IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received November 11, 2021, accepted December 14, 2021, date of publication December 20, 2021,
date of current version December 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3136567

Multi-View Collaborative Learning for Semi-Supervised Domain Adaptation

BA HUNG NGO, (Graduate Student Member, IEEE), JU HYUN KIM, YEON JEONG CHAE, AND SUNG IN CHO, (Member, IEEE)
Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea

Corresponding author: Sung In Cho (csi2267@dongguk.edu)

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1C1C1009662) and Korea Institute of Police Technology (KIPoT) grant funded by the Korea government (KNPA) (No.202100200902, AI driving ability test standardization and evaluation process development).

ABSTRACT Recently, Semi-supervised Domain Adaptation (SSDA) has become more practical because a small number of labeled target samples can significantly boost the empirical target performance when using SSDA. Several current methods focus on prototype-based alignment to achieve cross-domain invariance in which the labeled samples from the source and target domains are concatenated to estimate the prototypes. The model is then trained to assign the unlabeled target data to the prototype within the same class. However, such methods fail to exploit the advantage of using few labeled target data because the labeled source data dominate the prototypes in the supervision process. Moreover, a recent method (Yang *et al.*, 2021) showed that concatenating source and target samples for training can damage the semantic information of representations, which degrades the trained model's ability to generate discriminative features. To solve these problems, in this paper, we divide labeled source and target samples into two subgroups for training. One group includes a large number of labeled source samples, and the other obtains a few labeled target samples. Then, we propose a novel SSDA framework that consists of two models. A model trained on the group that has the labeled source samples to provide an ''inter-view'' on the unlabeled target data is called the inter-view model. A model trained on a few labeled target samples that provides an ''intra-view'' of the unlabeled target data is called the intra-view model. Finally, both of these models collaborate to fully exploit information on the unlabeled target data. To the best of our knowledge, our proposed method achieves the state-of-the-art classification performance of SSDA in extensive experiments conducted on several visual benchmark domain adaptation datasets that utilize the advantages of multiple views and collaborative training.

INDEX TERMS Semi-supervised domain adaptation, classification, multiple views, collaborative learning.

I. INTRODUCTION

With large-scale labeled data samples and the growth of computing power, supervised learning methods have shown empirical results in various computer vision applications such as image classification [2]–[4], image semantic segmentation [5]–[7], and object detection [8]–[10]. These methods assumed that the training and the test sets come from the same distribution; however, the training data (a source domain) and test data (a target domain) in most real-world applications are related but follow different distributions. Therefore, if a model trained on the source domain directly performs on the target domain, it poorly generalizes on the domain due to domain shift, which degrades the accuracy of the target application (most representatively, in image classification) for the target domain. To solve this problem, DANN [11] introduced an adversarial learning strategy to minimize the difference between the source and target distributions to achieve domain-invariant knowledge across domains. In the DA setting, a model is often trained on plentiful labeled data from the source domain to successfully perform the target task on the target domain, with little-to-no labeled data with a different distribution. Depending on the availability of labeled target samples during training, DA can be categorized as unsupervised domain adaptation (UDA) [11]–[16] or semi-supervised domain adaptation (SSDA) [17]–[25].

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

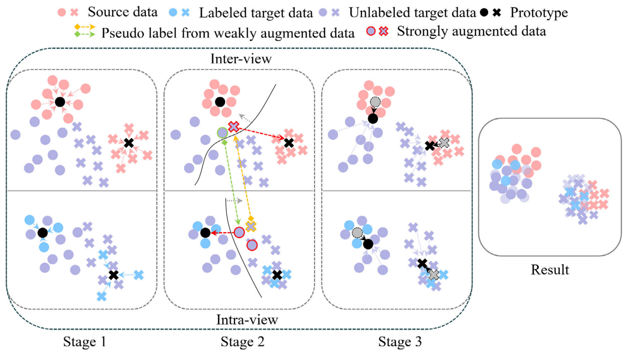


FIGURE 1. Inter-view and intra-view models to consider the correlation between labeled source data and labeled target data with unlabeled target data.

UDA [11]–[16] attempts to achieve domain-invariant representation for image classification tasks by minimizing the distribution between source and target domains; however, these works still have room for performance improvement because they focus on domain-invariant features without considering specific representations in each class. Thus, recently, SSDA [17]–[25] for image classification has received significant attention, in which the model is trained to give large amounts of labeled source data and has access to a few labeled target data. Similar to MME [17], UODA [20] is the SSDA method developed based on prototypes that are estimated from given labeled source and target samples. Then, the model of these methods is trained to encourage the unlabeled features clustered around the estimated prototypes. However, these methods cannot alleviate the domain gap because the large amount of source data, much larger than the target data, dominates when creating the class prototypes. In addition, [1] argued that, in many real-world applications, samples of different classes are often similarly expressed in the feature space. Therefore, when a model is trained on integrated source and target samples, the discriminative feature representation ability can be reduced because the semantic information of representations can be damaged. Therefore, in this paper, instead of integrating all labeled source and target domains for training, as in MME [17], UODA [20], and STar [24], we divided the labeled samples into two subsets to train two different models. The model trained on the labeled source samples is called an "inter-model," and the model trained on the labeled target domain is called an "intra-model." These models provide different views to predict unlabeled target data.

Although the inter-view model is trained on large amounts of labeled data from the source domain, it may not provide satisfactory image classification accuracy on unlabeled target data due to the domain shift problem. The intra-view model has an overfitting problem for classification of the target data because this model is trained on a few labeled target samples, which cannot generalize the unlabeled target data. Therefore, to solve the above problems, we proposed a novel framework called Multiple-view Collaborative Learning (MVCL). As shown in Figure 1, the training in our approach progresses in three stages. First, the inter- and intra-view models are trained on the labeled source and labeled target samples, respectively, and work as an inter-view and intra-view to extract information from the unlabeled target samples in the second stage. In this stage, the inter- and intra-view models will alternate, offering their pseudo labels selected from the highest prediction scores to teach the other model. This process, called collaborative learning (Co-learning), allows both models to exchange their mutually complementary information to make consistent predictions on unlabeled target data. Finally, we used adversarial learning via the minimax entropy strategy [17] to encourage the unlabeled target features clustered around the prototypes in the third stage.

Our main contributions are summarized as follows:

- To solve the bias in learning problem in the supervision phase of previous SSDA works, we divided the labeled samples into two subsets which are labeled source samples and labeled target samples instead of integrating them into one set, as in [17], [20], and [24]. We use two models to simultaneously extract unique features on these two subsets. The first model trained on labeled source samples is called the inter-view model because it provides an inter-view on the unlabeled target data. Similarly, the second model trained on labeled target samples is called the intra-view model because it corresponds to an intra-view on the unlabeled target data.
- Then, we successfully unified collaborative learning and domain adaptation into a framework for SSDA. Specifically, each individual model obtained partial information of the target data. The inter- and intra-view models then exchange their knowledge to comprehensively represent the target data via collaborative learning. Furthermore, MVCL can extract intrinsic information from each view and adaptively balance it complementarily to achieve consistency among different views. Therefore, it can alleviate the problem of feature degeneration and enhance the reasonability of using a consensus representation for multiple views.
- We conducted extensive MVCL experiments on several domain adaptation datasets, including *Office-31*, *Office-Home*, *VisDA2017*, and *DomainNet*, to show that our method achieved SOTA classification performance in SSDA.

II. RELATED WORK

In this section, we review the related works in SSDA and multiple views.

A. SEMI-SUPERVISED DOMAIN ADAPTATION

Recently, the SSDA approach has received a lot of attention [17]–[25], where a few labeled target samples work as leverage to improve the domain adaptation performance for image classification. SSDA via minimax entropy (MME) [17] is the first method that aligns the representations of source

