MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Incomplete Multi-view Domain Adaptation via Channel Enhancement and Knowledge Transfer

Xia, Haifeng; Wang, Pu; Ding, Zhengming

TR2022-134 October 25, 2022

Abstract

Unsupervised domain adaptation (UDA) borrows well-labeled source knowledge to solve the specific task on unlabeled target domain with the assumption that both domains are from a single sensor, e.g., RGB or depth images. To boost model performance, multiple sensors are deployed on new-produced devices like autonomous vehicles to benefit from enriched information. However, the model trained with multi- view data difficultly becomes compatible with conventional devices only with a single sensor. This scenario is defined as incomplete multi-view domain adaptation (IMVDA), which considers that the source domain consists of multi-view data while the target domain only includes single- view instances. To overcome this practical demand, this paper proposes a novel Channel Enhancement and Knowledge Transfer (CEKT) frame- work with two modules. Concretely, the source channel enhancement module distinguishes view-common from view-specific channels and explores channel similarity to magnify the representation of important channels. Moreover, the adaptive knowledge transfer module attempts to enhance target representation towards multi-view semantic through implicit missing view recovery and adaptive cross-domain alignment. Extensive experimental results illustrate the effectiveness of our method in solving the IMVDA challenge.

European Conference on Computer Vision (ECCV) 2022

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Incomplete Multi-view Domain Adaptation via Channel Enhancement and Knowledge Transfer

Haifeng Xia[1], Pu Wang[2], and Zhengming Ding[1]

[1] Department of Computer Science, Tulane University, USA
[2] Mitsubishi Electric Research Laboratories, USA
{hxia, zding1}@tulane.edu, pwany@merl.com

Abstract. Unsupervised domain adaptation (UDA) borrows well-labeled source knowledge to solve the specific task on unlabeled target domain with the assumption that both domains are from a single sensor, e.g., RGB or depth images. To boost model performance, multiple sensors are deployed on new-produced devices like autonomous vehicles to benefit from enriched information. However, the model trained with multi-view data difficultly becomes compatible with conventional devices only with a single sensor. This scenario is defined as incomplete multi-view domain adaptation (**IMVDA**), which considers that the source domain consists of multi-view data while the target domain only includes single-view instances. To overcome this practical demand, this paper proposes a novel Channel Enhancement and Knowledge Transfer (**CEKT**) framework with two modules. Concretely, the source channel enhancement module distinguishes view-common from view-specific channels and explores channel similarity to magnify the representation of important channels. Moreover, the adaptive knowledge transfer module attempts to enhance target representation towards multi-view semantic through implicit missing view recovery and adaptive cross-domain alignment. Extensive experimental results illustrate the effectiveness of our method in solving the IMVDA challenge.

Keywords: Multi-View Fusion, Domain Adaptation

1 Introduction

Deep neural network (DNN) recently becomes the dominate technique in computer vision community due to its success on the real-world applications such as image classification [49, 49, 20], object detection [38] and image segmentation [13, 33]. As a data-driven learning strategy, DNN generally requires considerable training samples with high-quality annotations to capture the intrinsic semantic knowledge. However, the data collection and manual annotation tend to be expensive and time-consuming [4, 46, 19]. To benefit from external resources, recent solutions pay more attentions to transfer learning, especially for unsupervised domain adaptation (UDA) [21, 36, 2, 6].

UDA aims to transfer well-supervised source knowledge to assist the specific tasks in target domain without any annotation information [29, 44]. However,

data collection typically occurring in varying environments easily triggers the significant distribution discrepancy across source and target samples [12, 45]. The main challenge of UDA is how to learn domain-invariant feature representations. Along with this direction, the UDA algorithms mainly explore metric-based scheme and adversarial training fashion. Specifically, one of the classical and effective metric-based strategies transforms target samples into source latent space and explore their sample-wise association to eliminate domain mismatch [1]. However, the alignment method needs to observe all data to accurately estimate the relation of source and target instances, which difficultly adjusts to the mini-batch training manner in DNN. In addition, the basic UDA setting considers that the images of source and target domain are merely captured by one sensor. But the practical application always deploys multiple sensors such as the autonomous vehicle to obtain more sufficient information to boost the model performance.

A few efforts [5, 16] have explored multi-view domain adaptation (MVDA), where source and target data are both collected from multiple sensors. The intuitive idea is to convert MVDA into a UDA problem by independently aligning source and target instances within each view and fusing multi-view semantic information within individual domain. They have achieved promising performance on solving MVDA and abundant empirical studies illustrate that the simple alignment-and-fusion promotes model performance on identifying target samples with more enriched data collected by multiple sensors. However, equipment rehabilitation to upgrade previous single-sensor devices with multiple sensors causes additional cost overhead, which makes MVDA to be invalid for several practical application scenarios. Instead, we post a question that *"Can we develop more effective domain adaptation algorithms to benefit single-sensor target data from enriched source data with multiple sensors?"*. This problem is defined as incomplete multi-view domain adaptation (**IMVDA**), where there are multi-view complete data in source domain, while single-view instances in target domain. This problem is under insufficient exploration in the literature.

To overcome IMVDA challenge, we propose a novel method named Channel Enhancement and Knowledge Transfer (**CEKT**) shown in Figure 1, which not only conducts multi-view semantic fusion within source domain but also transfers the integrated knowledge for the use of target domain. Concretely, CEKT explores the sparse attribution of channel to distinguish view-common from view-specific feature maps and exchanges view-specific channels across multiple views to fuse their semantic information. Furthermore, we develop a metric of channel similarity to highlight the representation of significant channels, which assists model learning with more discriminative features. Moreover, we introduce a parallel target model taking source and target samples from the same view as input. The source model trained in the first step teaches the target model to produce multi-view semantic only with single view data. In addition, we propose a novel adaptive subspace alignment to gradually mitigate domain discrepancy in an end-to-end training manner. To sum up, the main contributions of this work are highlighted in three folds:

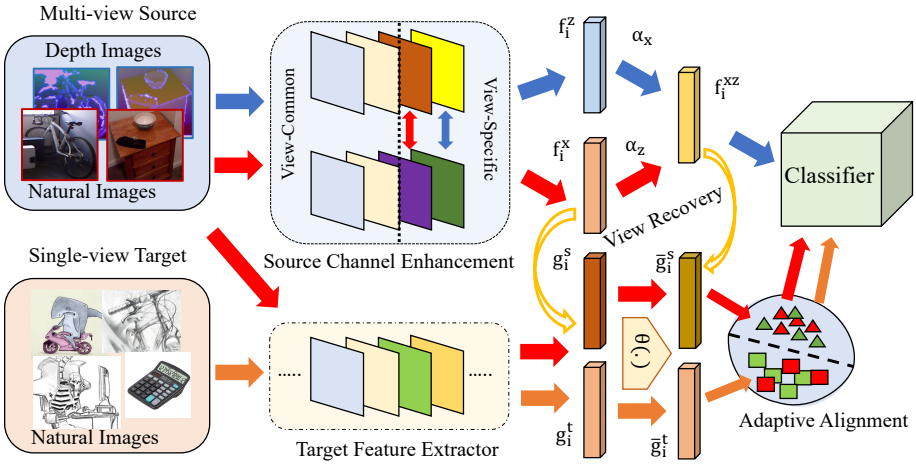

Fig. 1. Overview of our channel enhancement and knowledge transfer framework (CEKT) for incomplete multi-view domain adaptation (IMVDA). Specifically, the source channel enhancement module distinguishes view-common from view-specific channels and explores the channel similarity to emphasis essential representation. The source triggered missing view recovery teaches target model how to generate multi-view knowledge. And the adaptive alignment module aims to eliminate domain mismatch within the identical subspace.

- First, our proposed CEKT introduces a novel channel enhancement mechanism to preserve considerable view-common semantic knowledge and exchange view-specific semantic to enrich the representation of each view. This module not only effectively achieves feature fusion but also emphasizes more discriminative features for the classification task.
- Second, the adaptive knowledge transfer module explores the supervision of source model to supervise the target model to approximate multi-view semantic information, which mitigates the negative influence of missing view on target domain. Simultaneously, we present a novel adaptive subspace alignment method to learn domain-invariant representations.
- Finally, we exploit many public-available real-world image datasets to imitate the IMVDA scenario and conduct abundant experiments to evaluate the performance of our CEKT. The corresponding experimental results and analysis fully demonstrate the effectiveness of our method.

2 Related Work

2.1 Domain Adaptation

Unsupervised domain adaptation (UDA) aims to borrow well-supervised source knowledge to assist the target learning without any label information [6, 44, 46]. And source and target instances belong to different distributions, yet share

the identical label space [34]. The core task of UDA is to learn domain-invariant representations by gradually eliminating distribution mismatch. The mainstream learning mechanisms are considered as two types. One is metric-based alignment [32] which enforces source and target domains to share the identical statistics (e.g. mean and co-variance) and transform target samples into the source subspace via the estimation of cross-domain sample-wise association [1]. Another mature exploration adopts generative adversarial game between feature generator and discrimination to mitigate domain mismatch in latent feature space. In addition, [14] extends the conventional UDA by introducing more source domains to improve model generalization, which is named multi-source domain adaptation (MSDA). Similar with UDA based methods, [37] deploys multiple discriminators for arbitrary one source and target domains to independently achieve distribution alignment. However, the above problems generally assume that the instances per domain are captured with only one sensor, which prevents the development of technique. To improve model performance, abundant devices as autonomous vehicles are installed with multiple sensors to comprehensively perceive the open world. Thus, this work explores a practical and challenging IMVDA scenario, where source data are collected from multiple sensors while target samples are captured by the single sensor.

2.2 Multi-view Learning

Multi-view learning expects to access sufficient semantics via the joint utilization of multiple data sets [30, 52, 5]. Extensive empirical studies show significant performance improvement on object classification tasks [48, 50] by using multi-view data. The intuitive learning strategy is to discover the consistent hypothesis space across various views [22]. Specifically, [47] adopts a co-regularization manner to compress the search space of hypothesis function. Similarly, [28] presents an efficient dictionary learning and [24] utilizes a large-margin Gaussian process to find the intrinsic basis across multiple views. In addition, the clustering technique is introduced to discover complementary semantic knowledge from different views [43]. And the multi-view spectral embedding is developed to integrate feature representation. Although these multi-view methods produce positive effect given complete views, they assume the multi-view samples are from the identical distribution, which is not the case for the real-world applications. Instead, we not only consider multi-view knowledge fusion but also conduct simple yet effective knowledge transfer across multiple domains to address the IMVDA problem.

3 Proposed Method

3.1 Preliminary & Motivation

Formally for the IMVDA problem, we are given the well-annotated source domain with enriched views[3] as $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{z}_i^s, y_i)\}_{i=1}^{n_s}$ and the unlabeled target

[3] This paper considers the case that the source domain contains two views while target domain includes only single view.

domain with only single view as $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, where $\mathbf{x}$ and $\mathbf{z}$ denote two view-paired samples, y represents the corresponding source label, and n_s and n_t are the number of source and target samples, respectively. The goal of IMVDA is to transfer the enriched view information and well-annotated label information in the source domain to improve the single-view target recognition.

Therefore, two-fold challenges should be considered: 1) How to effectively integrate multi-view semantics to boost performance of model, and 2) How to transfer knowledge from multi-view source domain to single view target one. To address these questions, we propose a novel solution named Channel Enhancement and Knowledge Transfer (CEKT) framework as Figure 1. Concretely, CEKT involves two components, i.e., a source channel enhanced network and an adaptive knowledge transfer network. The former one aims to distinguish view-common channels from view-specific channels where semantic fusion occurs and exploit cross-view channel similarity to enhance the representation of necessary channels. The latter one attempts to adaptively learn a target-to-source projection to mitigate the domain mismatch.

