

Cross-domain Contrastive Learning for Unsupervised Domain Adaptation

Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, *Fellow, IEEE*, and Yu-Gang Jiang

Abstract—Unsupervised domain adaptation (UDA) aims to transfer knowledge learned from a fully-labeled source domain to a different unlabeled target domain. Most existing UDA methods learn domain-invariant feature representations by minimizing feature distances across domains. In this work, we build upon contrastive self-supervised learning to align features so as to reduce the domain discrepancy between training and testing sets. Exploring the same set of categories shared by both domains, we introduce a simple yet effective framework CDCL, for domain alignment. In particular, given an anchor image from one domain, we minimize its distances to cross-domain samples from the same class relative to those from different categories. Since target labels are unavailable, we use a clustering-based approach with carefully initialized centers to produce pseudo labels. In addition, we demonstrate that CDCL is a general framework and can be adapted to the data-free setting, where the source data are unavailable during training, with minimal modification. We conduct experiments on two widely used domain adaptation benchmarks, *i.e.*, Office-31 and VisDA-2017, for image classification tasks, and demonstrate that CDCL achieves state-of-the-art performance on both datasets.

Index Terms—Contrastive Learning, Unsupervised Domain Adaptation, Source Data-free.

I. INTRODUCTION

At the heart of many machine learning and computer vision tasks is to learn robust feature representations that generalize well to novel testing samples. However, state-of-the-art deep learning models still suffer from significant performance drops even when the testing distribution slightly drifts from the training distribution. To mitigate this issue, unsupervised domain adaptation [1]–[6] aims to reduce the discrepancy between training and testing, which is also known as domain shifts. This is generally achieved by aligning the distribution of a labeled training set (source domain) with that of an unlabeled testing set (target domain) [4], [7]. In particular, feature alignment aims to minimize carefully designed metrics like Maximum Mean Discrepancies (MMD) [8], covariances [9], [10], and adversarial loss functions [5], [11] such that the distances between training and testing distributions are reduced.

The idea of reducing feature distance in UDA tasks is similar in spirit to recent advances in self-supervised contrastive learning, which pulls an image to be closer to its

own augmented copy on a hypersphere compared to other images. In this paper, we ask the following question: can we leverage contrastive learning that produces decent feature representations in a variety of downstream tasks [12]–[14] for domain alignment in unsupervised domain adaptation? While appealing, this is non-trivial as in standard contrastive learning a positive pair can be naturally generated considering two related views of the same image, since they contain the same content but are transformed with different augmentations. In domain adaptation, it is not clear how to form positive and negative pairs in order to align feature distributions.

Exploring the fact that categories are shared between the source and target domain, we propose to align feature representations conditioned on class information to learn domain-invariant features. In particular, we argue that samples within the same class should be closer to each other while samples from different categories should lie far apart, even when they are from different domains.

In light of this, we introduce CDCL, a simple yet effective framework for unsupervised domain adaptation under both standard and data-free settings. As shown in Figure 1, given an anchor image from the source domain, we randomly select samples from the target domain that belong to the same class as the anchor to form positive pairs, based on pseudo labels of target samples in lieu of manual labels. We minimize the distance of all positive pairs relative to negative pairs, which are formed by cross-domain samples from different categories. Since labels are not available for the target domain, we generate pseudo labels with k-means clustering, whose initial clusters are set to class prototypes learned on the source domain. Through minimizing feature distance with the proposed cross-domain contrastive loss, CDCL produces domain-invariant features. We further show CDCL can be conveniently adapted to the newly proposed data-free scenario [15], where the source data are not available, by replacing sample features with prototypical features.

We conduct extensive experiments on two widely used domain adaptation benchmarks, *i.e.*, Office-31 [6] and VisDA [16] and demonstrate that our method achieves state-of-the-art performance on both datasets. We further show that CDCL can effectively produce domain-invariant features even when source data are not available. We also conduct a set of ablation experiments to validate the effectiveness of different components of our approach.

II. RELATED WORK

Unsupervised domain adaptation (UDA). Existing UDA methods focus on learning domain-invariant feature represen-

Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen and Yu-Gang Jiang are with Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University and Shanghai Collaborative Innovation Center on Intelligent Visual Computing.

Guo-Jun Qi is with the Seattle Cloud Lab, Futurewei Technologies, Bellevue, WA 98004.

Corresponding author: Zuxuan Wu.

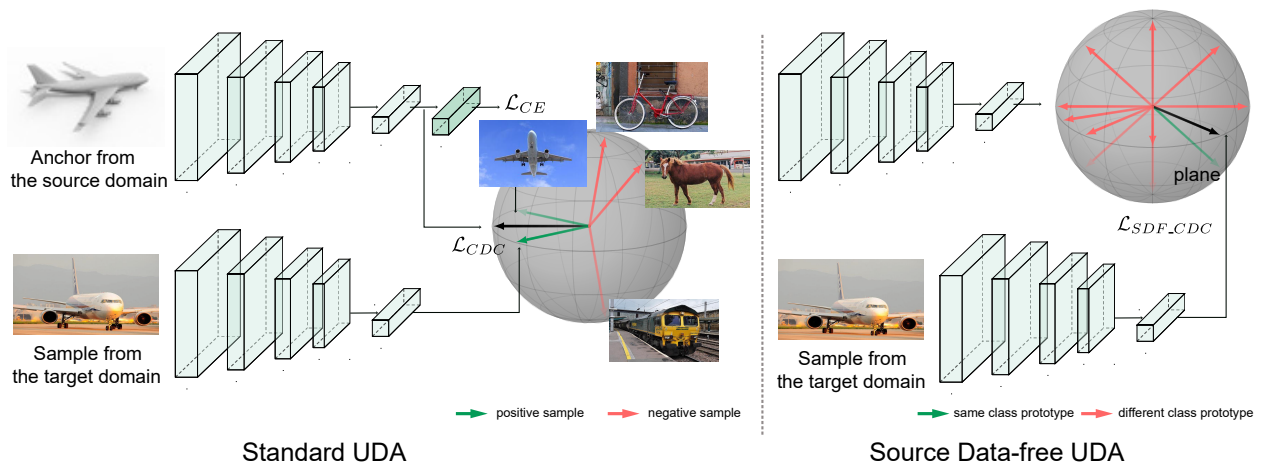


Fig. 1: A conceptual overview of our approach in the standard UDA setting (left) and the data-free setting (right). Left: Given an anchor image from the source domain, we pull its feature to be closer to samples of the same category from the target domain while pushing apart its feature with those from a different class in the target domain. Right: when source data are not available, we replace source samples with prototypical features derived from a pre-trained model in the source domain. See texts for more details.