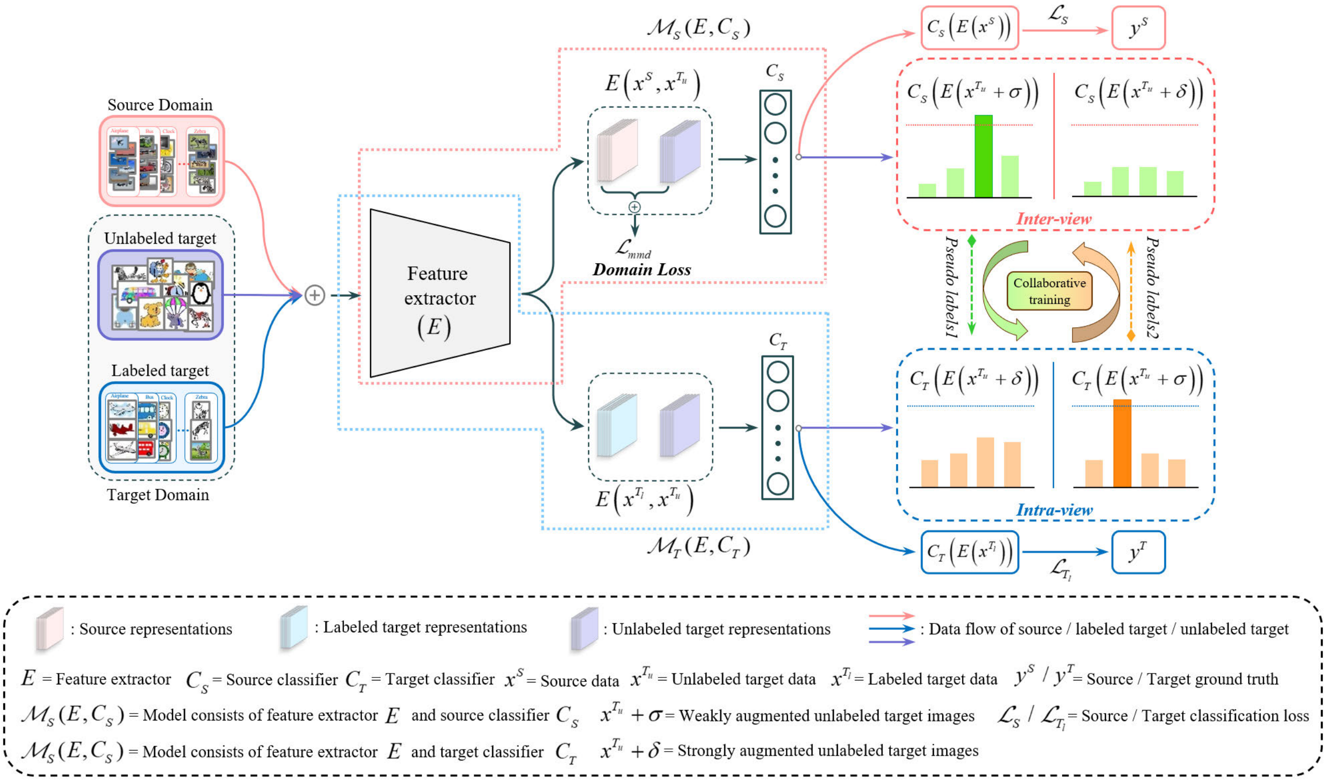

$\mathcal{M}_S(E,C_S)$
$E(x^S, x^{T_u})$
C_S
$C_S(E(x^S))$
$\mathcal{L}_S$
y^S
$C_S(E(x^{T_u}+\sigma))$
$C_S(E(x^{T_u}+\delta))$
Inter-view
Source Domain
Unlabeled target
Labeled target
Target Domain
Feature extractor (E)
$\mathcal{L}_{mmd}$
Domain Loss
Pseudo labels1
Collaborative training
Pseudo labels2
C_T
$C_T(E(x^{T_u}+\delta))$
$C_T(E(x^{T_u}+\sigma))$
Intra-view
$E(x^{T_l}, x^{T_u})$
$\mathcal{M}_T(E,C_T)$
$C_T(E(x^{T_l}))$
$\mathcal{L}_{T_l}$
y^T
: Source representations
: Labeled target representations
: Unlabeled target representations
: Data flow of source / labeled target / unlabeled target
E = Feature extractor
C_S = Source classifier
C_T = Target classifier
x^S = Source data
x^{T_u} = Unlabeled target data
x^{T_l} = Labeled target data
y^S / y^T = Source / Target ground truth
$\mathcal{M}_S(E,C_S)$ = Model consists of feature extractor E and source classifier C_S
$x^{T_u}+\sigma$ = Weakly augmented unlabeled target images
$\mathcal{L}_S$ / $\mathcal{L}_{T_l}$ = Source / Target classification loss
$\mathcal{M}_S(E,C_S)$ = Model consists of feature extractor E and target classifier C_T
$x^{T_u}+\delta$ = Strongly augmented unlabeled target images

FIGURE 2. Illustration of the proposed method's architecture. The inter-view model $\mathcal{M}_S(E, C_S)$ as indicated by the dotted red box is trained on a lot of labeled source data and provides an inter-view on the target domain. However, due to the discrepancy between two domains, if this model is directly applied to the target domain, it cannot achieve satisfactory classification performance on the target domain. The intra-view model $\mathcal{M}_T(E, C_T)$ as indicated by the dotted blue box is trained on a few labeled target samples and provides an intra-view on the target domain. Since this model is trained on limited labeled samples, it cannot generalize on the target domain. Therefore, we proposed a co-learning algorithm that allows the two models to exchange their knowledge to achieve the generalization view on the target domain and mitigate the negative transfer due to domain shift.

and target domains using adversarial learning called minimax entropy. Specifically, in this approach, they train their framework with concatenated labeled source and target samples to create prototypes represented by the weight vectors of a classifier. Then, in the first adversarial training step, they re-weight these vectors by maximizing the entropy of unlabeled target samples to estimate domain-invariant prototypes. In the second step of adversarial training, they use the minimizing entropy strategy on the unlabeled target samples to update the feature extractor and encourage the target features clustered around the prototypes. The classification accuracy in the target domain significantly increases when applying the MME method. However, recent studies [18] and [19] show that this method still has room for improvement in terms of classification accuracy in the target domain. APE [18] argued that only unlabeled target instances that have features that close the relationship with those of the labeled target samples move to the estimated prototypes. Other unlabeled targets are misaligned, which leads to an intra-domain discrepancy issue in the target domain. APE proposed a method that includes attention, perturbation, and exploration schemes to solve this problem; however, their method does not provide a solution for when the feature representations of the trained model are dominated by a large number of labeled source samples in the supervision process. Another method [19] successfully significantly improved the classification accuracy in the target domain by developing a novel mapping function (MAP-F) that could solve the bias in the learning of MME. To alleviate the bias problem caused by the unbalanced number of available source and target samples, MAP-F divides the labeled samples into two subgroups: labeled source and labeled target samples. Then, first, they train a model with labeled source samples to obtain a well-organized source distribution. Next, they introduce a novel mapping function to minimize the distance of target class centroids from the source samples within a class to reproduce the well-clustered features of the source domain in the target domain. Although MAP-F showed outstanding classification accuracy in the target domain, it uses two feature extractors simultaneously that require a lot of training time.

B. MULTI-VIEW DOMAIN ADAPTATION

Multi-view representation learning emerges as a promising direction in machine learning for classification. The key that plays an essential role in this approach is to extract multiple elements of knowledge from the input data by taking multiple views and then integrating them to find a

TABLE 1. Important notations.

Notation	Description
S/T	The source/target domain
$\mathcal{M}_S/\mathcal{M}_T$	The inter-view/intra-view model
$E(.)$	The common feature extractor
C_S/C_T	The source/target classifier
$\mathcal{L}_{MMD}$	The domain alignment loss
(x_i^S, y_i^S)	The *i*-th element source sample and its label
$(x_i^{T_l}, y_i^{T_l})$	The *i*-th element target sample and its label
$x_i^{T_u}$	The *i*-th element sample of the unlabeled target data
N_S	The number of samples in the source domain
N_{T_u}	The number of unlabeled target samples
N_{T_l}	The number of labeled target samples
D_S	A set of source samples
D_{T_l}	A set of labeled target samples
D_{T_u}	A set of unlabeled target samples

good representation of the input data. Thus, a few domain adaptation methods [26]–[29] have successfully improved the learning performance by integrating multiple views for training. Multi-View Transfer Learning with a Large Margin Approach (MVTL-LM) [26] introduces multiple views for a DA framework. This approach uses two views to exploit the source and target information; one view extracts the labeled source information and is then used to construct another view for the target samples by minimizing the consistency loss between the two views. Multi-view Discriminant Transfer (MDT) [27] is a multiple views-based approach for DA. The main concept of this method is to determine the optimal discriminative weight vectors for each view by maximizing the correlation between the two views while simultaneously minimizing the domain discrepancy loss. The Maximum Classifier Discrepancy (MCD) method [28] consists of two classifiers trained on labeled source samples where both views classify the unlabeled target data. In this method, they introduce a discrepancy loss to maximize the disagreement between predictions of these classifiers to classify the target samples far from the support of the source domain. In contrast, the feature extractor learns to extract target representations that are near the support of the source samples to minimize their discrepancy. [29] provides a short survey that discusses and analyzes the framework for the unification of multiple-view learning and domain adaptation.

III. PROPOSED METHOD

In this section, we introduce the definition of the problem in our method and notations of SSDA. Then, we present the training processes with the loss functions for the proposed method.