3.2 Source Channel Enhanced Network

Cross-view Channel Enhancement

Batch normalization (BN) [17] is widely used in deep neural networks to scale the hidden features of the specific layer to accelerate convergence and avoid the model collapse as:

$$\hat{h}_c = \gamma_c \frac{(h_c - \mu_c)}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c, \tag{1}$$

where h_c, $\hat{h}_c$ mean the input and output of the BN module, μ_c, σ_c are the mean and variance of the c-th channel, and γ_c, β_c are trainable parameters. However, from the perspective of channel exchange [43], the model training gradually neglects the representation of task-irrelevant channels as $\gamma_c \to 0$, and multi-view data cause the channels ($h_{x,c}$, $h_{z,c}$) from $(\mathbf{x}^s, \mathbf{z}^s)$ to be activated differently. Then, Wang et. al. proposed channel exchange for two views to compensate each other as [43]:

$$\hat{h}_{x/z,c} = \gamma_{z/x,c} \frac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} + \beta_{z/x,c}, \quad \text{if} \quad \gamma_{x/z,c} < \delta, \tag{2}$$

where δ is an adjustable threshold, and a sparse regularization term $\sum_{c=1}^{C} |\gamma_{x/z,c}|$ is introduced to encourage more channel exchanges. Such channel exchange totally relies on the learned $\gamma_{x/z,c}$, which makes channel exchange in an unsupervised fashion without considering sharing channels across views.

Thus, we develop a Cross-view Channel Enhancement (C^2E) module. Specifically, for one concrete layer, all channels are divided into two groups: view-common channels and view-specific ones. Under this condition, we suppose view-common channels tend to involve considerable shared semantics, where the cor-

responding parameters $\gamma_{x/z,c}$ should be compact rather than sparse, and view-specific channels carry the unique information for each view and should be exchanged and enhanced. With this consideration, the ℓ_1-norm over the parameters is a promising manner to highlight the difference across view-specific channels. In implementation, we consider the first half of all feature maps as the view-common channels and the remaining ones as view-specific parts. Thus, we adopt the following constraint for parameters $\gamma_{x/z}$ as:

$$\min_{\gamma_{l,c}} \mathcal{L}_\gamma = \sum_{l=1}^{L} \left(\sum_{c=1}^{\lfloor C/2 \rfloor} \gamma_{l,c}^2 + \sum_{c=\lfloor C/2 \rfloor}^{C} |\gamma_{l,c}| \right), \tag{3}$$

where we omit the superscript (x, z) for convenience, C and $\lfloor C/2 \rfloor$ mean the number of channel and the rounding or flooring operation, and L is the number of network layers attached with the BN module. It is worth noting that only the view-specific channels participate in the channel exchange via Eq. (2). Through the above strategy, we not only achieve feature fusion but also preserve as much view-common semantics as possible. Hence, $\gamma_{x/z,c} \geq \delta$ illustrates that this channel can contribute to the classification task.

To further enhance the channels shared across views, we propose a strategy to identify those channels and amplify their presence during batch normalization. As two views data present the identical content in various forms, their representations to the necessary information such as the contour of object tend to be similar or even consistent. In other words, the c-th channel with a high similarity across two views should be considered as an important component with a high confidence. Thus, the similarity (s_c) of two views at channel c is defined as:

$$s_c = \frac{\exp(-\|\mu_{x,c} - \mu_{z,c}\|_2/\eta)}{\sum_{c=1}^{C} \exp(-\|\mu_{x,c} - \mu_{z,c}\|_2/\eta)}, \tag{4}$$

where $\sum_c s_c = 1$ and η controls the change of scale. Then, we first adjust the importance of channel with $\hat{h}_{x/z,c} = (1 + s_c)\hat{h}_{x/z,c}$ before the channel exchange in Eq. (2). For instance, when the two channels are very different, corresponding s_c plays a small fraction of the similarity vector and, hence, the importance of the c-th channel is not augmented with a relatively small s_c.

Data-dependant Cross-view Fusion

For now, our module is easily applied into most deep neural network $\mathcal{F}(\cdot)$ mapping the original image into the high-level features $\mathbf{f}_x = \mathcal{F}(\mathbf{x})$ or $\mathbf{f}_z = \mathcal{F}(\mathbf{z})$. To further learn robust features, we adopt a data-dependant fusion manner to obtain these high-level representations as:

$$\mathbf{f}_{xz} = \alpha_x \mathcal{F}(\mathbf{x}) + \alpha_z \mathcal{F}(\mathbf{z}), \tag{5}$$

where $\alpha_{x/z}$ are the probability score for two views and we plug in the softmax layer $\mathbf{s}(\cdot)$ to $\mathcal{F}(\mathbf{x})$ and $\mathcal{F}(\mathbf{z})$ to learn the data-dependant fusion weights.

Finally, the multi-class source classifier $\mathcal{C}(\cdot)$ takes the fused features as input to generate the prediction. The objective function for training the source model is formulated as:

$$\min_{\mathcal{F},\mathcal{C},\mathbf{s},\gamma} \mathcal{L}_s = \sum_{i=1}^{n_s} \mathcal{L}_c\Big(\mathcal{C}(\mathbf{f}_{xz}^i), y_i\Big) + \lambda_\gamma \mathcal{L}_\gamma, \tag{6}$$

where λ_γ is a trade-off parameter and $\mathcal{L}_c(\cdot,\cdot)$ is the classical cross-entropy loss.

3.3 Adaptive Knowledge Transfer Network

The target domain lacks one view and exists considerable distribution difference with source domain, which makes it unreasonable to directly identify target samples with multi-view source model. Thus, the current challenge is how to effectively transfer source fused knowledge to the target domain. Along with this direction, we construct a novel adaptive knowledge transfer network (AKT), whose core is to associate two domains with source view data x_i^s as the bridge. Concretely, we introduce an additional target network $\mathcal{G}(\cdot)$ with the same network architecture to source and the conventional BN module.

Source Triggered Missing View Recovery. To guide the target network with the ability for missing view, we allow source sample $\mathbf{x}_i^s$ and target sample $\mathbf{x}_j^t$ to pass through the target network $\mathcal{G}(\cdot)$ so that we can obtain the high-level features, i.e., $\mathbf{g}_i^s = \mathcal{G}(\mathbf{x}_i^s)$ and $\mathbf{g}_j^t = \mathcal{G}(\mathbf{x}_j^t)$. Following that, we deploy one dimensionality-identical full-connection layer with trainable parameter θ to obtain $\bar{\mathbf{g}}_i^s$ and $\bar{\mathbf{g}}_j^t$, which aims to recover the missing view information for the target network by mapping one view to two-view fused representation.