tations. One direction of UDA methods is minimizing the discrepancy between different domains [4], [7], [9], [10], [17]. In early approaches [4], [7], MK-MMD and joint MMD are employed to measure the discrepancy between the source domain and target domain. Besides, higher-order statistics and other well-designed discrepancies are utilized in [17]–[21]. To leverage category information for domain alignment, CAN [22] introduces intra- and inter-class domain discrepancies. In [22], class information from the target domain is obtained by K-means clustering, which is similar to the generation of pseudo labels in our method. However, CAN is unpractical for source data-free UDA due to the utilization of source data when minimizing the domain discrepancy between source samples and target samples. Another direction is to design an adversarial optimization objective for a domain discriminator and to obtain domain-invariant representations by adversarial learning [5], [23]. GVB-GD [24] promotes adversarial domain adaptation with a gradually vanishing bridge mechanism. GSDA [25] implements hierarchical domain alignment with multiple adversarial discriminators. Recently, in addition to discrepancy-based methods and adversarial-based methods, there are other UDA methods, e.g., domain-adaptive dictionary learning [26], multi-modality representation learning [27] and feature disentanglement [28]. Some recent approaches [29]–[31] also explore feature norm or batch norm for UDA. SAFN [29] enlarges feature norms of different domains to improve the transferability of features. [30] uses the domain-specific batch normalization and [31] performs batch nuclear-norm maximization to generate discriminative and diverse predictions. To improve feature discriminability, BSP [32] penalizes the largest singular values and ADR [33] applies dropout on the classifier. To avoid ambiguous target features, MCD [34] maximizes the discrepancy between two classifiers and STAR [35] samples classifiers from Gaussian distributions without more parameters. In this paper, we align features with

contrastive learning, which is simple and easy to optimize.

Contrastive learning. Great progress in unsupervised representation learning has been achieved by self-supervised contrastive learning [12]–[14], [36]. The standard approach of contrastive learning is to learn discriminative representations by pulling together positive pairs and pushing apart negative pairs. In self-supervised learning methods [12]–[14], the positive pairs are produced by creating different augmented views of each sample, while negative pairs can be randomly chosen from different samples. Instance discriminative representations learned by self-supervised contrastive learning can be transferred well to downstream tasks with fine-tuning. However, without task-specific semantic information, representations with intra-class compactness and inter-class discrimination can not be learned through instance-level contrastive learning. Recently, supervised contrastive learning [37] leverages category labels to compose positive and negative pairs and achieves promising performance on fully-supervised image classification. [38] proposes a self-paced contrastive learning framework for domain adaptive object re-ID with multi-level supervision in each domain. Nonetheless, feature alignment, which is critical for domain adaptation methods, is not considered in these contrastive learning methods. There are some approaches [39]–[41] applying contrastive learning for other UDA tasks. [39] simply performs contrastive learning on each domain independently and minimizes MMD to reduce the domain gap. Nevertheless, neither domain alignment nor class alignment is considered in the contrastive loss of [39]. CoSCA [40] enhances the MCD [34] framework with contrastive loss that separates the ambiguous target samples, and uses MMD to obtain better global domain alignment. As mentioned in [34], the classification loss on the source data is necessary for MCD when maximizing the discrepancy of two classifiers on the target data. Therefore, CoSCA can not be directly transferred to source data-free UDA. [41] proposes an effective feature clustering-based strategy to capture the different semantic

modes of the feature distribution and group features of the same class into tight and well-separated clusters for improved unsupervised domain adaptation in semantic segmentation. However, it confuses the data of the source and target domains during the clustering-based training process, making it impossible to fit in source data-free setting. It is worth noting a concurrent work explores the idea of contrastive learning for domain adaptation [42]. However, the approach mixes samples from all domains and thus can only be used in the standard UDA setting. Besides, the results of our ablation study show that if ignoring the domain-level information and simply considering label information in the contrastive loss like [42], the performance will degrade.

Source Data-free UDA. Recently, due to the concern of source data privacy in the realistic applications of UDA methods, source data-free UDA has been proposed by [15]. The main challenge of source data-free UDA is that a pre-trained model on the source domain should be adapted to the target domain without access to source data. Based on hypothesis transfer learning, [15] proposes a self-training framework with mutual information maximization and pseudo-labeling strategy. In [43], a collaborative class conditional generative adversarial network is employed to avoid the usage of source data with target data generation and model adaptation. In this paper, instead of using entropy minimization, we leverage contrastive learning for cross-domain alignment.

III. METHODOLOGY

Unsupervised domain adaptation aims to transfer models learned on a labeled source domain to an unlabeled target domain, whose data distribution is different from that of the source domain. During training, UDA assumes access to all labeled samples in the source domain as well as unlabeled images from the target domain. Formally, given a fully-labeled source domain dataset with N_s image and label pairs $D_s = (\mathcal{X}_s, \mathcal{Y}_s) = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$, and an unlabeled dataset in a target domain with N_t images $D_t = \mathcal{X}_t = \{x_t^i\}_{i=1}^{N_t}$, both $\{x_s^i\}$ and $\{x_t^i\}$ belong to the same set of M predefined categories. We use $y_s^i \in \{0, 1, \dots, M-1\}$ to represent the label of the i -th source sample while the labels of target samples are unknown during training. UDA aims to predict labels of testing samples in the target domain using a model $f_t : \mathcal{X}_t \rightarrow \mathcal{Y}_t$ trained on $D_s \cup D_t$. The model, parameterized by θ consists of a feature encoder $g : \mathcal{X}_t \rightarrow \mathbb{R}^d$ and a classifier $h : \mathbb{R}^d \rightarrow \mathbb{R}^M$, where d is the dimension of features produced by the encoder.