A. DEFINITION OF THE PROBLEM AND NOTATIONS

In the SSDA setting, we have the set of the source domain denoted as $D_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$ where N_S is the number of labeled source (S) samples, $y_i^S \in \mathbb{R}^K$ is the label of the samples x_i^S and K is the number of classes. In addition, the set of labeled target samples (T_l) is denoted as $D_{T_l} = \{(x_i^{T_l}, y_i^{T_l})\}_{i=1}^{N_{T_l}}$, where N_{T_l} is the number of labeled target samples, $y_i^{T_l} \in \mathbb{R}^K$ is the label of sample $x_i^{T_l}$. The set of unlabeled target samples (T_u) is denoted as $D_{T_u} = \{(x_i^{T_u})\}_{i=1}^{N_{T_u}}$, where N_{T_u} is the number of unlabeled target samples and $N_{T_u} \gg N_{T_l}$. Table 1 lists all important symbols.

The goal of SSDA is to design a framework that not only formally reduces the domain shift between both domains but also tries to achieve class-wise matching by leveraging a few labeled samples in each class of the target domain.

As shown in Figure 2, the proposed method indicates the inter-view model $\mathcal{M}_S(E, C_S)$ using a dotted red box and the intra-view model $\mathcal{M}_T(E, C_T)$ using a dotted blue box. These two models have classifiers C_S and C_T and share feature extractor E. The training process in the proposed method has three stages: the first uses supervised learning, the second uses collaborative learning (Co-learning) on the unlabeled target domain, and the third uses the minimax entropy strategy, as shown in Figure 1. Each stage is explained in detail in the following subsections.

B. SUPERVISED TRAINING

In the first stage, the two models $\mathcal{M}_S(E, C_S)$ and $\mathcal{M}_T(E, C_T)$ are trained using the labeled data. Specifically, the labeled source samples are fed into the inter-view model $\mathcal{M}_S(E, C_S)$, which consists of a shared CNN-based feature extractor E that obtains their corresponding representations. Then, these features are categorized by the task-specific classifier C_S by minimizing the standard cross-entropy loss on the ground-truth labels as follows:

$$\mathcal{L}_S = -\frac{1}{N_S}\sum_{i=1}^{N_S}\sum_{k=1}^{K}\mathbb{1}_{[k=y_i^S]}\log\left(C_S(E(x_i^S))\right), \tag{1}$$

where $\mathbb{1}_{[.]} \in [0, 1]$ is the indication function that receives a value of 1 or 0 when the input [] is true or false, respectively. The *i*-th source image x_i^S has a label $y_i^S = k \in K$.

Similarly, the feature extractor E and target classifier C_T of the intra-view model $\mathcal{M}_T(E, C_T)$ are trained using the standard cross-entropy over the limited labeled target samples as follows:

$$\mathcal{L}_{T_l} = -\frac{1}{N_{T_l}}\sum_{i=1}^{N_{T_l}}\sum_{k=1}^{K}\mathbb{1}_{[k=y_i^{T_l}]}\log\left(C_{T_l}(E(x_i^{T_l}))\right). \tag{2}$$

Consequently, the parameters of the shared feature extractor E are optimized by minimizing the following loss:

$$\mathcal{L}_{E_{cls}} = \mathcal{L}_S + \mathcal{L}_{T_l}. \tag{3}$$

The parameters of the classifiers C_S and C_T are updated by adopting Eqs. (1) and (2), respectively. The model $\mathcal{M}_S$

trained on the labeled source data provides an inter-view aspect while the model $\mathcal{M}_T$ trained on a few labeled target samples works as an intra-view when they are used to extract information from the unlabeled target data.

C. DOMAIN ALIGNMENT

The domain shift [14] means that the inter-view model trained according to (1), which has only source label information, cannot provide satisfactory classification accuracy on the target domain. To solve this, we use the maximum mean discrepancy (MMD) approach [30] to minimize the distance between the source and target distributions. Noticeably, the MMD is a useful metric that compares the distributions of data in the source and target domains by mapping the data to a high-dimensional embedding in Reproducing Kernel Hilbert Space (RKHS). Regarding the source distribution (**P**) and target distribution (**Q**) for domain adaptation, the MMD between **P** and **Q** is equivalent to the distance between the means of the samples in the source and target in RKHS, using the mapping function ϕ. Through the MMD, the inter-view model $\mathcal{M}_S$ is trained to minimize the distance between **P** and **Q** so that it can achieve domain-invariant feature representations. The MMD loss of the source and target can be estimated as follows:

$$\mathcal{L}_{MMD} = \left\| \frac{1}{N_S} \sum_{i=1}^{N_S} \phi(x_i^S) - \frac{1}{N_{T_u}} \sum_{i=1}^{N_{T_u}} \phi(x_i^{T_u}) \right\|_{\mathcal{H}}^2, \qquad (4)$$

where $\mathcal{H}$ is the RKHS and $\phi(.) \in \mathcal{H}$ is the mapping function of $\mathcal{X}$ to the RKHS. The MMD loss can be expressed in terms of a kernel method. We exploit the Gaussian kernel function that satisfies the condition of the MMD to project samples of the source and target domains instead of the mapping function. The impact of $\mathcal{L}_{MMD}$ on the classification performance in the target domain is reported in the ablation study in Section IV. E.

D. CONSISTENCY CLASS ALIGNMENT WITH MULTI-VIEW CO-LEARNING

For the inter-view, after finishing stage 1, the model $\mathcal{M}_S$ holds rich information from the source domain to transfer to the target domain. However, this model lacks information from the target domain. For the intra-view, the model $\mathcal{M}_T$ poorly generalizes on the target domain because it has only been trained on limited labeled target samples. Therefore, we solve this problem in the second stage using multiple views with co-learning. Specifically, the multiple views-based strategy lets us fully exploit the information of the unlabeled target data; co-learning encourages these two models to exchange knowledge to alleviate each of their shortcomings. For example, the inter-view model $\mathcal{M}_S$ generates two predictions for two augmented versions of an unlabeled image: a weakly augmented version $x_i^{T_u} + \sigma$ and a strongly augmented version $x_i^{T_u} + \delta$. The weak augmentation used a simple transformation method, such as random cropping or flipping. The strong augmentation used RandAugment [31], which randomly selects from a list of 14 various augmentation schemes, such as rotations, translations, and color/brightness enhancements. The two predictions by $\mathcal{M}_S$ over the weakly and strongly augmented images are computed as follows:

$$\begin{aligned} p_S^w\left(x_i^{T_u}\right) &= softmax\big(C_S(E(x_i^{T_u} + \sigma))\big), \\ p_S^{str}\left(x_i^{T_u}\right) &= softmax\big(C_S(E(x_i^{T_u} + \delta))\big), \end{aligned} \qquad (5)$$

where $p_S^w(x_i^{T_u})$ and $p_S^{str}(x_i^{T_u})$ are the predictions of the weak and strong augmentation versions of an unlabeled target image $x_i^{T_u}$ generated by $\mathcal{M}_S$. Similarly, $\mathcal{M}_T$ also provides its two predictions $p_T^w(x_i^{T_u})$ and $p_T^{str}(x_i^{T_u})$ on the same input image.

Then, the co-learning process aims to enforce the consistency regularization that is conducted by minimizing the cross-entropy of the selected pseudo label from the inter-view prediction ($PS = \max_{PS} p_S^w(x_i^{T_u})$) and each intra-view prediction $p_T^{str}(x_i^{T_u})$ of the strongly augmented image. This entire process is conducted as follows: in the inter-view, the model $\mathcal{M}_S$ offers its pseudo labels selected from the highest confident prediction of the weakly augmented version of an unlabeled target image, $x_i^{T_u} + \sigma$. This is then converted into a one-hot encoded label for the calculation of cross-entropy, with the prediction of the strongly augmented version of the same unlabeled target image $x_i^{T_u} + \delta$, as predicted by the model $\mathcal{M}_T$. Simultaneously, the model $\mathcal{M}_T$ provides its pseudo labels are generated over a weak augmentation image to match with the prediction of the strongly augmented transformation predicted by the model $\mathcal{M}_S$ over the same unlabeled target image. Finally, the generalization performance on the target domain is improved by integrating both views' complementary information.

The incorrect pseudo labels can negatively impact the performance of the models in the target domain; therefore, only the prediction of unlabeled target samples that have a high probability over a given threshold value is selected as a pseudo label, $max p_j^w(x_i^{T_u}) > \tau$, where τ is the threshold value and j is the index for the domain. The consistency losses between $\mathcal{M}_S$ and $\mathcal{M}_T$ are calculated as follows:

$$\mathcal{L}_{inter} = \sum_{i=1}^{N_{T_u}} \mathbb{1}\left[\max p_T^w\left(x_i^{T_u}\right) > \tau_{intra}\right] \hat{y}_i^{intra} \log\left(p_S^{str}(x_i^{T_u})\right), \qquad (6)$$

$$\mathcal{L}_{intra} = \sum_{i=1}^{N_{T_u}} \mathbb{1}\left[\max p_S^w\left(x_i^{T_u}\right) > \tau_{inter}\right] \hat{y}_i^{inter} \log\left(p_T^{str}(x_i^{T_u})\right), \qquad (7)$$

where $\mathbb{1}_{[.]}$ is an indication function and $\mathcal{M}_T$ offers $\hat{y}_i^{inter} = argmax\big(p_S^w(x_i^{T_u})\big)$ while $\mathcal{M}_T$ provides $\hat{y}_i^{intra} = argmax\big(p_T^w(x_i^{T_u})\big)$, which are converted into a one-hot encoded label to use as a pseudo label for supervised learning. τ_{inter} and τ_{intra} are the threshold values used to select pseudo labels of $\mathcal{M}_S$ and $\mathcal{M}_T$. These are studied in the ablation study in Section IV. E.