Since the target model does not directly touch $\mathbf{z}_i^s$, we expect to learn the fused semantic only with one source view data. As DNN manifests strong approximation capability by using the convolution layers and non-linear mapping [8], it fits better to the given target. Inspired by this observation, when accessing the fused representations with fixed source model, we make $\mathbf{g}_i^s$ and $\bar{\mathbf{g}}_i^s$ approximate $\mathbf{f}_x^i$ and $\mathbf{f}_{xz}^i$, respectively, to mimic the fused semantic features. Hence, we propose a source triggered missing view recovery term as:

$$\min_{\mathcal{G},\theta} \mathcal{L}_g = \sum_{i=1}^{n_s} \Big(\|\mathbf{g}_i^s - \mathbf{f}_x^i\|_2^2 + \|\bar{\mathbf{g}}_i^s - \mathbf{f}_{xz}^i\|_2^2 \Big). \tag{7}$$

In this way, the source model teaches the target one to offset the absence of the other view. Moreover, as $\mathbf{x}^s$ and $\mathbf{x}^t$ belong to the same view, the imitative manner brings semantics of the other view to feature learning of target samples. Certainly, the significant domain shift across $\mathbf{x}^s$ and $\mathbf{x}^t$ obstructs the delivery of additional semantic knowledge to the target domain. Thus, the target model needs to achieve distribution alignment by gradually eliminating the cross-domain discrepancy.

Adaptive Cross-Domain Alignment. The direct alignment approach is first to transform all source and target instances into the shared latent space and then

to reduce the sample-wise distance with the manifold theory. The formulation of this classical strategy [1] is:

$$\min_{\mathbf{A}^{st}} \|\bar{\mathbf{G}}^s - \mathbf{A}^{st}\bar{\mathbf{G}}^t\|_{\mathrm{F}}^2 + \Omega(\mathbf{A}^{st}), \tag{8}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm, $\bar{\mathbf{G}}^{s/t}$ is the feature matrix of all samples $\bar{\mathbf{g}}_i^{s/t}$, and $\mathbf{A}^{st}$ is defined as the transformation matrix mapping target features into the source feature subspace, and $\Omega(\mathbf{A}^{st})$ denotes a regularization term over $\mathbf{A}^{st}$ such as the ℓ_2-norm or ℓ_1-norm. This strategy achieves promising performance on domain adaptation with shallow feature extractors [9]. However, the feature transformation requires simultaneous access to all samples, which the mini-batch training mechanism used in DNN hardly satisfies. Meanwhile, a direct computation of $\mathbf{A}^{st}$ within each mini-batch is unreasonable since the insufficient samples fail to accurately capture the association of samples. To break the bottleneck, we present an adaptive alignment solution involving two fully connected layers without bias terms. The features $\bar{\mathbf{g}}_i^s$ and $\bar{\mathbf{g}}_j^t$ are fed into it to calculate the similarity $\mathbf{A}_{ij}^{st}$ via:

$$\mathbf{A}_{ij}^{st} = \delta\Big(\langle W_s\bar{\mathbf{g}}_i^s, W_t\bar{\mathbf{g}}_j^t\rangle\Big), \tag{9}$$

where $W_{s/t}$ is the projection matrix, $\delta(\cdot)$ denotes an activation function such as ReLU, and $\langle\cdot,\cdot\rangle$ denotes the inner product operation. During the update of $W_{s/t}$, the inputs are fixed. As the model training, $W_{s/t}$ gradually learns the intrinsic distribution information of overall dataset and can accurately estimate the sample-wise relationship.

On the other hand, we can access to the category probability of sample with $\mathbf{p}_i^{s/t} = \mathcal{C}(\bar{\mathbf{g}}_i^{s/t})$. As $\mathbf{p}_i^{s/t}$ with more discriminative information can reflect the structural relation of hidden features via $\bar{\mathbf{A}}_{ij}^{st} = \langle\mathbf{p}_i^s, \mathbf{p}_j^t\rangle$, we propose the adaptive cross-domain alignment as:

$$\min_{\mathcal{G},\theta,W_{s/t}} \mathcal{L}_a = \|\bar{\mathbf{G}}^s - \mathbf{A}^{st}\bar{\mathbf{G}}^t\|_{\mathrm{F}}^2 + \|\mathbf{A}^{st} - \bar{\mathbf{A}}^{st}\|_{\ell_1}, \tag{10}$$

where $\|\cdot\|_{\ell_1}$ denotes the ℓ_1-norm. $\mathbf{A}^{st}$ and $\bar{\mathbf{A}}^{st}$ are normalized along the row dimension. According to the guidance of adaptive similarity $\mathbf{A}^{st}$, the source features can be represented by the similar ones in target domain, and Eq. (10) effectively reduces their divergence to mitigate the domain mismatch.

3.4 Overall Objective

We first finalize the objective function for the target model. To preserve abundant source knowledge, we adopt source annotations to supervise the target model training. Similar to [27], the pseudo labels of target samples are explored to make target features more discriminative. Specifically, for each epoch, the predictions of target samples (y_j^t) with the fixed target model are used to calculate the class centers, $\mathcal{O}_k = \frac{1}{n_k}\sum_{j=1}^{n_t}\mathbb{I}(y_j^t = k)\bar{\mathbf{g}}_j^t$, where n_k is the number of target samples

from the k-th class and $\mathbb{I}(\cdot)$ is the indicator function. With the class centers, the K-means clustering is adopted to reassign the optimized labels $\hat{y}_j^t$ to target samples. The loss function to the target model is defined as:

$$\min_{\mathcal{G},\theta,\mathcal{C},W_{s/t}} \mathcal{L}_t = \mathcal{L}_c^s + \lambda_g \mathcal{L}_g + \lambda_\tau(\mathcal{L}_a + \mathcal{L}_c^t), \tag{11}$$

where $\mathcal{L}_c^s$ denotes source supervision loss as $\sum_{i=1}^{n_s} \mathcal{L}_c(\mathcal{C}(\bar{\mathbf{g}}_i^s), y_i^s)$, $\mathcal{L}_c^t$ denotes the pseudo target supervision loss as $\sum_{j=1}^{n_t} \mathcal{L}_c(\mathcal{C}(\bar{\mathbf{g}}_i^t), \hat{y}_j^t)$, and λ_g, λ_τ are trade-off parameters. To avoid the negative effect in the beginning, we define λ_τ as $\frac{1-\exp(-10\tau)}{1+\exp(-10\tau)}$ with the changing of epoch number (τ).

