about the proposed method, including its convergency, the visualization of representations in the common shared space, the ablation study, and the parameter analysis. In the following experiments, we use both the labeled data in the source domain and the unlabeled data in the target domain for training.

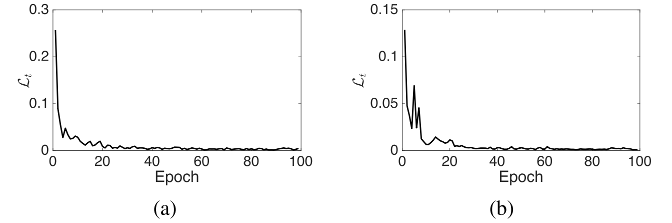
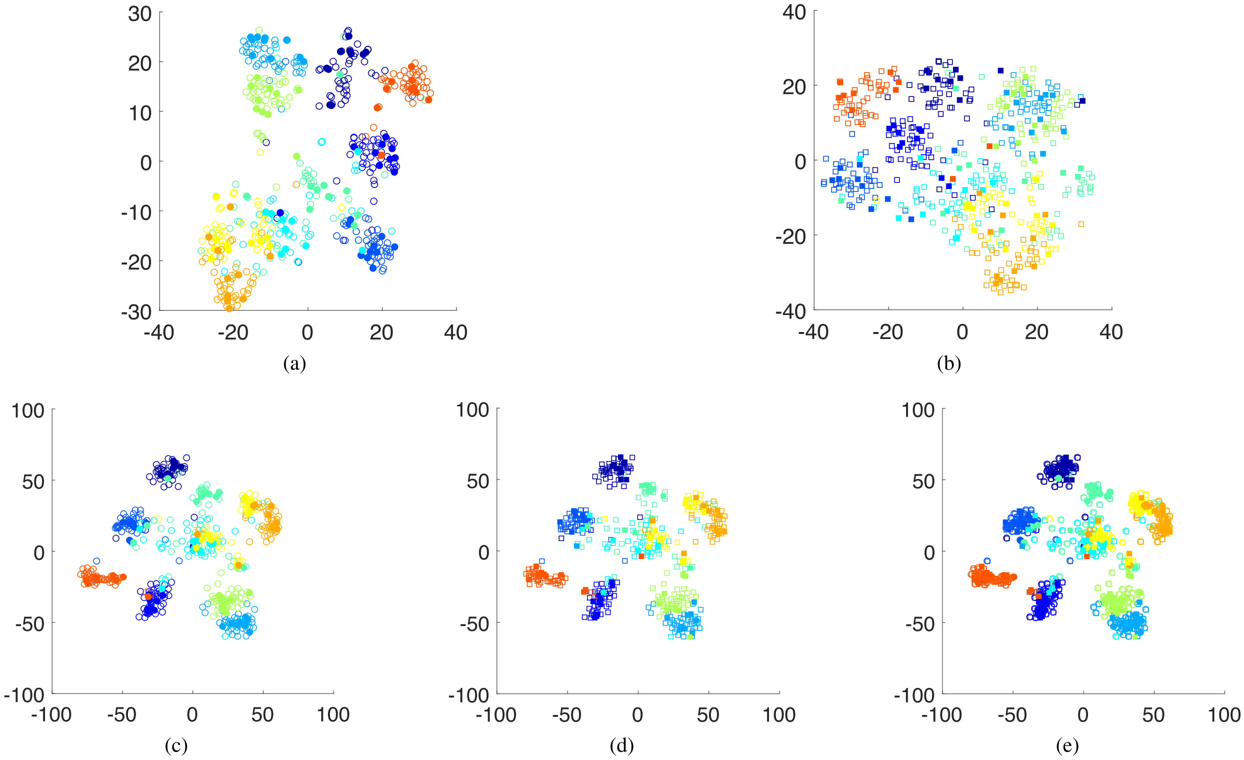


Fig. 5. Value of the loss term $\mathcal{L}_t$ of DMTL versus the number of training epochs on (a) Pascal Sentences and (b) Wikipedia.

1) Convergency: Fig. 4 shows the objective function of our method versus the different number of training epochs on Pascal Sentences, Wikipedia, NUS-WIDE, and XMediaNet. From the results, we find that during the entire training procedure, the value of the objective function decreases almost monotonously and converges smoothly on the four data sets. The value of the objective function becomes stable after 50 epochs, which indicates that DMTL can be efficiently trained by adopting the stochastic gradient descent optimization algorithm Adam [54].