The loss for the multi-view co-learning process is used to update the parameters of feature extractor E, which are calculated as follows:

$$\mathcal{L}_{E_{Co}} = \mathcal{L}_{inter} + \mathcal{L}_{intra}. \tag{8}$$

The classifiers C_S and C_T are trained using (6) and (7), respectively.

In the third stage, the parameters of the feature extractor E, classifiers C_S and C_T are updated using the minmax strategy following [17]. Deep convolution neural networks (CNNs), such as Alexnet [32], VGG16 [33], or ResNet-34 [34], are used as the backbone network of feature extractor E. Each classifier consists of two fully connected layers (FCs). The last linear layer is replaced by a K-way linear classification (K represents the number of classes) that aims to exploit the cosine similarity-based classifier architecture. Therefore, they are also called "cosine classifiers." Each cosine classifier is presented by K class-specific weight vectors $W = [w_1, w_2, \ldots, w_K]$, where each weight vector w_i represents the i-th class prototype. The final probability of each cosine classifier is denoted by $p(x) = softmax\left(\frac{1}{T}W^T f\right)$ where $\frac{1}{T}W^T f$ is the output of this classifier, T is a fixed temperature (0.05), and f is the input feature. The minimax entropy strategy is conducted as follows: In an entropy maximization process, each cosine classifier is trained such that each w_i is similar to the generated target features **f** to achieve domain-invariant prototype generation. Then, in the entropy minimization process, feature extractor E is trained to achieve discriminative features on the unlabeled target data by assigning the extracted features of the unlabeled target samples to a certain prototype.

The minimax strategy is applied to both the inter- and intra-view models. For the inter-view model, the conditional entropy loss of the unlabeled target samples to the target-clustering classifier is C_S determined as follows:

$$H_{inter} = -\mathbb{E}_{x_i^{T_u} \sim D_{T_u}} \sum_{k=1}^{K} p_{inter}(y = k \mid x_i^{T_u}) \log p_{inter}(y = k \mid x_i^{T_u}), \tag{9}$$

where $p_{inter}(y = k \mid x_i^{T_u})$ represents the probability of $x_i^{T_u}$ belonging to class k predicted by the inter-view model.

Similarly, for the intra-view model, the conditional entropy loss of the unlabeled target samples to the target-clustering classifier C_T is determined as follows:

$$H_{intra} = -\mathbb{E}_{x_i^{T_u} \sim D_{T_u}} \sum_{k=1}^{K} p_{intra}(y = k \mid x_i^{T_u}) \log p_{intra}(y = k \mid x_i^{T_u}), \tag{10}$$

TABLE 2. **Description of datasets.**

Datasets	Domains	Type	Instance	Classes
Visda2017	Synthetic Real	Object	152,397 55,388	12
Office-31	DSLR Webcam Amazon	Object	498 795 2,817	31
Office-Home	Artistic Clip Art Product Real-World	Object	2,427 4,365 4,439 4,357	65
DomainNet	Clipart Infograph Painting Quickdraw Real Sketch	Object	48,837 53,201 75,759 172,500 175,327 70,386	345

where $p_{intra}(y = k \mid x_i^{T_u})$ is the probability that the intra-view model predicts that $x_i^{T_u}$ belongs to class k.

The total loss for feature extractor E over three stages is computed as follows:

$$\mathcal{L}_E = \mathcal{L}_{E_{cls}} + \mathcal{L}_{MMD} + \mathcal{L}_{E_{Co}} + \lambda H_{inter} + \lambda H_{intra}, \tag{11}$$

where λ is a balancing parameter and was set as in [17].

The total losses for classifier C_S and C_T are calculated as follows:

$$\begin{aligned} \mathcal{L}_{C_S} &= \mathcal{L}_S + \mathcal{L}_{inter} - \lambda H_{inter}, \\ \mathcal{L}_{C_T} &= \mathcal{L}_{T_l} + \mathcal{L}_{intra} - \lambda H_{intra}. \end{aligned} \tag{12}$$

E. INFERENCE ON UNLABELED TARGET DATA

The prediction on the unlabeled target data is calculated by taking an ensemble value of the *softmax* outputs for the two models as follows:

$$y_{pred} = argmax\Big(C_S\big(E(x^{T_u})\big) + C_T\big(E(x^{T_u})\big)\Big). \tag{13}$$

IV. EXPERIMENT

In this section, we show the experimental details of our investigation of the proposed method's efficiency for the SSDA setting. We implemented the one-/three-shot settings as in [17] on four commonly used domain adaptation benchmark datasets, namely *Office-31* [35], *Office-Home* [36], *VisDA2017* [37], and *DomainNet* [38]. The transfer tasks are denoted as: source domain→target domain.

A. DATASETS

- The *Office-31* dataset has three different domains: *Amazon* (**A**), *Webcam* (**W**), and *DSLR* (**D**), which have approximately 2,800, 800, and 500 images, respectively, and 31 classes. Following [17], we evaluated our method in two cases, **D**→**A** and **W**→**A**.
- *Office-Home* is a new visual domain adaptation dataset that contains images from four different domains:

TABLE 3. Classification accuracy (%) on the *VisDA2017* dataset with over 12 tasks for the 3-shot using ResNet34 as the backbone network. The highest and second-highest accuracies are demonstrated by red and blue, respectively.

Method	Plane	Bicycle	Bus	Car	House	Knife	Motorcycle	Person	Plant	Skateboard	Train	Struck	Mean
S+T	86.0	55.6	62.4	55.0	60.6	55.5	79.1	52.2	68.1	59.9	82.2	38.0	62.9
ENT	90.2	54.4	72.5	76.1	86.3	95.5	89.1	76.4	84.8	47.9	87.7	7.1	72.3
UODA	87.7	49.9	87.5	73.0	81.4	81.4	91.3	72.9	89.0	63.9	81.6	15.5	72.9
MME	87.3	65.1	74.0	72.1	83.2	89.8	93.0	78.2	90.8	70.5	87.0	34.0	77.1
APE	94.9	64.9	77.7	80.1	93.5	81.0	94.8	82.4	90.6	81.2	87.1	5.6	77.8
MAP-F	97.3	85.7	87.1	75.0	96.4	89.0	94.3	85.3	95.6	87.9	89.8	57.4	86.7
Ours	97.8	88.9	89.4	79.4	97.2	93.3	95.6	84.2	96.2	90.5	93.0	61.5	88.9

TABLE 4. Classification accuracy (%) on the *DomainNet* dataset for the 1-shot and 3-shot settings using Alexnet and ResNet34 as the backbone networks. The highest and second-highest accuracies are illustrated in red and blue, respectively.