Then, for the overall training strategy, we adopt an iterative training manner to optimize both source and target networks. Concretely, Eq. (6) is used to optimize the parameters of source model with the fixed target network $\mathcal{G}(\cdot)$ and then we update target model via Eq. (11) with the frozen source network $\mathcal{F}(\cdot)$.

3.5 Theoretical Analysis

In Eq. (3), we adopt two different constraints on the scaling factors $\gamma_{l,c}$, which enable the network to **actively** learn view-specific and view-common knowledge in various channels, respectively. Similar with [42], we deduce the following theorem to explain why the $\sum_{c=\lfloor C/2 \rfloor}^{C} |\gamma_{l,c}|$ can assist the model to capture view-specific information and the function of $\sum_{c=1}^{\lfloor C/2 \rfloor} \gamma_{l,c}^2$.

Theorem 1. *The proposed $\sum_{c=\lfloor C/2 \rfloor}^{C} |\gamma_{l,c}|$ will definitely make the corresponding scaling factors towards zero with the probability $2\Phi\big(\lambda_\gamma(\frac{\partial \mathcal{L}_c}{\partial h_c})^{-1}\big) - 1$, where the $\Phi(\cdot)$ denotes the cumulative probability of standard Gaussian. To be simple, the subscript l of $\gamma_{l,c}$ is mitigated.*

Proof. According to Eq. (6), it is straightforward to deduce the derivative of $\mathcal{L}_s$ with respect to $\gamma_c, c \in [C/2, C]$ as the following:

$$\frac{\partial \mathcal{L}_s}{\partial \gamma_c} = \begin{cases} \dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c} \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} + \lambda_\gamma \dfrac{\partial \mathcal{L}_\gamma}{\partial \gamma_c}, & \gamma_c > 0 \\ \dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c} \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} - \lambda_\gamma \dfrac{\partial \mathcal{L}_\gamma}{\partial \gamma_c}, & \gamma_c < 0 \end{cases} \tag{12}$$

When the model training approaches convergence, the derivative of $\mathcal{L}_c$ w.r.t $\hat{h}_c$ approximates zero. Due to $\lambda\gamma > 0$, we easily achieve the following inequality:

$$\begin{cases} \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} > -\lambda_\gamma (\dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1}, & \gamma_c > 0 \\ \dfrac{(h_{z/x,c} - \mu_{z/x,c})}{\sqrt{\sigma_{z/x,c}^2 + \epsilon}} < \lambda_\gamma (\dfrac{\partial \mathcal{L}_c}{\partial \hat{h}_c})^{-1}, & \gamma_c < 0 \end{cases} \tag{13}$$

With the central limit theorem, we can convert the above inequality into the probability formulation:

$$\mathbb{P}\Big(-\lambda_\gamma(\frac{\partial\mathcal{L}_c}{\partial\hat{h}_c})^{-1} < \frac{(h_{z/x,c}-\mu_{z/x,c})}{\sqrt{\sigma^2_{z/x,c}+\epsilon}} < \lambda_\gamma(\frac{\partial\mathcal{L}_c}{\partial\hat{h}_c})^{-1}\Big) = 2\Phi\big(\lambda_\gamma(\frac{\partial\mathcal{L}_c}{\partial h_c})^{-1}\big) - 1. \tag{14}$$

The model convergence means $\frac{\partial\mathcal{L}_c}{\partial\hat{h}_c} \to 0$ so that the above probability approximates one. It suggests the scaling factors to these channels will become zero with high-probability. Multi-view images are likely to activate different channels in this part for the classification task. Thus, we consider these channel information as view-specific content. Inversely, benefit from the ℓ_2-norm analysis [51], the $\gamma_c, c \in [1, C/2)$ will be dense non-zero values with the constraint $\sum_{c=1}^{\lfloor C/2\rfloor}\gamma^2_{l,c}$. These channels across various views are both activated to learn semantic from the identical location of images or feature maps and tend to include the similar even consistent patterns, which are defined as view-common channels.

4 Experiments

4.1 Experimental Details

♢ **Datasets:** i). **RGB-D** dataset [23] is a large-scale household objects dataset including 51 categories and each specific object is captured by Kinect style 3D camera (30Hz) generating RGB and depth images at the same time. ii). **B3DO** [18] is a popular 3D benchmark database with RGB and depth image pairs from 83 object categories. And these images are collected from real domestic and office-environments by Microsoft Kinect sensor. iii). **Office-31** [39] is a standard multi-domain RGB image benchmark including Amazon (**A**), Webcam (**W**) and DSLR (**D**), which are gathered with different cameras. And all domains share the identical label space with 31 categories. iv). **Office-Home** [41] as a large-scale cross-domain dataset involves four domains as Art Painting (**Ar**), Clipart (**Cl**), Product (**Pr**) and Real World (**Rw**) with significant image style difference. And each domain includes the same 65 object classes. v). **Caltech-256** (**C**) [11] is a classical natural image database with 30,607 images from 257 objects.

In IMVDA experiments, we consider RGB-D and B3DO as two multi-view (RGB and Depth) well-annotated source domains, while the Caltech-256 or each domain of Office-31 and Office-Home as the unlabeled target domain to mimic the incomplete multi-view scenario. For each specific adaptation task, we select the shared categories across source and target domains. Concretely, the number of categories for tasks **RGB-D→Office31**, **RGB-D→Office-Home**, **RGB-D→Caltech-256** are 8, 13 and 10, respectively, while that for **B3DO→Office31**, **B3DO→Office-Home**, **B3DO→Caltech-256** are 27, 14 and 8, respectively.
♢ **Implementation Details:** The implementation of our model is based on pytorch platform. And we adopt the pre-trained ResNet-50 [15] without the last FC layer as the feature extractor for source and target models, and $W_{s/t} \in \mathbb{R}^{64\times256}$, $\{\mathbf{F}_i^{x/z}, \mathbf{F}_i^{xz}, \mathbf{G}_i^{s/t}, \bar{\mathbf{G}}_i^{s/t}\} \in \mathbb{R}^{256}$. Moreover, the stochastic gradient descent (SGD)

Table 1. Object Classification Accuracy (%) of target domain with **RGB-D** datasets as multi-view source domain. We adopt **bold** to highlight the best result and show the second best one with underline.