(a)

(b)

(c)

(d)

(e)

Fig. 6. Visualization for the new category data from the Pascal Sentences data set by using the t-SNE method [58]. The samples that come from the same semantic category are marked with the same color. Different shapes denote different modalities, and the filled shapes represent the samples in the training set and the unfilled shapes represent the samples in the test set. (a) Original image samples represented by the VGGNet [48] features. (b) Original text samples represented by the Doc2Vec [49] features. (c) Image representations in the common representation space. (d) Text representations in the common representation space. (e) Image and text representations in the common space.

Also, we investigate the convergence of the loss term $\mathcal{L}_t$ on the pseudolabels of the data in the target domain on Pascal Sentences and Wikipedia, as shown in Fig. 5. It shows that the loss decreases along with the increase of the training epoch and becomes stable after 40 epochs and almost equals zero.

2) Visualization of the Learned Representations: To visually investigate the effectiveness of DMTL, we adopt the t-SNE approach [58] to embed the representations of the image and text samples (in the common space) into a 2-D plane. The results of the original images represented by the 4096-D (VGGNet [48]) features and the text samples represented by the 300-D (Doc2Vec [49]) features (after the embedding process) are shown in Fig. 6(a) and (b), respectively. We can see that the distributions of the image modality and the text modality in the Pascal Sentences data set are largely different and the interclass samples from both the image modality and the text modality are hard to be distinguished in the original input space. Fig. 6(c) and (d) shows the 2-D distributions of the image and text representations in the common space. From the results, we can see that our proposed method can model the discrimination between the samples from different semantic categories. It effectively separates the representations into several semantically discriminative clusters. We can also find that a small number of the representations of different semantic categories are mixed together, which makes DMTL prone to return some irrelevant results for a query. These results are in accordance with the retrieval results shown in Table II. Furthermore, the distributions of image modality and text modality in Fig. 6(e) are well mixed together and are difficult to be separated from each other. It means that the cross-modal discrepancy is largely reduced by using the proposed method. At last, for each category (marked with different colors), most of the filled shapes and the unfilled shapes are overlapped. It means that DMTL has effectively connected the test data set with the training data set for the new categories, which can be helpful to achieve high CMR performance on the target data set of new categories.

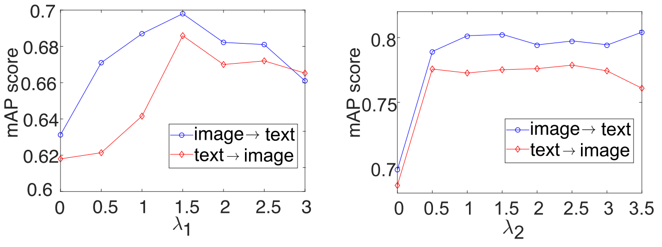
3) Impact of Different Components: The objective function of our DMTL consists of three terms: the discrimination loss for the source data set $\mathcal{L}_s$, the discrimination loss for the target data set $\mathcal{L}_t$, and the domain-invariance loss $\mathcal{L}_m$. To investigate the impacts of these terms on the performance of DMTL, we develop and evaluate its three variations: DMTL with $\mathcal{L}_m$ only, DMTL with $\mathcal{L}_s$ only, and DMTL without $\mathcal{L}_t$. Table VI reports the performance comparisons of DMTL and its three variations on the four data sets. From the results, we have the following observations.

1) The full DMTL outperforms the variant of DMTL (without $\mathcal{L}_t$), which demonstrates the advantages of using pseudolabels to feedback information to learn discriminative features for the new categories. The $\mathcal{L}_t$ term achieves an improvement of 9.7% on NUS-WIDE in terms of the average mAP score.
2) The variant of DMTL (with $\mathcal{L}_m$ only) can achieve promising results, which demonstrates the effectiveness of our proposed distribution approximation strategy.
3) The variant of DMTL (with $\mathcal{L}_s$ only) results in the lowest mAP scores on the four data sets. The potential reason