Net	Method	R→C		R→P		P→C		C→S		S→P		R→S		P→R		Mean	
		1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
AlexNet	S+T	43.3	47.1	42.4	45.0	40.1	44.9	33.6	36.4	35.7	38.4	29.1	33.3	55.8	58.7	40.0	43.4
	DANN	43.3	46.1	41.6	43.8	39.1	41.0	35.9	36.5	36.9	38.9	32.5	33.4	53.5	57.3	40.4	42.4
	ENT	37.0	45.5	35.6	42.6	26.8	40.4	18.9	31.1	15.1	29.6	18.0	29.6	52.2	60.0	29.1	39.8
	MME	48.9	55.6	48.0	49.0	46.7	51.7	36.3	39.4	39.4	43.0	33.3	37.9	56.8	60.7	44.2	48.2
	BiAT	54.2	58.6	49.2	50.6	44.0	52.0	37.7	41.9	39.6	42.1	37.2	42.0	56.9	58.8	45.5	49.4
	APE	47.7	54.6	49.0	50.5	46.9	52.1	38.5	42.6	38.5	42.2	33.8	38.7	57.5	61.4	44.6	48.9
	PAC	55.4	61.7	54.6	56.9	47.0	59.8	46.9	52.9	38.6	43.9	38.7	48.2	56.7	59.7	48.3	54.7
	CDAC	56.9	61.4	55.9	57.5	51.6	58.9	44.8	50.7	48.1	51.7	44.1	46.7	63.8	66.8	52.1	56.2
	Ours	57.1	61.2	58.3	57.2	52.0	58.8	46.6	50.0	50.9	54.6	44.7	45.5	69.0	69.9	54.1	56.7
ResNet34	S+T	55.6	60.0	60.6	62.2	56.8	59.4	50.8	55.0	56.0	59.5	46.3	50.1	71.8	73.9	56.9	60.0
	DANN	58.2	59.8	61.4	62.8	56.3	59.6	52.8	55.4	57.4	59.9	52.2	54.9	70.3	72.2	58.4	60.7
	ENT	65.2	71.0	65.9	69.2	65.4	71.1	54.6	60.0	59.7	62.1	52.1	61.1	75.0	78.6	62.6	67.6
	MME	70.0	72.2	67.7	69.7	69.0	71.7	56.3	61.8	64.8	66.8	61.0	61.9	76.1	78.5	66.4	68.9
	BiAT	73.0	74.9	68.0	68.8	71.6	74.6	57.9	61.5	63.9	67.5	58.5	62.1	77.0	78.6	67.1	69.7
	UODA	72.7	75.4	70.3	71.5	69.8	73.2	60.5	64.1	66.4	69.4	62.7	64.2	77.3	80.8	68.5	71.2
	APE	70.4	76.6	70.8	72.1	72.9	76.7	56.7	63.1	64.5	66.1	63.0	67.8	76.6	79.4	67.8	71.7
	ELP	72.8	74.9	70.8	72.1	72.0	74.4	59.6	4.3	66.7	69.7	63.3	64.9	77.8	81.0	69.0	71.6
	STar	74.1	77.1	71.3	73.2	71.0	75.8	63.5	67.8	66.1	69.2	64.1	67.9	80.0	81.2	70.0	73.2
	PAC	74.9	78.6	73.0	74.3	72.6	76.0	65.8	69.6	67.9	69.4	68.7	70.2	76.7	79.3	71.4	73.9
	MAP-F	75.3	77.0	74.0	75.0	74.3	77.0	65.8	69.5	73.0	73.3	67.5	69.2	81.7	83.3	73.1	74.9
	CDAC	77.4	79.6	74.2	75.1	75.5	79.3	67.6	69.9	71.0	73.4	69.2	72.5	80.4	81.9	73.6	76.0
	Ours	78.8	79.8	76.0	77.4	78.0	80.3	70.8	73.0	75.1	76.7	72.4	74.4	82.4	85.1	76.2	78.1

TABLE 5. Classification accuracy (%) on the *DomainNet* dataset for the 5-shot and 10-shot settings using ResNet34 as the backbone network. We illustrate the highest and second-highest accuracies in red and blue, respectively.

Method	R→C		R→P		P→C		C→S		S→P		R→S		P→R		Mean	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
S+T	64.5	68.5	63.1	66.4	64.2	69.2	59.2	64.8	60.4	64.2	56.2	60.7	75.7	77.3	63.3	67.3
DANN	63.7	70.0	62.9	64.5	60.5	64.0	55.0	56.9	59.5	60.7	55.8	60.5	72.6	75.9	61.4	64.6
ENT	77.1	79.0	71.0	72.9	75.7	78.0	61.9	68.9	66.2	68.4	64.6	68.1	81.1	82.6	71.1	74.0
MME	75.5	77.1	70.4	71.9	74.0	76.3	65.0	67.0	68.2	69.7	65.5	67.8	79.9	81.2	71.2	73.0
APE	77.7	79.8	73.0	75.1	76.9	78.9	67.0	70.5	71.4	73.6	68.8	70.8	80.5	82.9	73.6	75.9
MAP-F	78.4	82.6	76.2	77.6	78.0	82.3	71.5	74.5	74.1	76.7	70.9	73.0	83.9	86.8	76.1	79.1
CDAC	80.8	83.1	75.3	77.2	79.9	81.7	72.1	74.3	74.7	76.3	72.9	74.6	83.2	84.7	76.9	78.9
Ours	81.1	83.0	78.2	79.2	81.7	82.7	74.7	76.0	77.2	78.1	74.6	75.9	86.3	87.0	79.1	80.3

Real (**R**), *Product* (**P**), *Clipart* (**C**), and *Art* (**A**), each of which has 65 categories. We used 1-shot and 3-shot splits and evaluated the adaptation performance on the target domain for 12 pairs (source→target) as in [17]. We evaluated the proposed method on 12 transfer tasks.

- *VisDA2017* is a challenging synthetic-to-real dataset that consists of 152,397 *Synthetic* images as the source domain and 5,388 *Real* images as the target domain; they share 12 categories. We randomly selected one and three real images from each of the 12 classes corresponding to the 1-shot and 3-shot SSDA settings, respectively.
- *DomainNet* is a large-scale domain adaptation dataset that consists of six domains divided into 345 categories. However, for fair comparison with previous studies, we selected *Real* (**R**), *Painting* (**P**), *Clipart* (**C**), and *Sketch* (**S**) as the four domains to evaluate via seven domain adaptation tasks with 126 classes, as follows: **R**→**C** (*Real* is selected as the source domain to adapt

TABLE 6. Classification accuracy (%) on the *Office-Home* dataset for the 3-shot setting using Alexnet, VGG16 and ResNet34 as the backbone network. The highest and second-highest accuracy are represented by red and blue in the table.

Net	Method	R→C	R→P	R→A	P→R	P→C	P→A	A→P	A→C	A→R	C→R	C→A	C→P	Mean
AlexNet	S+T	44.6	66.7	47.7	57.8	44.4	36.1	57.6	38.8	57.0	54.3	37.5	57.9	50.0
	DANN	47.2	66.7	46.6	58.1	44.4	36.1	57.2	39.8	56.6	54.3	38.6	57.9	50.3
	ENT	44.9	70.4	47.1	60.3	41.2	34.6	60.7	37.8	60.5	58.0	31.8	63.4	50.9
	MME	51.2	73.0	50.3	61.6	47.2	40.7	63.9	43.8	61.4	59.9	44.7	64.7	55.2
	PAC	58.9	72.4	47.5	61.9	53.2	39.6	63.8	49.9	60.0	54.5	36.3	64.8	55.2
	APE	51.9	74.6	51.2	61.6	47.9	42.1	65.5	44.5	60.9	57.1	44.3	64.8	55.6
	STar	51.9	74.1	54.0	65.1	48.5	43.0	66.0	45.8	63.0	61.5	46.1	65.3	57.0
	CDAC	54.9	75.8	51.8	64.3	51.3	43.6	65.1	47.5	63.1	63.0	44.9	65.6	56.8
	Ours	55.4	73.1	54.6	65.6	49.9	44.7	66.0	47.9	64.5	59.7	42.9	63.3	57.3
VGG16	S+T	49.6	78.6	63.6	72.7	47.2	55.9	69.4	47.5	73.4	69.7	56.2	70.4	62.9
	DANN	56.1	77.9	63.7	73.6	52.4	56.3	69.5	50.0	72.3	68.7	56.4	69.8	63.9
	ENT	48.3	81.6	65.5	76.6	46.8	56.9	73.0	44.8	75.3	72.9	59.1	77.0	64.8
	MME	56.9	82.9	65.7	76.7	53.6	59.2	75.7	54.9	75.3	72.9	61.1	76.3	67.6
	APE	56.0	81.0	65.2	73.7	51.4	59.3	75.0	54.4	73.7	71.4	61.7	75.1	66.5
	PAC	63.5	82.3	66.8	75.8	58.6	57.1	75.9	56.7	72.2	70.5	57.7	75.3	67.7
	ELP	57.1	83.2	67.0	76.3	53.9	59.3	75.9	55.1	76.3	73.3	61.9	76.1	68.0
	UODA	57.6	83.6	67.5	77.7	54.9	61.0	77.7	55.4	76.7	73.8	61.9	78.4	68.9
	Ours	62.8	82.5	69.8	77.6	59.5	62.3	77.0	58.3	77.5	74.4	62.0	75.9	70.0
ResNet34	S+T	55.7	80.8	67.8	73.1	53.8	63.5	73.1	54.0	74.2	68.3	57.6	72.3	66.2
	DANN	57.3	75.5	65.2	69.2	51.8	56.6	68.3	54.7	73.8	67.1	55.1	67.5	63.5
	ENT	62.6	85.7	70.2	79.9	60.5	63.9	79.5	61.3	79.1	76.4	64.7	79.1	71.9
	MME	64.6	85.5	71.3	80.1	64.6	65.5	79.0	63.6	79.7	76.6	67.2	79.3	73.1
	APE	66.4	86.2	73.4	82.0	65.2	66.1	81.1	63.9	80.2	76.8	66.6	79.9	74.0
	CDAC	67.8	85.6	72.2	81.9	67.0	67.5	80.3	65.9	80.6	80.2	67.4	81.4	74.2
	Ours	69.6	88.1	76.4	80.9	66.0	71.2	82.0	66.0	79.6	79.6	67.4	80.7	75.6

on the target domain *Clipart*), **R→P**, **P→C**, **C→S**, **S→P**, **R→S**, and **P→R**. For each task, we evaluated our method on the 1-, 3-, 5-, and 10-shot settings, where one, three, five, and ten are the labeled samples that were randomly selected from the target domain, respectively. Table 2 summarizes the dataset description.