Method	A	D	W	Ar	Cl	Pr	Rw	C	Avg
ResNet[15]	61.75	79.37	81.73	35.90	28.86	48.01	52.68	74.82	57.89
DANN [10]	67.98	81.51	82.35	46.42	35.50	48.99	63.15	75.42	62.67
CDAN+E [31]	66.15	84.37	85.06	46.95	34.42	51.04	63.30	78.32	63.70
SRDC [40]	68.28	87.70	87.77	51.57	35.96	58.00	66.44	81.45	67.14
CGDM [7]	65.48	84.57	84.59	43.26	36.80	53.54	63.20	77.49	63.62
FixBi [35]	69.07	85.04	86.59	50.29	**38.33**	61.53	65.58	81.14	67.19
M3SDA [37]	66.11	85.70	85.86	45.10	37.00	56.53	64.96	80.71	65.25
DRT [26]	67.86	86.79	86.57	46.00	35.55	57.28	64.97	80.62	65.71
Ours	**70.79**	**89.68**	**90.87**	**56.17**	35.46	**66.86**	**70.33**	**84.21**	**70.55**

Table 2. Object Classification Accuracy (%) of target domain with **B3DO** datasets as multi-view source domain. We adopt **bold** to highlight the best result and show the second best one with underline.

Method	A	D	W	Ar	Cl	Pr	Rw	C	Avg
ResNet[15]	31.98	49.54	44.35	48.54	35.53	50.56	57.70	48.56	45.85
DANN [10]	44.05	63.53	62.35	59.61	40.05	67.09	74.98	68.18	59.98
CDAN+E [31]	47.70	66.75	64.69	62.00	43.93	70.29	77.93	71.35	63.08
SRDC [40]	49.47	68.67	66.74	64.44	45.85	72.77	79.73	73.55	65.15
CGDM [7]	47.19	66.07	64.09	61.23	43.15	69.70	76.97	70.42	62.35
FixBi [35]	49.67	68.59	66.69	63.97	45.41	71.72	80.23	72.82	64.89
M3SDA [37]	47.76	66.55	64.92	62.01	44.86	71.17	77.51	71.96	63.34
DRT [26]	47.75	67.59	66.01	63.00	44.22	70.84	78.62	72.82	63.86
Ours	**50.02**	**71.87**	**70.23**	**68.00**	**47.40**	**76.61**	**82.81**	**77.21**	**68.02**

optimizer with momentum 0.9 is used to optimize all parameters. The learning rate and batch size are 1e-3 and 96. The ϵ and δ are set as 1e-6 and 0.02 for all experiments. Our source code is available https://github.com/HaifengXia/IMVDA.

$\diamondsuit$ **Baselines:** In term of IMVDA, since source and target domains both involve one identical view data, the conventional unsupervised domain adaptation methods can exploit these samples to achieve alignment and identify target samples. Thus, we evaluate the DANN [10], CDAN+E [31], SRDC [40], CGDM [7], FixBi [35] under IMVDA scenario. Moreover, each view data of source domain can be considered as one independent domain. The multi-source domain adaptation methods M3SDA [37] and DRT [26] are used to solve IMVDA challenges. And we adopt their published source code and empirically search optimal parameters to conduct experiments.

4.2 Comparison of Results

The main experimental results in terms of target recognition accuracy are summarized in Table 1 and Table 2. According to the evaluation performance, we

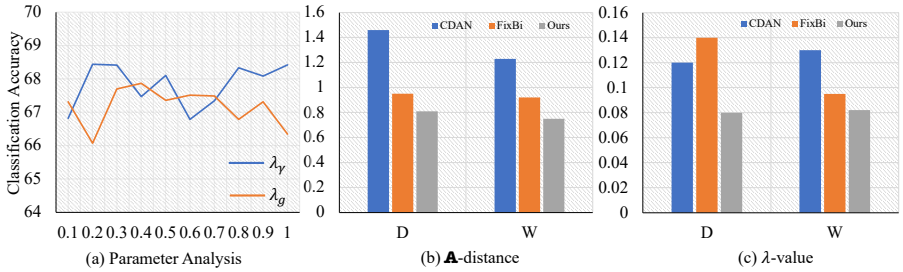

Fig. 2. Parameter analysis & Transfer ability. (a) Target classification accuracy with the varying parameters λ_γ and λ_g from 0.1 to 1.0 with B3DO as source domain. (b) $\mathcal{A}$-distance of source and target features from the same view data with RGB-D as source domain. (c) λ-value of three methods with tasks from RGB-D to D and W.

can easily achieve several significant conclusions. **First**, our method outperforms other baselines by a large margin on the average classification accuracy. Specifically, with RGB-D dataset as source domain, our CEKT surpasses the second best comparison (i.e., FixBi) by 3.36%. It illustrates the deployment of multi-view information effectively boosts the model performance on target domain even with considerable distribution shift. **Second**, we notice that our CEKT obtains much higher classification accuracy than others on the task RGB-D→Ar. As we all know, the images of Art Painting domain in Office-Home include lots of texture information to describe each object. On the other hand, depth sensor integrates more spatial information into depth images to clearly show the contour of object, which provides more discriminative semantic to the classification task. However, M3SDA and DRT, taking advantage of depth images to train the model, still fail to effectively assist the recognition of unlabeled target samples. These observations demonstrate our proposed solution not only emphasizes the specific semantic of depth images via source cross-view channel enhancement but also transfers such knowledge from source domain to target domain by reducing the negative influence of missing view with adaptive knowledge transfer network. **Third**, comparison of Table 1 and Table 2 shows that B3DO has larger distribution difference than RGB-D to the other target domain in Office-31, Office-Home and Caltech-256, as we achieve worse results by directly recognizing the target based on ResNet features. However, our proposed CEKT model can still achieve very close results no matter which source is used. In details, we improve the average accuracy from 57.89% to 70.55% by using RGB-D as source, while promote the average accuracy from 45.85% to 68.02% by using B3DO as source.