Our goal is to align feature distributions between the source and the target domain through contrastive self-supervised learning. To this end, we first briefly review contrastive learning, and then introduce CDCL that forms positive and negative pairs in a cross-domain manner to learn domain-invariant features. Finally, we show that the proposed approach is not only suitable for standard UDA but can also be applied to data-free scenarios, where the source data are unavailable during training.

A. Contrastive Learning with InfoNCE

State-of-the-art contrastive learning frameworks typically use the N-pair loss [44], also known as the InfoNCE [12], [36] and

NT-Xent loss [13], to minimize the distance of a positive pair relative to all other pairs. More formally, let u and v denote ℓ_2 -normalized feature representations of a pair and the loss function is then defined as:

$$\mathcal{L} = - \sum_{v^+ \in V^+} \log \frac{\exp(u^\top v^+ / \tau)}{\exp(u^\top v^+ / \tau) + \sum_{v^- \in V^-} \exp(u^\top v^- / \tau)}, \quad (1)$$

where $v^+ \in V^+$ and $v^- \in V^-$ represent the positive and negative samples with respect to u , and τ is a temperature parameter that is manually set. In practice, a positive pair is derived by two random data augmentations (*i.e.*, blurring and color jittering, *e.t.c.*) operated on the same image sample, resulting in two correlated views. In contrast, in domain adaptation, it remains unclear how to obtain positive and negative pairs for feature alignment.

B. Cross-domain Contrastive Learning

We now introduce how to form pairs to learn domain-invariant features with contrastive learning. Since samples from the source domain and the target domain belong to the same set of classes in current UDA settings, we build upon this assumption to reduce domain shift. More specifically, we hypothesize that samples within the same category are close to each other while samples from different classes lie far apart, regardless of which domain they come from. More formally, we consider the ℓ_2 -normalized features z_i^i from the i -th sample x_i^i in the target domain as an anchor, and it forms a positive pair with a sample in the same class from the source domain, whose features are denoted as z_s^p , we formulate the cross-domain contrastive loss as:

$$\mathcal{L}_{CDC}^{t,i} = - \frac{1}{|P_s(\hat{y}_t^i)|} \sum_{p \in P_s(\hat{y}_t^i)} \log \frac{\exp(z_i^{i\top} z_s^p / \tau)}{\sum_{j \in I_s} \exp(z_i^{i\top} z_s^j / \tau)} \quad (2)$$

where I_s denotes the set of source samples in a mini-batch and $P_s(\hat{y}_t^i) = \{k \mid y_s^k = \hat{y}_t^i\}$ indicates the set of positive samples from the source domain that share the same label with the target anchor x_t^i . Since we do not have access to labels of target samples, we use estimated pseudo labels \hat{y}_t^i (as will be introduced below) to generate pairs. The cross-domain loss forces intra-class distance to be smaller than inter-class distance for samples from different domains so as to reduce domain shift. It is worth pointing out that compared to the standard InfoNCE loss, we sum over all samples in a mini-batch from the source domain that belong to the same category as the anchor x_t^i , which could reduce sampling variance.

In Eqn. 2, we consider samples from the target domain as anchors. Alternatively, we can use source samples as anchors and compute $\mathcal{L}_{CDC}^{s,i}$ similarly by setting $P_s(y_s^i) = \{k \mid \hat{y}_t^k = y_s^i\}$. Then, we combine $\mathcal{L}_{CDC}^{s,i}$ with $\mathcal{L}_{CDC}^{t,i}$ to derive the cross-domain contrastive loss as follows:

$$\mathcal{L}_{CDC} = \sum_{i=1}^{N_s} \mathcal{L}_{CDC}^{s,i} + \sum_{i=1}^{N_t} \mathcal{L}_{CDC}^{t,i}. \quad (3)$$

The cross-domain contrastive loss aligns features in a bi-directional manner by using anchors from both domains for improved performance. Finally, combining the cross-domain contrastive loss with a standard cross-entropy loss \mathcal{L}_{CE} enforced on the source domain, we have the final objective function for training:

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}_{CE}(\theta; D_s) + \lambda \mathcal{L}_{CDC}(\theta; D_s, D_t), \quad (4)$$

where λ controls the trade-off between the two loss terms and θ denotes the parameters to be optimized.

Algorithm 1: Pseudo code of CDCL for standard UDA.

Result: θ for the prediction model f
Input: unlabeled target dataset $D_t = \mathcal{X}_t$, source dataset $D_s = (\mathcal{X}_s, \mathcal{Y}_s)$, model $f = h \circ g$, max epoch E , iterations per epoch K
Initialize encoder g with ImageNet pre-trained weights
for $e = 1$ **to** E **do**
 Initialize cluster centers with source class prototypes using Eqn. 7
 Perform K-means clustering on target data \mathcal{X}_t , obtain pseudo labels \hat{y}_t^i
 for $k = 1$ **to** K **do**
 Sample batch (x_s^i, y_s^i) from D_s and compute \mathcal{L}_{CE}
 Sample batches (x_s^j, y_s^j) and (x_t^j, \hat{y}_t^j) from D_s and D_t
 Compute \mathcal{L}_{CDC} using Eqn. 3
 Back-propagate and update model f via Eqn. 4
 end
end

Algorithm 2: Pseudo code of CDCL for source data-free UDA.