B. EXPERIMENT SETTINGS

Similar to previous SSDA approaches [17], [18], [24], we used Alexnet, VGG16, and ResNet-34 as backbones for the shared feature extractor; they were pre-trained on the ImageNet dataset [39]. The two classifiers of $\mathcal{M}_S$ and $\mathcal{M}_T$ have the same architecture as [17]. We used a Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of $\eta_0 = 0.01$, weight decay of 0.0005, and momentum of 0.9. The strategy to update the learning rate η was conducted following [11]: $\eta = (\eta_0/((1+10p)^{0.75}))$, where $p \in [0, 1]$. The threshold values for selecting the pseudo labels of the inter- and intra-view models were set to 0.96, as detailed in Section IV. E. The batch size (b) was set to 128 in the experiments. We conducted all experiments on the widely used Pytorch [40] framework.

All results of the *Office-31*, *Office-Home*, and *DomainNet* datasets for the benchmark methods were collected from previous works [17], [22], [23] based on Alexnet, VGG16, and ResNet34. For the results on the *VisDa2017* dataset, we implemented the benchmark methods ourselves using the codes released by the authors.

For the experiments on the *Office-31* and *Office-Home* datasets, the model was trained for 10,000 training steps to collect the best accuracy for target validation. For the experiments on the *VisDA2017* and *DomainNet* datasets, we trained all models for 50,000 training steps to observe the best classification accuracy on the target domain.

TABLE 7. Classification accuracy (%) on the *Office-31* dataset for the 1-shot and 3-shot settings using Alexnet and VGG16 as the backbone networks. The highest and second-highest accuracies are illustrated in red and blue.

Net	Method	D→A		W→A		Mean	
		1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
AlexNet	S+T	50.0	62.4	50.4	61.2	50.2	61.8
	DANN	54.5	65.2	57.0	64.4	55.8	64.8
	ENT	50.0	66.2	50.7	64.0	50.4	65.1
	MME	55.8	67.8	57.2	67.3	56.5	67.6
	APE	-	69.0	-	67.6	-	68.3
	PAC	54.7	66.3	53.6	65.1	54.2	65.7
	Ours	59.3	69.1	56.7	69.0	58.0	69.1
VGG16	S+T	68.2	73.3	69.2	73.2	68.7	73.3
	DANN	70.4	74.6	69.3	75.4	69.9	75.0
	ENT	72.1	75.1	69.1	75.4	70.6	75.3
	MME	73.6	77.6	73.1	76.3	73.4	77.0
	PAC	72.4	75.6	70.2	76.0	71.3	75.8
	Ours	72.9	78.3	74.9	77.9	73.9	78.1

C. COMPARISON WITH STATE-OF-THE-ART APPROACHES

We compared the proposed method with previous SOTA SSDA approaches, including the following:

1) Minimax entropy (MME) [17].
2) Attract, perturb, and explore (APE) [18].
3) Mapping function (MAP-F) [19].

TABLE 8. Ablation study on *DomainNet* using ResNet34 for R→P, P→C, S→P with the 3-shot setting. The table reports the impact of each component of the proposed method on the classification performance in the target domain.

Baseline	DA	SL	CL	R→P		P→C		S→P		Mean	
				Inter-view	Intra-view	Inter-view	Intra-view	Inter-view	Intra-view	Inter-view	Intra-view
✓				68.8	60.4	67.0	60.2	65.5	55.9	67.1	58.8
✓	✓			70.2	61.9	69.3	60.7	67.3	56.3	68.9	59.6
✓	✓	✓		75.3	64.8	72.5	65.5	70.9	66.4	72.9	65.6
✓	✓		✓	78.3	78.3	80.2	80.2	76.7	76.7	78.4	78.4

Note: DA: Domain Alignment, SL: Self-learning, CL: Co-learning

TABLE 9. Experiments for all adaptation tasks on *DomainNet* using ResNet34 with the 3-shot setting.

Method	Task	R→C	R→P	P→C	C→S	S→P	R→S	P→R	Mean
BL+DA+SL	Inter-view	73.4	75.3	72.5	64.9	70.9	68.1	81.2	72.2
	Intra-view	65.7	64.8	65.5	57.7	66.4	59.2	75.1	64.8
	Ensemble	68.7	70.0	69.6	61.6	69.5	62.7	78.2	68.6
BL+DA+CL	Inter-view	79.7	78.3	80.2	72.9	76.7	73.0	85.0	78.0
	Intra-view	79.7	78.3	80.2	72.9	76.7	73.0	85.0	78.0
	Ensemble	79.8	78.4	80.3	73.0	76.7	74.4	85.1	78.2

4) Unifying the learning of opposite structures (UODA) [20].
5) Effective label propagation (ELP) [21].
6) Cross-domain adaptive clustering (CDAC) [22].
7) Pretraining and consistency (PAC) [23].
8) Select target (STar) [24].
9) Bidirectional adversarial training (BiAT) [25].

Additionally, we made a comparison to S+T [41], which is trained with the labeled source and target samples without using the unlabeled target samples. DANN [11] and ENT [42] are both methods widely used in UDA. DANN is the domain adversarial learning method, which employs a domain classifier for matching feature distributions of the source and target domains. ENT is trained on labeled data using the standard cross-entropy loss and unlabeled data using entropy minimization. We modified these methods to be suitable for the SSDA setting, as in [17].

D. ANALYSIS OF RESULTS

Results on the VisDa2017 dataset: Table 3 reports comparisons of the results of our method and SOTA SSDA approaches on the *VisDA2017* dataset in the 3-shot setting. We could see that the proposed method showed remarkable improvement in classification accuracy on the target domain for almost all tasks. Our method achieved the best classification accuracy in the target domain, 88.9%, and was 2.2% better than the current SOTA method MAP-F [19]. This means over 11.8% and 11.1% improvements in the average classification results compared to MME [17] and APE [18], respectively.

Results on the DomainNet dataset: The mean classification accuracy of our method achieved the best performance on the *DomainNet* dataset in both cases using Alexnet and ResNet34 as the backbone network. The details of the comparison results are listed in Table 4. Specifically, compared to the most popular SSDA method, MME [17], the average classification results on the target domain of our method were boosted by 9.2% and 8.5% of the 1-shot and 3-shot settings, respectively, when using Alexnet as the backbone network. Using ResNet34 as the backbone network, the proposed method achieved the best accuracy of the target domain in all tasks and surpassed the current best results obtained by CDAC [22] by 2.6% and 2.1% in the 1-shot and 3-shot settings, respectively.

We extended the experiments on the *DomainNet* dataset to evaluate the proposed method in the 5-shot and 10-shot settings. Compared to the existing SOTA SSDA counterparts, our method showed outstanding results in almost all adaptation scenarios.

Results on Office-Home and Office-31 datasets: The average classification results of our method showed the best performance for Alexnet, VGG16, and ResNet34 on the target domain under the 3-shot setting for the *Office-Home* dataset, as presented in Table 6. It performed 0.3%, 1.1%, and 1.4%, respectively, better than the current SOTA method. The classification accuracy on the target domain of various methods for the *Office-31* dataset is listed in Table 7; our approach reported remarkable classification results in the target domain when using Alexnet and VGG16 as backbone networks on both the 1-shot and 3-shot settings.

E. ABLATION STUDIES

1) THE CONTRIBUTIONS OF EACH COMPONENT OF THE PROPOSED METHOD

We investigated the contributions of each component of the proposed method on the classification performance on the target domain, including baseline (BL), domain alignment (DA), self-learning (SL), and co-learning (CL). We used the MME [17] as the baseline of models $\mathcal{M}_S$ and $\mathcal{M}_T$. The baseline results on the target data were computed by taking

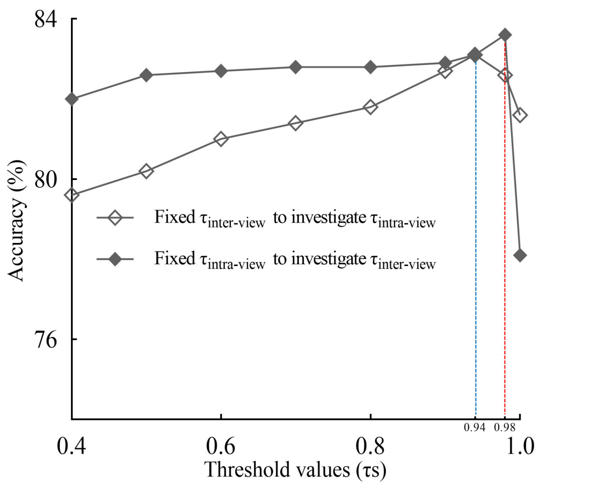


FIGURE 3. Sensitivity analysis of both inter-view threshold value (τ_{inter}) and intra-view threshold value (τ_{intra}) for the classification accuracy of the target domain.