4.3 Empirical Analysis

Parameter Sensitivity. During training model, there are two parameters (λ_γ, λ_g) in our designed CEKT framework which are manually adjusted. These two parameters are changed from 0.1 to 1.0 with step size 0.1. To analyse the model

sensitivity to them, we record the classification accuracy of target domain with various parameter selection on task from B3DO to **Ar**, which is shown in Figure 2 (a). On the whole, the model is not sensitive to the change of parameters. However, larger λ_γ can easily bring more benefits to the model, while the smaller λ_g results in better performance, which further illustrates the proposed channel enhancement module effectively assists model to learning discriminative features. Note that for the selection of parameters, we randomly select 10% source samples as validation set for each tentative and use it to evaluate the model performance.

Transfer Ability. In addition, Ben-David theoretically points out the learning bound of domain adaptation [3] is determined by three parts: 1) the expected error $\varepsilon_s(h)$ of hypothesis h on source domain; 2) the $\mathcal{A}$-distance defined as $d_{\mathcal{H}\triangle\mathcal{H}} = 2(1-2\xi)$ measuring the domain mismatch, where ξ is the error from a trained domain classifier distinguishing source from target ones; 3) the error λ produced by the ideal hypothesis on both two domains. Inspired by this theoretical analysis, we report the $\mathcal{A}$-distance and λ-value over the shared-view data across source and target domains and show the results in Figure 2 (b)-(c). Compared with CDAN and FixBi, our proposed method obtains relative smaller $\mathcal{A}$-distance and λ-value on two tasks from RGB-D to **D** and **W**, which suggests that CEKT learns a model with a higher generalization ability.

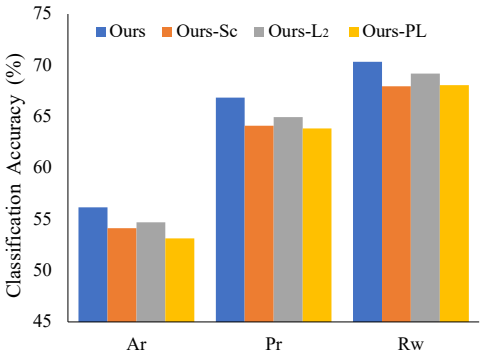

Fig. 3. Ablation study of model variants on three tasks with RGB-D as source domain.

Ablation Study. To clearly reflect the contribution of each component to the model performance, we carry out experiments on three knowledge transfer tasks with RGB-D as source domain by removing the corresponding operations. As previous mentioned, the source channel enhanced network actively discovers the view-common and view-specific parts via Eq. (3) and encourages the representation of important channels with Eq. (9). Thus, we replace Eq. (3) with $\sum_{c=1}^{C} |\gamma_{x/z,c}|$ (Ours-L_2) and attempt to remove Eq. (9) as Ours-Sc to study their effect. In addition, the model training adopts pseudo labels to facilitate feature with more discriminative power, and we further add a variant without the pseudo labeling as Ours-PL. Figure 3 reports the corresponding results with various methods on three tasks. According to it, we discover the enhancement with channel similarity and pseudo labels both produce significant and positive influence on improving model performance on target domain. Moreover, the sparse constraint for parameters $\gamma_{x/z,c}$ as [43] also results in the performance degradation, which further verifies the necessity of the preservation for the view-common channel split in multi-view data analysis.

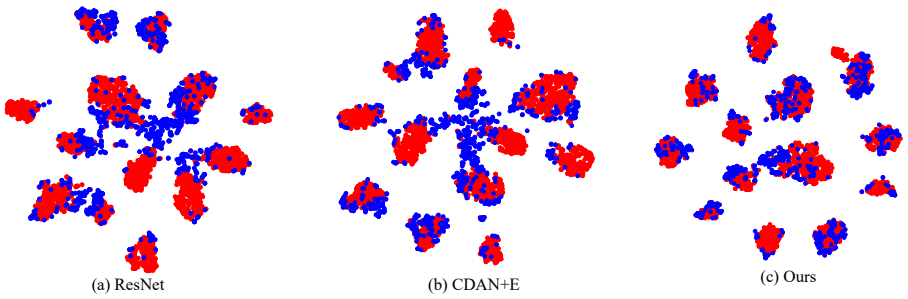

Fig. 4. Feature Visualization with t-SNE in 2D plane. The source and target features are represented by red and blue, respectively. And the experiment aims to transfer knowledge from RGB-D to **Ar** in Office-Home.

Feature Visualization To further understand the situation of distribution alignment, we follow [25] to visualize source and target features from the same view in 2D-plane, shown in Figure 4. Concretely, we access to the high-level features $\bar{\mathbf{G}}_i^{s/t}$ from the well-trained target model and adopt t-SNE technique to draw them in the canvas. Moreover, the experiment is carried out on adaptation task from RGB-D to **Pr** and ResNet as well as CDAN+E are considered as the competitors. According to the visualization results, it is easy to observe that there exist more overlaps between source and target features, compared with other baselines, which shows our method successfully mitigates the domain shift and better align them. Moreover, we notice that the classification boundary is more explicit than that in ResNet and CDAN+E. It suggests CEKT effectively learns the discriminative features for classification task.

5 Conclusion

Unsupervised domain adaptation (UDA) aims to learn the domain-invariant knowledge across well-supervised source and unlabeled target samples to enhance the model generalization ability. However, UDA assumes the instances per domain are captured by single sensor, which difficultly matches the practical scenario with multi-view data. This paper considered a practical and challenging problem named incomplete multi-view domain adaptation (IMVDA) which access to multi-view source data and single-view target samples. To overcome the challenge, we proposed a novel learning framework channel enhancement and knowledge transfer (CEKT). Concretely, CEKT first explored channel attributions to conduct semantic fusion and enhance the representation of view-common channels to learn more discriminative features. Moreover, adaptive knowledge transfer module not only brought multi-view knowledge to single-view feature learning but also achieved simple yet effective alignment across source and target domains. Considerable experimental results and analysis fully demonstrated our CEKT effectively broke the bottleneck of IMVDA by improving the performance.