Result: θ for the prediction model f
Input: unlabeled target dataset $D_t = \mathcal{X}_t$, model $f = h \circ g$ pre-trained on source dataset D_s , max epoch E , iterations per epoch K
Freeze the parameters of classifier h
for $e = 1$ **to** E **do**
 Initialize cluster centers with source class prototypes using Eqn. 5
 Perform K-means clustering on target data \mathcal{X}_t , obtain pseudo labels \hat{y}_t^i
 for $k = 1$ **to** K **do**
 Sample batch (x_t^i, \hat{y}_t^i) from D_t and compute $\mathcal{L}_{SDF-CDC}^{t,i}$ using Eqn. 6
 Back-propagate and update model f via Eqn. 8
 end
end

C. Pseudo Labels for the Target Domain

Ground-truth labels from the target domain are not available during training, and thus we leverage k-means clustering to produce pseudo labels [15], [22], forming pairs for cross-domain

contrastive learning. Since K-means is sensitive to initialization, using randomly generated clusters fails to guarantee related semantics with respect to predefined categories. To mitigate this issue, we set the number of clusters to the number of classes M and use class prototypes from the source domain as initial clusters. The benefits of initializing the cluster centers with class prototypes are twofold: i) source class prototypes can be seen as the approximation of target class prototypes, since features used are high-level and contain semantics information (ii) with the alignment of samples in the same category by CDCL, this approximation will be more accurate as the training continues. More formally, we first compute the centroid of source samples in each category as the corresponding class prototype and the initial cluster center O_t^m for the m -th class is defined as:

$$O_t^m \leftarrow O_s^m = \mathbb{E}_{i \sim D_s, y_s^i = m} z_s^i. \quad (5)$$

Given features from the target domain, we then perform spherical K-means clustering using these carefully initialized centers. When determining the assignment of each target sample, cosine similarity is adopted to measure the distance between the target feature z_t^i and the m -th cluster center O_t^m . Once clustering is finished, each sample in the target domain x_t^i is associated with a pseudo label \hat{y}_t^i . To reduce the noise in target pseudo labels, we remove the ambiguous samples far from its assigned clustered centers. Concretely, one target sample will be removed when the cosine similarity between its feature and its assigned cluster center is below a manually set threshold d .

D. Source Data-free UDA

In this section, we demonstrate that CDCL can be easily adapted to a newly introduced source data-free setting [15], where a model trained on the source domain is provided yet source data are unavailable due to corruption or privacy concerns. Formally, the goal is to learn a model $f_t : X_t \rightarrow Y_t$ and predict $\{y_t^i\}_{i=1}^{N_t}$ with only unlabeled target data D_t and a pre-trained source model $f_s : X_s \rightarrow Y_s$. The pre-trained source model is obtained by minimizing the cross-entropy loss on the source samples.

It is difficult for most previous UDA methods to adapt to source data-free UDA. For discrepancy-based methods, the predefined domain discrepancies are statistics that should be measured between source samples and target samples. Similarly, for adversarial-based methods, the adversarial discriminators need to be trained with source samples and target samples. Without access to source data, both strategies are not applicable to perform adaptation to the target domain. Besides, for the standard UDA setting, many UDA methods assume that the same feature encoder is shared on the source and target domains. However, this constraint may be hard to implement under source data-free setting since the feature encoder can not be trained on the source and target domain simultaneously. Some UDA methods, e.g., DSBN [30], prove that a domain-specific module in the feature encoder can improve the performance of domain adaptation. Therefore, it is practical to remove the parameter-sharing constraint for feature encoders under source data-free setting.

For CDCL, the lack of samples from the source domain D_s makes it challenging to (1) form positive and negative pairs and (2) to compute source class prototypes. We address this issue by replacing source samples with classifier weights from the trained model f_s . The intuition is that the weight vectors in the classifier layer of a pre-trained model can be regarded as prototypical features of each class learned on the source domain. In particular, we first remove the bias of the fully-connected layer and perform normalization for the classifier.

We use w_s^m to denote the weight vector of the m -th class in the classification layer $\mathbf{W}_s = [w_s^1, \dots, w_s^M]$ learned on the source domain. Since the weights are normalized, we use them as class prototypes. When adapting to the target domain, we freeze the parameters of the classifier layer to keep the source class prototypes and only train the feature encoder. Through replacing the source samples with source class prototypes, the cross-domain contrastive loss under the source data-free setting can be written as:

$$\mathcal{L}_{SDF-CDC}^{t,i} = - \sum_{m=1}^M \mathbf{1}_{\hat{y}_i^t=m} \log \frac{\exp(z_i^{t\top} w_s^m / \tau)}{\sum_{j=1}^M \exp(z_i^{t\top} w_s^j / \tau)}. \quad (6)$$

Similarly, we estimate labels for samples in the target domain with clustering. However, it is not feasible to compute class prototypes using samples anymore. Instead, we replace Eqn. 5 with class weights:

$$O_t^m \leftarrow O_s^m = w_s^m \quad (7)$$

The final objective of source data-free UDA is:

$$\text{minimize} \sum_{i=1}^{N_t} \mathcal{L}_{SDF-CDC}^{t,i}. \quad (8)$$

Compared to Eqn. 4, the cross-entropy loss is not used since the source data are no longer available for supervised training.

IV. EXPERIMENTS

A. Datasets and Compared Approaches

We use two public benchmarks to evaluate our method for unsupervised domain adaptation under both standard and data-free settings.

VisDA-2017 [16] is a challenging large-scale benchmark including 12 classes from two domains: the source domain with 152,397 synthetic images, and the target domain contains 55,388 real-world images. Our method is evaluated on the synthesis-to-real domain adaptation task.

Office-31 [6] is a common DA benchmark which contains 4,110 images from three distinct domains, *i.e.*, Amazon (**A** with 2,817 images), DSLR (**D** with 498 images) and Webcam (**W** with 795 images). Each domain consists of 31 object categories. Our method is evaluated by performing domain adaptation on each pair of domains, which generates 6 different tasks.