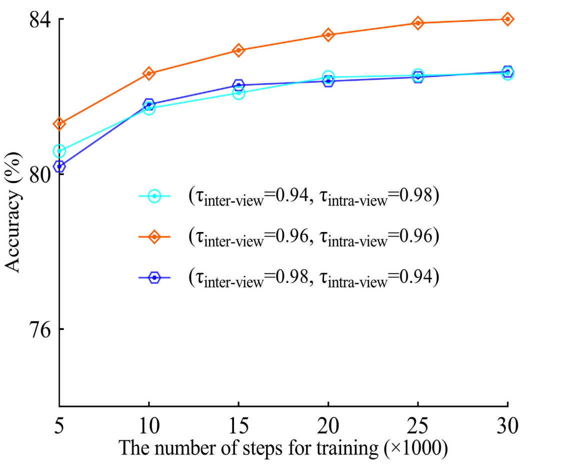


FIGURE 4. Investigation of an optimal threshold value pair (τ_{inter}, τ_{intra}).

an ensemble of softmax outputs from both models. We added a domain loss for DA to match the feature distributions of the source and target domains extracted by the inter-view model $\mathcal{M}_S$. For the SL process, each model exploited the target information by generating the pseudo labels on the unlabeled target samples to train itself. In the CL process, both the inter-view and intra-view models exchanged their knowledge by alternatively providing their pseudo labels, selecting from the highest confidence prediction to teach the other model. Consequently, the shortage in each view is alleviated, leading to improved classification performance in the target domain.

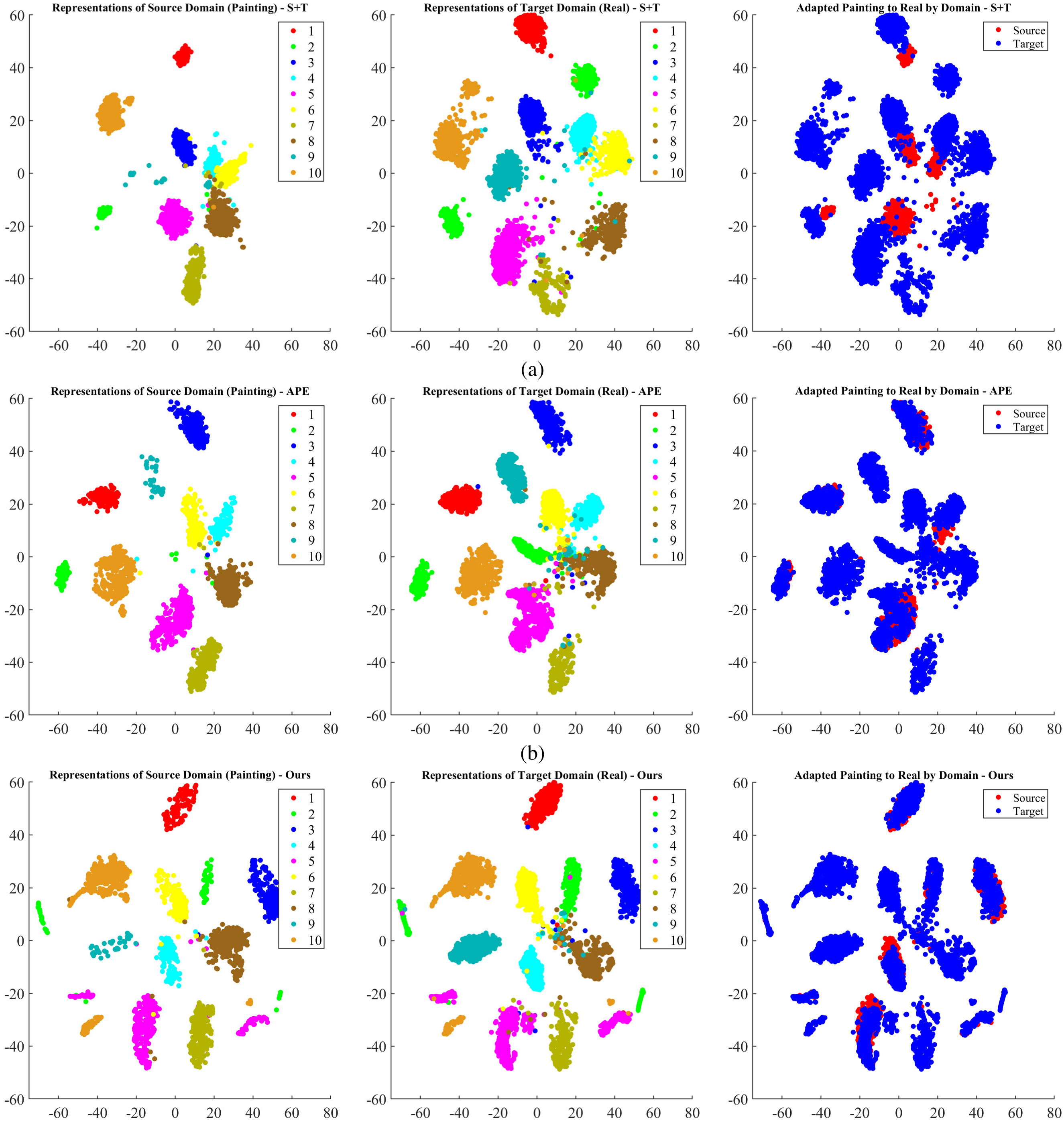
Table 8 recorded three adaptation tasks (**R**→**P**, **P**→**C**, **S**→**P**) on *DomainNet* with the 3-shot setting using ResNet34 as the backbone network. The average classification results on the target domain of the inter-view and the intra-view models were significantly different when only using BL, with 8.3% gapping. The average accuracy of the inference results on the target domain was improved when we added DA (BL+DA) to reduce the domain discrepancy between the source and target domains in the feature space, which is expressed in Section III. C. Because model $\mathcal{M}_S$ and model $\mathcal{M}_T$ share feature extractor E, in the classification accuracy on the target domain of the intra-view model, $\mathcal{M}_T$, also slightly increased. Compared to the baseline, in scenario BL+DA+SL, the average classification results on the target domain of both inter-and intra-view models surpassed 5.8% and 6.8%, respectively, when they were complementary to the target information from the unlabeled samples by using SL. However, as shown in Table 8, the prediction results on the target domain of the inter-view model over BL, BL+DA, and BL+DA+SL methods were significantly higher than the prediction results provided by the intra-view model. This is because the inter-view model was trained on the large amounts of labeled source samples, while the intra-view model was trained on the small amounts of labeled target samples. Therefore, the intra-view model was poorly generalized to the target domain compared to the inter-view model.

We could observe that in the case BL+DA+CL, the bias prediction of both models was removed when using CL, which encourages both inter-view and intra-view models to make similar predictions on an unlabeled target sample by mutually exchanging their knowledge. To demonstrate the efficiency of CL in the SSDA setting, we extended it to all adaptation tasks on the *DomainNet* dataset, and the results are listed in Table 9. Both models provided similar prediction accuracies in all tasks. The average accuracy of the ensemble result of the case BL+DA+CL was nearly 10.0% higher than that of the case BL+DA+SL. The intra-view model $\mathcal{M}_T$ inherited abundant ground-truth information from the inter-view model $\mathcal{M}_S$ via CL. Simultaneously, the inter-view model $\mathcal{M}_S$ supplemented the labeled target information during training. The experiments showed that the classification accuracy of the target domain could be significantly improved by using the proposed method for two reasons. First, the target information was extracted efficiently via multiple views. Second, CL successfully distilled useful class information of each model to transfer to the other model. Therefore, it could maximize the within-class correlation and simultaneously minimize the correlation between each class. We provide empirical evidence from feature visualization analysis in a later section.

2) SENSITIVITY OF THE PROPOSED METHOD WITH VARYING THRESHOLD VALUES

In this section, we explain how to select an appropriate threshold value. As mentioned in Section III. D, each model generated pseudo labels by choosing the highest confidence predictions on unlabeled target samples. Simultaneously, its weights were updated from the other's pseudo labels.