TABLE VI

PERFORMANCE COMPARISONS OF THE PROPOSED DMTL AND ITS THREE VARIATIONS IN TERMS OF MAP SCORES

Dataset	Method	image→text	text→image	Average
Pascal Sentences	DMTL (with $\mathcal{L}_m$ only)	0.646	0.630	0.638
	DMTL (with $\mathcal{L}_s$ only)	0.296	0.298	0.297
	DMTL (without $\mathcal{L}_t$)	0.653	0.628	0.641
	DMTL	**0.669**	**0.652**	**0.660**
Wikipedia	DMTL (with $\mathcal{L}_m$ only)	0.473	0.491	0.482
	DMTL (with $\mathcal{L}_s$ only)	0.306	0.298	0.302
	DMTL (without $\mathcal{L}_t$)	0.487	0.537	0.512
	DMTL	**0.499**	**0.557**	**0.528**
NUS-WIDE	DMTL (with $\mathcal{L}_m$ only)	0.616	0.610	0.613
	DMTL (with $\mathcal{L}_s$ only)	0.451	0.443	0.447
	DMTL (without $\mathcal{L}_t$)	0.698	0.686	0.692
	DMTL	**0.802**	**0.775**	**0.789**
XMediaNet	DMTL (with $\mathcal{L}_m$ only)	0.749	0.664	0.706
	DMTL (with $\mathcal{L}_s$ only)	0.116	0.125	0.121
	DMTL (without $\mathcal{L}_t$)	0.754	0.665	0.709
	DMTL	**0.754**	**0.665**	**0.710**



(a)

(b)

Fig. 7. mAP scores versus hyperparameters on the NUS-WIDE data set. (a) mAP versus λ_1 when $\lambda_2 = 0$. (b) mAP versus λ_2 when $\lambda_1 = 1.5$.

is that it suffers the domain shift problem. However, the variant of DMTL (without $\mathcal{L}_t$) outperforms the variant of DMTL (with $\mathcal{L}_m$ only), which indicates that using the labeled known category data can improve the retrieval performance on the new category data set.

The above analysis indicates that all of the three terms in the objective function contribute to the final accuracy.

4) Parameter Sensitivity Analysis: In this section, we investigate the parameter sensitivity of our proposed method. There are two hyperparameters λ_1 and λ_2. We conduct the analysis by varying the value of one parameter while fixing the value of another parameter. Specifically, we first set the value of λ_2 as zero to search the optimal value of λ_1 and then fix λ_1 and search the optimal value of λ_2. The results on the NUS-WIDE data set are shown in Fig. 7, from which we can see that DMTL obtains the highest mAP score under $\lambda_1 = 1.5$ when fixing the value of λ_2 as zero, and then, its mAP scores dropdown along with the increase of the value of λ_1. In addition, if we fix λ_1 as 1.5, the average mAP scores of DMTL increase first and then reduce slightly along with the growth of the value of λ_2. The mAP scores of DMTL are higher than 65% in a large range of values for λ_1 and λ_2 and that DMTL is robust against the two hyperparameters.

V. CONCLUSION

In this article, we proposed a novel approach (DMTL) to achieve the CMR on the new categories by transferring the knowledge from the known categories. To overcome the domain shift problem, we designed a joint learning paradigm to exploit the annotated labels and the pseudolabels to learn discriminative features for the target data set of new categories. Following the proposed learning paradigm, we optimize the modality-specific networks and update the pseudolabels alternatively to achieve the multimodal transfer learning. Furthermore, to bridge the heterogeneity gap between different modalities, we designed two coupled networks to learn two conditional distribution spaces to match the ideal distribution spaces induced by the available data set of two different modalities. It can preserve the latent structure of the data set and achieve better performance. Extensive experimental results and the comprehensive analysis have validated the effectiveness of the proposed joint learning strategy and the designed distribution approximation networks. Our method achieves the promising CMR performance on the new categories compared with state-of-the-art methods.

Despite its high competitiveness, our method faces the following two limitations: 1) it cannot handle increasing categories efficiently and 2) it assumes that the pairwise information between different modalities is available, which is not always able to meet in some real-world applications. In this regard, investigating how to leverage incremental learning and unpaired cross-modal learning to address these two issues is the focus of our subsequent study.

REFERENCES

[1] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, “A comprehensive survey on cross-modal retrieval,” *CoRR*, vol. abs/1607.06215, pp. 1–20, 2016. [Online]. Available: http://arxiv.org/abs/1607.06215

[2] Y. Peng, X. Huang, and Y. Zhao, “An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[3] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, “From deterministic to generative: Multimodal stochastic RNNs for video captioning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.