Compared Approaches. We first report the results of a model trained on the source domain only, and compare with the following state-of-the-art approaches:(a) DANN [5], which utilizes a domain discriminator with adversarial optimization

objective to reduce the domain gap. (b) DAN [4] and JAN [7], which learn domain-invariant features by minimizing MK-MMD and Joint MMD. (c) ADR [33], which encourages the encoder to generate more discriminative features by using dropout on the classifier. (d) SAFN [29], which adapts the feature norms of different domains to a large range of values. (e) SWD [18], which measures the dissimilarity between the output of classifiers with sliced Wasserstein discrepancy. (f) MMAN [27], which introduces semantic multi-modality representation learning into adversarial domain adaptation and captures fine-grained category information by multi-channel constraint. (g) CDAN [23], which aligns the conditional distribution in adversarial learning. (h) DSBN [30], which adopts the domain-specific batch normalization in models. (i) BSP [32], which improves the feature discriminability by penalizing the largest singular values. (j) BNM [31], which achieves the discriminability and diversity of the predictions with batch nuclear-norm maximization. (k) MDD [19], which proposes a hypothesis-induced discrepancy for domain adaptation. (l) GVB-GD [24], which proposes a gradually vanishing bridge mechanism for adversarial-based domain adaptation. (m) GSDA [25], which aims to learn domain invariant representations by hierarchical domain alignment. (n) STAR [35], which tries to employ more classifiers by sampling from Gaussian distribution without more parameters. (o) CAN [22], which introduces class information into domain alignment by minimizing the contrastive domain discrepancy. (p) SHOT [15], which develops a framework for data-free UDA based on hypothesis transfer learning. (q) ModelAdapt [43], which adopts a collaborative class conditional generative adversarial networks to avoid using source data.

Among these methods, SHOT and ModelAdapt are developed under the source data-free UDA setting. We implement our method under both standard UDA and source data-free UDA settings.

B. Implementation Details

Network architecture. We adopt a ResNet-50 (for Office-31) and ResNet-101 (for VisDA-2017) pre-trained on ImageNet [46] as the feature encoder in the experiments of the standard UDA task. We replace the last FC layer with the task-specific FC classifier layer. All the network parameters are shared between different domains except those of the batch normalization (BN) layers as we utilize the domain-specific BN [30]. Under the source data-free UDA setting, following [15], we use one bottleneck layer containing a FC layer and a BN layer after convolutional layers in the feature encoder module g . Furthermore, we remove the bias of the task-specific FC classifier layer and perform normalization for the classifier.

Training details. The network is trained by using mini-batch SGD with a momentum of 0.9. The initial learning rate η_0 is set as $1e^{-3}$ for pre-trained convolutional layers and $1e^{-2}$ for newly added layers. We employ the same learning rate scheduler $\eta = \eta_0 \cdot (1 + 10 \cdot p)^{-b}$ as [4], [5], [7], where p denotes training process linearly increase from 0 to 1. Following [22], for Office-31, $b = 0.75$ while for VisDA-2017, $b = 2.25$. When pre-training models with source samples under source data-free

TABLE I: Accuracy(%) on VisDA-2017 for unsupervised domain adaptation (ResNet-101). † denotes that this method is developed under the source data-free UDA setting.

Method	plane	beycl	bus	car	horse	knife	meycl	person	plant	sktbrd	train	truck	Avg
ResNet-101 [45]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [5]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
DAN [4]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
ADR [33]	87.8	79.5	83.7	65.3	92.3	61.8	88.9	73.2	87.8	60.0	85.5	32.3	74.8
CDAN [23]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.7
CDAN+BSP [32]	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
SAFN [29]	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
SWD [18]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
MSTN+DSBN [30]	94.7	86.7	76.0	72.0	95.2	75.1	87.9	81.3	91.1	68.9	88.3	45.5	80.2
STAR [35]	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
CoSCA [40]	95.7	87.4	85.7	73.5	95.3	72.8	91.5	84.8	94.6	87.9	87.9	36.8	82.9
CAN [22]	97.0	87.2	82.5	74.3	97.8	96.2	90.8	80.7	96.6	96.3	87.5	59.9	87.2
JCL [42]	97.0	91.3	84.5	66.8	96.1	95.6	89.8	81.5	94.7	95.6	86.1	71.8	87.6
CDCL (ours)	97.4	89.5	85.9	78.2	96.4	96.8	91.4	83.7	96.3	96.2	89.7	61.6	88.6
SHOT [15] †	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
ModelAdapt [43] †	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
CDCL (ours) †	97.3	90.5	83.2	59.9	96.4	98.4	91.5	85.6	96.0	95.8	92.0	63.8	87.5

TABLE II: Accuracy(%) on Office-31 for unsupervised domain adaptation (ResNet-50). † denotes that this method is developed under the source data-free UDA setting.

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg
ResNet-50 [45]	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DAN [4]	78.6	80.5	63.6	97.1	62.8	99.6	80.4
DANN [5]	79.7	82.0	68.2	96.9	67.4	99.1	82.2
JAN [7]	84.7	85.4	68.6	97.4	70.0	99.8	84.3
MMAN [27]	85.8	85.8	70.3	97.4	71.2	100.	85.1
SAFN+ENT [29]	92.1	90.3	73.4	98.7	71.2	100.	87.6
CDAN [23]	92.9	94.1	71.0	98.6	69.3	100.	87.7
CDAN+BSP [32]	93.0	93.3	73.6	98.2	72.6	100.	88.5
CDAN+BNM [31]	92.9	92.8	73.5	98.8	73.8	100.	88.6
MDD [19]	93.5	94.5	74.6	98.4	72.2	100.	88.9
GVB-GD [24]	95.0	94.8	73.4	98.7	73.7	100.	89.3
GSDA [25]	94.8	95.7	73.5	99.1	74.9	100.	89.7
CAN [22]	95.0	94.5	78.0	99.1	77.0	99.8	90.6
CDCL (ours)	96.0	96.0	77.2	99.2	75.5	100.	90.6
SHOT [15] †	94.0	90.1	74.7	98.4	74.3	99.9	88.6
ModelAdapt [43] †	92.7	93.7	75.3	98.5	77.8	99.8	89.6
CDCL (ours) †	94.4	92.1	76.4	98.5	74.1	100	89.3

setting, following [15], we randomly split the source dataset into 0.9/0.1 train-validation sets and select the optimal source pre-trained model based on the validation set. We use one RTX 3090 with 24GB for experiments.