References

1. Aljundi, R., Emonet, R., Muselet, D., Sebban, M.: Landmarks-based kernelized subspace alignment for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 56–63 (2015)
2. Baktashmotlagh, M., Harandi, M.T., Lovell, B.C., Salzmann, M.: Unsupervised domain adaptation by domain invariant projection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 769–776 (2013)
3. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al.: Analysis of representations for domain adaptation. Advances in neural information processing systems **19**, 137 (2007)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Ding, Z., Fu, Y.: Low-rank common subspace for multi-view learning. In: 2014 IEEE international conference on Data Mining. pp. 110–119. IEEE (2014)
6. Ding, Z., Li, S., Shao, M., Fu, Y.: Graph adaptive knowledge transfer for unsupervised domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 37–52 (2018)
7. Du, Z., Li, J., Su, H., Zhu, L., Lu, K.: Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3937–3946 (2021)
8. Elbrächter, D., Perekrestenko, D., Grohs, P., Bölcskei, H.: Deep neural network approximation theory. IEEE Transactions on Information Theory **67**(5), 2581–2623 (2021)
9. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE international conference on computer vision. pp. 2960–2967 (2013)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
11. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)
12. Guan, D., Huang, J., Lu, S., Xiao, A.: Scale variance minimization for unsupervised domain adaptation in image segmentation. Pattern Recognition **112**, 107764 (2021)
13. Guan, D., Huang, J., Xiao, A., Lu, S., Cao, Y.: Uncertainty-aware unsupervised domain adaptation in object detection. IEEE Transactions on Multimedia (2021)
14. He, J., Jia, X., Chen, S., Liu, J.: Multi-source domain adaptation with collaborative learning for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11008–11017 (2021)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. He, Y., Tian, Y., Liu, D.: Multi-view transfer learning with privileged learning framework. Neurocomputing **335**, 131–142 (2019)
17. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
18. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3d object dataset: Putting the kinect to work. In: Consumer depth cameras for computer vision, pp. 141–165. Springer (2013)

19. Jing, T., Liu, H., Ding, Z.: Towards novel target discovery through open-set domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9322–9331 (2021)
20. Jing, T., Xia, H., Hamm, J., Ding, Z.: Augmented multi-modality fusion for generalized zero-shot sketch-based visual retrieval. IEEE Transactions on Image Processing (2022)
21. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4893–4902 (2019)
22. Kumar, A., Rai, P., Daume, H.: Co-regularized multi-view spectral clustering. Advances in neural information processing systems **24**, 1413–1421 (2011)
23. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: 2011 IEEE international conference on robotics and automation. pp. 1817–1824. IEEE (2011)
24. Li, J., Li, Z., Lu, G., Xu, Y., Zhang, B., Zhang, D.: Asymmetric gaussian process multi-view learning for visual classification. Information Fusion **65**, 108–118 (2021)
25. Li, R., Jiao, Q., Cao, W., Wong, H.S., Wu, S.: Model adaptation: Unsupervised domain adaptation without source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9641–9650 (2020)
26. Li, Y., Yuan, L., Chen, Y., Wang, P., Vasconcelos, N.: Dynamic transfer for multi-source domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10998–11007 (2021)
27. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 6028–6039. PMLR (2020)
28. Liu, B., Chen, X., Xiao, Y., Li, W., Liu, L., Liu, C.: An efficient dictionary-based multi-view learning method. Information Sciences **576**, 157–172 (2021)
29. Liu, X., Guo, Z., Li, S., Xing, F., You, J., Kuo, C.C.J., El Fakhri, G., Woo, J.: Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10367–10376 (2021)
30. Liu, Y., Wang, L., Bai, Y., Qin, C., Ding, Z., Fu, Y.: Generative view-correlation adaptation for semi-supervised multi-view learning. In: European Conference on Computer Vision. pp. 318–334. Springer (2020)
31. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. arXiv preprint arXiv:1705.10667 (2017)
32. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: International conference on machine learning. pp. 2208–2217. PMLR (2017)
33. Lu, X., Wang, W., Danelljan, M., Zhou, T., Shen, J., Van Gool, L.: Video object segmentation with episodic graph memory networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 661–679. Springer (2020)
34. Lu, Z., Yang, Y., Zhu, X., Liu, C., Song, Y.Z., Xiang, T.: Stochastic classifiers for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9111–9120 (2020)
35. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1094–1103 (2021)

36. Pan, Y., Yao, T., Li, Y., Wang, Y., Ngo, C.W., Mei, T.: Transferrable prototypical networks for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2239–2247 (2019)
37. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1406–1415 (2019)
38. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
39. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: European conference on computer vision. pp. 213–226. Springer (2010)
40. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8725–8735 (2020)
41. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5018–5027 (2017)
42. Wang, M., Wang, W., Li, B., Zhang, X., Lan, L., Tan, H., Liang, T., Yu, W., Luo, Z.: Interbn: Channel fusion for adversarial unsupervised domain adaptation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3691–3700 (2021)
43. Wang, Q., Cheng, J., Gao, Q., Zhao, G., Jiao, L.: Deep multi-view subspace clustering with unified and discriminative learning. IEEE Transactions on Multimedia (2020)
44. Xia, H., Ding, Z.: Structure preserving generative cross-domain learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4364–4373 (2020)
45. Xia, H., Ding, Z.: Cross-domain collaborative normalization via structural knowledge. In: AAAI 2022 (2022)
46. Xia, H., Jing, T., Ding, Z.: Maximum structural generation discrepancy for unsupervised domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
47. Xu, H., Zhang, X., Xia, W., Gao, Q., Gao, X.: Low-rank tensor constrained co-regularized multi-view spectral clustering. Neural Networks **132**, 245–252 (2020)
48. Zhang, C., Cui, Y., Han, Z., Zhou, J.T., Fu, H., Hu, Q.: Deep partial multi-view learning. IEEE transactions on pattern analysis and machine intelligence (2020)
49. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12203–12213 (2020)
50. Zhang, D., Yang, G., Zhao, S., Zhang, Y., Ghista, D., Zhang, H., Li, S.: Direct quantification of coronary artery stenosis through hierarchical attentive multi-view learning. IEEE Transactions on Medical Imaging **39**(12), 4322–4334 (2020)
51. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: 2011 International conference on computer vision. pp. 471–478. IEEE (2011)
52. Zhao, J., Xie, X., Xu, X., Sun, S.: Multi-view learning overview: Recent progress and new challenges. Information Fusion **38**, 43–54 (2017)