[4] L. Gao, X. Li, J. Song, and H. T. Shen, “Hierarchical LSTMs with adaptive attention for visual captioning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, May 2020.

[5] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, “Adversarial cross-modal retrieval,” in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[6] L. Zhen, P. Hu, X. Wang, and D. Peng, “Deep supervised cross-modal retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.

[7] P. Hu, L. Zhen, D. Peng, and P. Liu, “Scalable deep multimodal learning for cross-modal retrieval,” in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 635–644.

[8] Q.-Y. Jiang and W.-J. Li, “Deep cross-modal hashing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3270–3278.

[9] X. J. Zhu, “Semi-supervised learning literature survey,” Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. TR 1530, 2005.

[10] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, “Semi-supervised cross-media feature learning with unified patch graph regularization,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.

[11] J. Wang, G. Li, P. Pan, and X. Zhao, “Semi-supervised semantic factorization hashing for fast cross-modal retrieval,” *Multimedia Tools Appl.*, vol. 76, no. 19, pp. 20197–20215, Oct. 2017.

[12] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, “Generalized semi-supervised and structured subspace learning for cross-modal retrieval,” *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, Jan. 2018.

[13] Z. Ding, N. M. Nasrabadi, and Y. Fu, “Semi-supervised deep domain adaptation via coupled neural networks,” *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5214–5224, Nov. 2018.

[14] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 529–536.

[15] F. Zhong, Z. Chen, and G. Min, "An exploration of cross-modal retrieval for unseen concepts," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, 2019, pp. 20–35.

[16] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. 11th Annu. Conf. Comput. Learn. Theory (COLT)*, 1998, pp. 92–100.

[17] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.

[18] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 686–701.

[19] C. Li, C. Zhu, Y. Huang, J. Tang, and L. Wang, "Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 808–823.

[20] P. Hu, X. Wang, L. Zhen, and D. Peng, "Separated variational hashing networks for cross-modal retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1721–1729.

[21] P. Hu, D. Peng, Y. Sang, and Y. Xiang, "Multi-view linear discriminant analysis network," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5352–5365, Nov. 2019.

[22] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, Dec. 1936.

[23] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[24] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[25] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 188–194, Jan. 2016.

[26] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image-text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 11, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8994196, doi: 10.1109/TNNLS.2020.2967597.

[27] H. T. Shen *et al.*, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 29, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8974240, doi: 10.1109/TKDE.2020.2970050.

[28] 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.

[29] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 759–766.

[30] T. Scott, K. Ridgeway, and M. C. Mozer, "Adapted deep embeddings: A synthesis of methods for k-shot inductive transfer learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 76–85.

[31] Y. He, J. Yuan, and L. Li, "Enhancing RNN based OCR by transductive transfer learning from text to images," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–2.

[32] Z. Ding, S. Li, M. Shao, and Y. Fu, "Graph adaptive knowledge transfer for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 37–52.

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

[34] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7167–7176.

[35] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8156–8164.

[36] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

[37] D. Zhang, J. He, Y. Liu, L. Si, and R. Lawrence, "Multi-view transfer learning with a large margin approach," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2011, pp. 1208–1216.

[38] P. Yang and W. Gao, "Multi-view discriminant transfer learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1848–1854.

[39] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.

[40] J. Chi and Y. Peng, "Zero-shot cross-media embedding learning with dual adversarial distribution network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1173–1187, Apr. 2020.

[41] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical Turk," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data With Amazon's Mech. Turk.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 139–147.

[42] J. C. Pereira *et al.*, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.

[43] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from national University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr. (CIVR)*, 2009, p. 48.

[44] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1070–1076.

[45] J. Chi and Y. Peng, "Dual adversarial networks for zero-shot cross-media retrieval," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 663–669.

[46] M. Rohrbach, S. Ebert, and B. Schiele, "Transfer learning in a transductive setting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 46–54.

[47] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1641–1648.

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–14, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[49] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," in *Proc. RepL4NLP*, 2016, pp. 78–86.

[50] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning—The good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4582–4591.

[51] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.

[52] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 233–240.

[53] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.* Madison, WI, USA: Omnipress, 2010, pp. 807–814.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, pp. 1–15, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[55] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.

[56] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Proc. Conf. Data Mining Data Warehouses*, 2010, pp. 1–4.

[57] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 593–600.

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