C. Main Results

Table I and Table II summarize the results of our approach and comparisons with state-of-the-art methods. On both datasets, We can see that all domain adaptation methods achieve significantly better results compared to the source-only method, confirming the importance of feature alignment. On VisDA, we observe CDCL achieves a mean accuracy of 88.6% across all categories, outperforming all state-of-the-art approaches and boosting the accuracy of source-only baseline by 26.2%. In particular, CDCL is better than CAN [22] by

1.4% absolute point and beats JCL [42] by 1.0%, which is notable given that VisDA is a challenging benchmark.

When the source data are no longer available, CDCL achieves a mean accuracy of 87.5%, outperforming the state-of-the-art approach SHOT [15] by 4.6% point, highlighting the effectiveness of our approach. Comparing the data-free setting and the conventional UDA setting, we see that CDCL is slightly (0.9%) worse in the data-free setting. It is noteworthy that CDCL for source data-free setting still surpasses many UDA methods for standard setting. This suggests that one can simply explore a model trained on the source domain for effective transfer without the need to access the source data.

We observe similar trends on Office-31 under both settings. In particular, in the conventional UDA setting, CDCL offers a mean accuracy of 90.6% across 6 different tasks, which is on par with the best results in the literature. Under the data-free setting, CDCL achieves a mean accuracy of 89.3%, comparable to ModelAdapt. It is worth mentioning that results are better on Office-31 compared to VisDA since the dataset is smaller. Besides, due to the small domain gap between D and W, it is easy to achieve high accuracy on Office-31 tasks D→W and W→D, even for source-only models. Since domains **W** and **D** contain fewer samples compared to domain **A**, the performance on tasks W→A and D→A is relatively poor.

D. Ablation Studies and Discussions

In this section, we conduct a set of ablation experiments to justify the effectiveness of different components and provide discussions.

Positive and Negative Pairs. We mainly form positive pairs and negative pairs using cross-domain samples. We now discuss alternative ways to form pairs and report results on VisDA. In particular, we use the following approach to form pairs: (1) In-domain, where anchors come from both domains but samples that are used to form pairs are from the same domain and same category as the anchor; (2) Combined-domain, where the

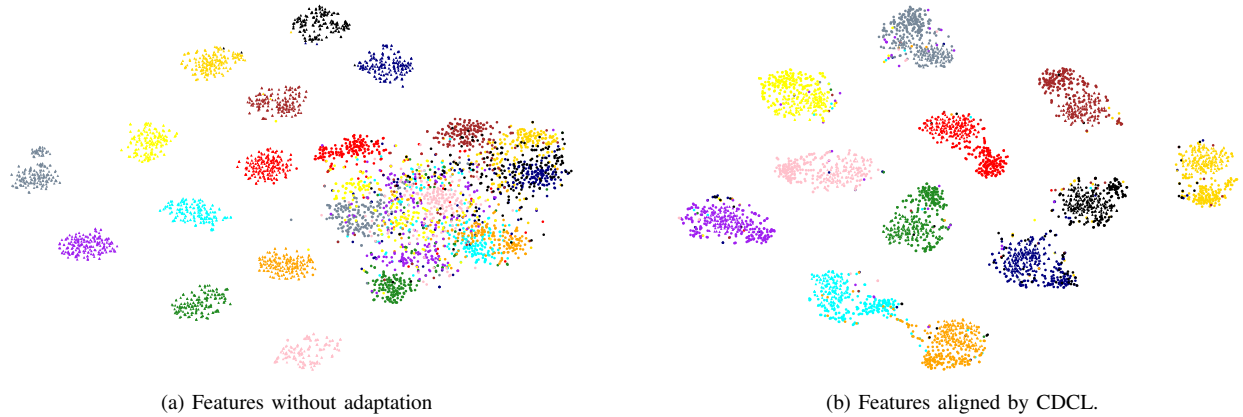


Fig. 2: The t-SNE [47] visualization of features from the source domain and the target domain before and after alignment. The triangle and circle markers indicate the source and target samples respectively, and different colors denote different classes.

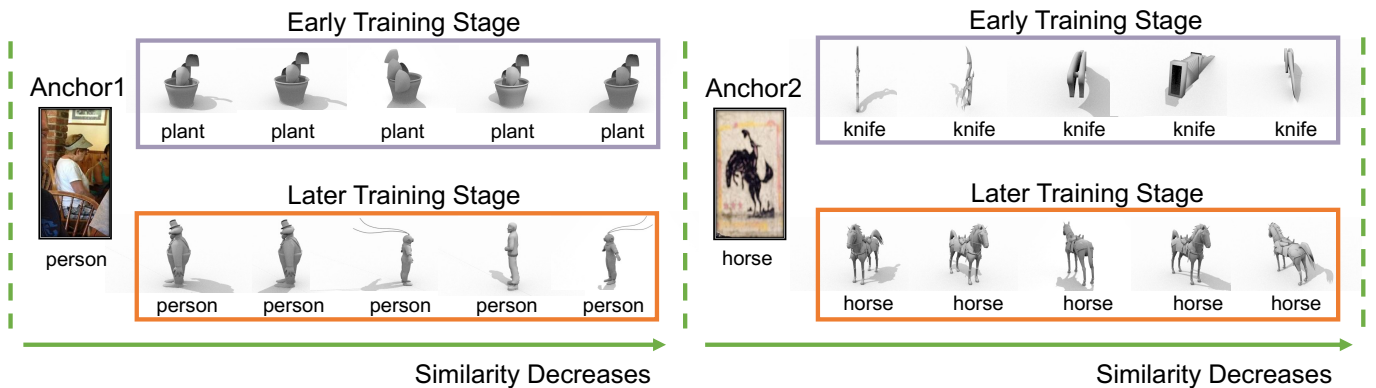


Fig. 3: Given an anchor from one domain, top-5 cross-domain samples are retrieved by comparing feature similarity after the first epoch of training (Top) and at the end of training (Bottom).

source and target domains are mixed and pairs are generated by simply considering label information; (3) Cross-domain ($\mathcal{L}_{CDC}^{s,i}$ only), which simply $\mathcal{L}_{CDC}^{s,i}$ in Eqn. 3 using samples in the source domain as anchors; (4) Cross-domain ($\mathcal{L}_{CDC}^{t,i}$ only), which uses samples in the target domain as anchors. From the results shown in Table III, we observe that cross-domain alignment achieves better results compared to performing a simply in-domain alignment. This suggests the importance of forming pairs from two domains in order to produce domain-invariant features. In addition, we observe that mixing both domains together is worse than CDCL, possibly due to the fact that jointly modeling intra-class and inter-class information is challenging. Moreover, the bi-directional use of anchors is better compared to using anchors simply from one domain.

The impact of hyper-parameters. We test the sensitivity of CDCL to the temperature τ on VisDA-2017 for standard UDA. As shown in Figure 4(a), the accuracies around $\tau = 0.05$ are not sensitive. When the τ grows larger, the accuracy steadily increases before decreasing. Additionally, we study the effect of λ on VisDA. Results in Figure 4(b) show that the accuracy around $\lambda = 1.6$ are also not sensitive and the gap of accuracies is smaller than 0.2 when $\lambda > 1.4$. Therefore, CDCL is not sensitive to its hyper-parameters.

TABLE III: Ablation study for the selection of anchors, positive samples and negative samples in contrastive loss.

Method	Anchor	Positive	Negative	VisDA
In-domain	all	same	same	86.5
Combined-domain	all	all	all	87.3
Cross-domain ($\mathcal{L}_{CDC}^{s,i}$ only)	source	different	different	87.5
Cross-domain ($\mathcal{L}_{CDC}^{t,i}$ only)	target	different	different	86.6
CDCL	all	different	different	88.6

Feature Visualization. We further use t-SNE [47] to visualize features from the source and target domain before and after alignment in Figure 2. We can see that before alignment, source and target features are separated into two clusters, which demonstrates the gap between the two domains. After alignment with CDCL, we see that features from different domains are mixed together and features from different classes are well separated.

Learned Feature Distance. We visualize in Figure 3 top-5 retrieved images given an anchor image at the beginning and the end of training. We observe that in early states, retrieved samples are similar to the anchor in shape but belong to a different category. As training continues, CDCL gradually pulls

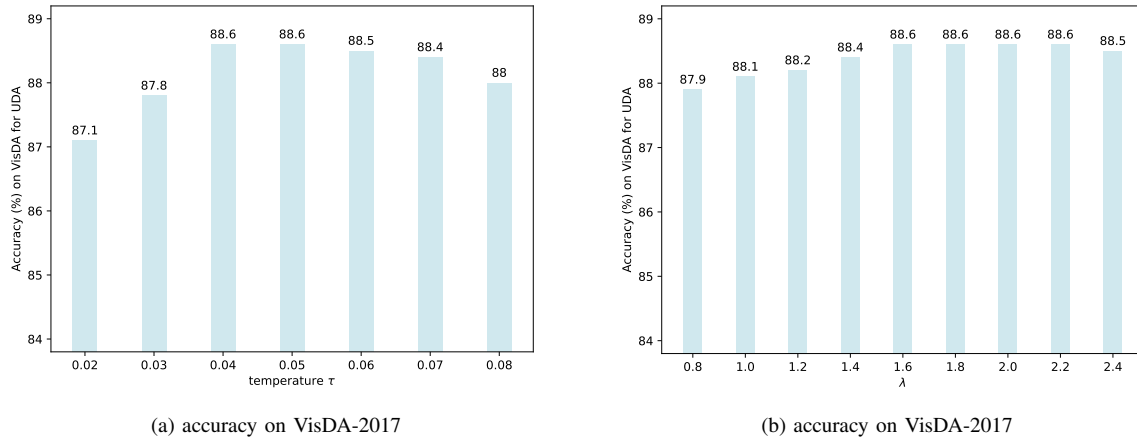


Fig. 4: Performance sensitivity of 2 hyper-parameters τ , λ in CDCL.

features from the same class to be closer. This highlights the effectiveness of CDCL in learning domain-invariant features.

V. CONCLUSION

In this paper, we presented CDCL, a simple yet effective framework for unsupervised domain adaptation. CDCL builds upon contrastive learning to align features for domain alignment and is suitable for both the standard UDA setting and the source data-free setting. In particular, given an image from one domain, we minimize its distance with respect to samples in the same class but from a different domain relative to all other cross-domain samples from different categories. Since labels are not available for the target domain, we generate pseudo labels using clustering. Further, we showed that CDCL can be easily adapted to the source data-free settings through considering classifier weights as class prototypes. We conducted extensive experiments on two widely used domain adaptation benchmarks and demonstrated that CDCL achieves state-of-the-art performance on both datasets.

ACKNOWLEDGMENT

This project was supported by NSFC under Grant No. 62102092.

REFERENCES

- [1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.
- [2] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," in *Advances in Neural Information Processing Systems*, 2007, pp. 137–144.
- [3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3722–3731.
- [4] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning*, 2015, pp. 97–105.
- [5] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.
- [6] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 213–226.
- [7] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International Conference on Machine Learning*, 2017, pp. 2208–2217.
- [8] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems*, 2006, pp. 513–520.
- [9] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- [10] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 443–450.
- [11] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.
- [12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.
- [14] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv:2003.04297*, 2020.
- [15] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*, 2020, pp. 6028–6039.
- [16] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," *arXiv:1710.06924*, 2017.
- [17] W. Zellingner, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," in *International Conference on Learning Representations*, 2017.
- [18] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 285–10 295.
- [19] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning*, 2019, pp. 7404–7413.
- [20] D. Li, Y. Lu, W. Wang, Z. Lai, J. Zhou, and X. Li, "Discriminative invariant alignment for unsupervised domain adaptation," *IEEE Transactions on Multimedia*, 2021.
- [21] H. Yan, Z. Li, Q. Wang, P. Li, Y. Xu, and W. Zuo, "Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2420–2433, 2019.
- [22] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.
- [23] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial

- domain adaptation,” in *Advances in Neural Information Processing Systems*, 2018, pp. 1647–1657.
- [24] S. Cui, S. Wang, J. Zhuo, C. Su, Q. Huang, and Q. Tian, “Gradually vanishing bridge for adversarial domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 455–12 464.
- [25] L. Hu, M. Kan, S. Shan, and X. Chen, “Unsupervised domain adaptation with hierarchical gradient synchronization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4043–4052.
- [26] Y. Zheng, X. Wang, G. Zhang, B. Xiao, F. Xiao, and J. Zhang, “Multi-kernel coupled projections for domain adaptive dictionary learning,” *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2292–2304, 2019.
- [27] X. Ma, T. Zhang, and C. Xu, “Deep multi-modality adversarial networks for unsupervised domain adaptation,” *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2419–2431, 2019.
- [28] X. Jin, C. Lan, W. Zeng, and Z. Chen, “Style normalization and restitution for domain generalization and adaptation,” *IEEE Transactions on Multimedia*, 2021.
- [29] R. Xu, G. Li, J. Yang, and L. Lin, “Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1426–1435.
- [30] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, “Domain-specific batch normalization for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7354–7362.
- [31] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, “Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3941–3950.
- [32] X. Chen, S. Wang, M. Long, and J. Wang, “Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation,” in *International Conference on Machine Learning*, 2019, pp. 1081–1090.
- [33] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, “Adversarial dropout regularization,” in *International Conference on Learning Representations*, 2018.
- [34] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, “Maximum classifier discrepancy for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.
- [35] Z. Lu, Y. Yang, X. Zhu, C. Liu, Y.-Z. Song, and T. Xiang, “Stochastic classifiers for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9111–9120.
- [36] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv:1807.03748*, 2018.
- [37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” in *Advances in Neural Information Processing Systems*, 2020, pp. 18 661–18 673.
- [38] Y. Ge, F. Zhu, D. Chen, R. Zhao, and h. Li, “Self-paced contrastive learning with hybrid memory for domain adaptive object re-id,” in *Advances in Neural Information Processing Systems*, 2020, pp. 11 309–11 321.
- [39] M. Thota and G. Leontidis, “Contrastive domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 2209–2218.
- [40] S. Dai, Y. Cheng, Y. Zhang, Z. Gan, J. Liu, and L. Carin, “Contrastively smoothed class alignment for unsupervised domain adaptation,” in *Proceedings of the Asian Conference on Computer Vision*, 2020, pp. 268–283.
- [41] M. Toldo, U. Michieli, and P. Zanuttigh, “Unsupervised domain adaptation in semantic segmentation via orthogonal and clustered embeddings,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1358–1368.
- [42] C. Park, J. Lee, J. Yoo, M. Hur, and S. Yoon, “Joint contrastive learning for unsupervised domain adaptation,” *arXiv:2006.10297*, 2020.
- [43] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, “Model adaptation: Unsupervised domain adaptation without source data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9641–9650.
- [44] K. Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” in *Advances in Neural Information Processing Systems*, 2016, pp. 1857–1865.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [47] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.



Rui Wang received the B.S. degree from Fudan University, Shanghai, China, in 2021. He is currently pursuing his Ph.D. degree in Computer Science at Fudan University. His research interests include video understanding, unsupervised learning and domain adaptation.



Zuxuan Wu received his Ph.D. in Computer Science from the University of Maryland with Prof. Larry Davis in 2020. He is currently an Associate Professor in the School of Computer Science at Fudan University. His research interests are in computer vision and deep learning. His work has been recognized by an AI 2000 Most Influential Scholars Honorable Mention in 2021, a Microsoft Research PhD Fellowship in 2019 and a Snap PhD Fellowship in 2017.



Zejia Weng received the B.S. degree from Fudan University, Shanghai, China in 2020. He is currently pursuing his M.S. degree in the School of Computer Science at Fudan University. His research interests are in computer vision and deep learning, especially large-scale video understanding.



Jingjing Chen is now a pre-tenured associate professor at the School of Computer Science, Fudan University. Before joining Fudan University, she was a postdoc research fellow at the School of Computing in the National University of Singapore. She received her Ph.D. degree in Computer Science from the City University of Hong Kong in 2018. Her research interest lies in diet tracking and nutrition estimation based on multi-modal processing of food images, including food recognition, cross-modal recipe retrieval.



Guo-Jun Qi (M14-SM18-F22) is the Chief Scientist and the director of Seattle Research Center in the OPPO Research USA since 2021. Before that, he is the Chief Scientist who led and oversaw an international R&D team in the domain of multiple intelligent cloud services, including smart cities, visual computing service, medical intelligent service, and connected vehicle service at Futurewei since 2018. He was a faculty member in the Department of Computer Science and the director of MACHine Perception and LEarning (MAPLE) Lab at the University of Central Florida since 2014. Prior to that, he was also a Research Staff Member at IBM T.J. Watson Research Center, Yorktown Heights, NY.



Yu-Gang Jiang received the Ph.D. degree in Computer Science from City University of Hong Kong in 2009 and worked as a Postdoctoral Research Scientist at Columbia University, New York during 2009-2011. He is currently a Professor and Dean at School of Computer Science, Fudan University, Shanghai, China. His research lies in the areas of multimedia, computer vision and trustworthy AI. His work has led to many awards, including the inaugural ACM China Rising Star Award, the 2015 ACM SIGMM Rising Star Award, and the research award for excellent young scholars from NSF China.