This is because the inter-view and intra-view models were trained on the differently labeled datasets. Therefore, the quality and quantity of their pseudo labels were quite different. To investigate the sensitivity of the proposed



(a)

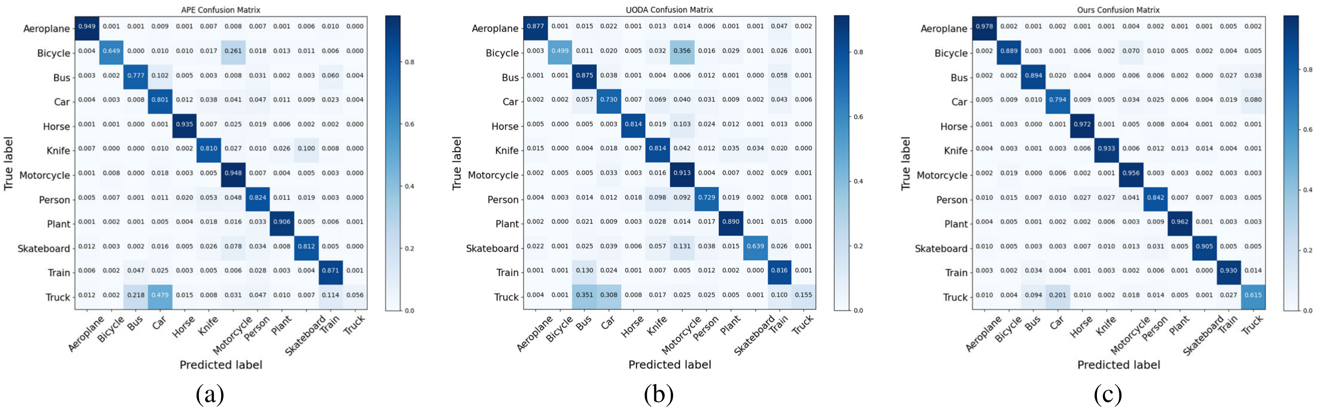
(b)

(c)

FIGURE 5. Visualization of the embedding space of different methods using t-SNE [44]. We show the representations of ten classes on the source and target domains of (a) S+T, (b) APE, and (c) our method for the P→R task considering the 3-shot setting on the *DomainNet* dataset. The left and middle columns of this figure visualize the source and target representations, respectively. The right column displays the adaptation efficiencies of the different methods.

method with pseudo labels generated from each model, we implemented the following: a model provided its pseudo labels with a fixed threshold value to observe the optimal threshold value of the rest model. We conducted the **P→R** task on *DomainNet* using ResNet34 as the backbone network under the 3-shot setting. We observed the inference results of the proposed method of all experiments during 30,000 training steps.

We could obtain a lot of pseudo labels by using a small threshold value. However, these pseudo labels contained noisy labels, leading to degradation in the classification performance on the target data. On the contrary, we could



(a) (b) (c)

FIGURE 6. Visualization of the confusion matrix of different models. These experiments were implemented on the *VisDa2017* dataset based on the ResNet-34 backbone network.

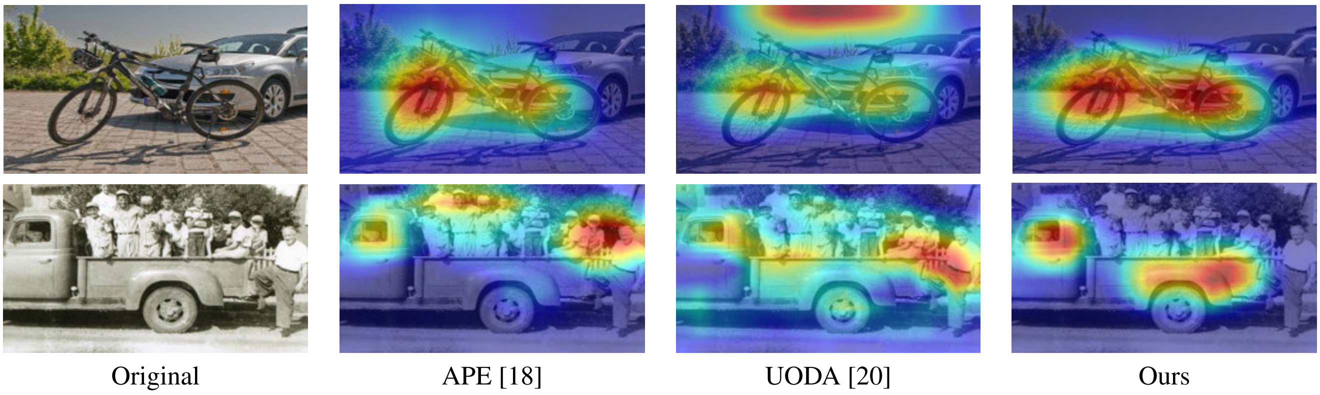


FIGURE 7. Grad-CAM results [45] of different models. We visualized the extracted features of the last convolutional layer in the ResNet34 backbone for attention map visualization analysis of the task *Synthetic→Real* on the *VisDA2017* dataset under the 3-shot setting.

get high-quality pseudo labels by setting a high threshold value. Nevertheless, in this way, the useful information of the target data could be discarded, which also led to a decrease in accuracy. Therefore, we should study the selection of a threshold value that can control the trade-off between the quantity and quality of pseudo labels. [43] suggested that the quality of pseudo labels should be considered more than the quantity to obtain better performance. Thus, we set $\tau_{inter} = 0.92$ and changed $\tau_{intra} = 0.4 \sim 1.0$ to evaluate the impact of the intra-view threshold value on the classification performance on the target domain. Similarly, we adjusted $\tau_{inter} = 0.4 \sim 1.0$ and fixed $\tau_{intra} = 0.92$ to investigate the sensitivity of the classification results of the target domain to the inter-view threshold value. Figure 3 shows the classification accuracies on the target domain of the proposed method, depending on the pair $(\tau_{inter}, \tau_{intra})$. In the case of fixed τ_{inter} and changed τ_{intra}, the classification accuracy on the target domain of the proposed method was over 83.0% when τ_{intra} was 0.94, which is indicated by a green dashed line. In the case of fixed τ_{intra} and changed τ_{inter}, the highest classification accuracy on the target domain of our method on the unlabeled target data could achieve nearly 84.0% with $\tau_{inter} = 0.98$, which is indicated by the red dashed line. As shown in this figure, we could see that the classification result on the target domain provided by the inter-view model was higher than that of the intra-view model with the same fixed threshold value, 0.92. That means the inter-view model trained on the large amounts of labeled source samples generated pseudo labels more accurately than the intra-view model trained on small amounts of labeled target samples when the threshold value changed from 0.4 to 0.9. These results are concordant with the classification accuracy of the target domain shown in Table 8.

As shown in Figure 3, we determined that the optimal threshold value was in an interval from 0.94 to 0.98. We observed the variation of the classification performance of the target domain corresponding to three pairs, $(\tau_{inter} = 0.94, \tau_{intra} = 0.98)$, $(\tau_{inter} = 0.96, \tau_{intra} = 0.96)$, and $(\tau_{inter} = 0.98, \tau_{intra} = 0.94)$, to decide the optimal threshold value for the proposed method. Figure 4 displays the test accuracies of these three tasks. As shown in this figure, in case $(\tau_{inter} = 0.96, \tau_{intra} = 0.96)$, the prediction results for the unlabeled target data still increased and achieved 84.0% after 30,000 training iterations. On the contrary, after 20,000 training iterations, in cases $(\tau_{inter} = 0.94, \tau_{intra} = 0.98)$ and $(\tau_{inter} = 0.98, \tau_{intra} = 0.94)$, the classification results of the target domain showed almost no change. Both models of the proposed framework obtained the best prediction results on the target domain with the same threshold value that was easy to understand. With CL, the quality of both inter-view and intra-view models was similar in terms of performance on the unlabeled target data, which is also demonstrated by the results listed in Table 9; the predictions of both models were the same for all tasks.

F. FEATURE VISUALIZATION ANALYSIS

1) VISUALIZATION ANALYSIS OF CLASSIFICATION FEATURES

Figure 5 shows the t-SNE visualization [44] of the source and target features categorized by the different methods for the **P→R** task on *DomainNet* under the 3-shot setting using the ResNet-34 backbone network. Different colors indicate different classes. The figures on the left side and in the middle columns show representations of the source and target domains, respectively. A different color denotes each class. The figure in the right-side column presents the effects of distribution matching between the source and target domains by showing the t-SNE visualization of domain representations in the shared feature space. The red color denotes the source domain, while the blue color indicates the target domain. We could see that the distribution matching of the S+T method was less efficient than APE and our method, as shown in the figure in the right-side column. The target features in each class were discriminated by our method more clearly compared to APE, as shown in the middle columns.

2) CONFUSION MATRIX VISUALIZATION

The confusion matrixes of APE, UODA, and our method are represented in Figures 6 (a)–(c), respectively. These experiments were implemented on the *VisDa2017* dataset based on the ResNet-34 backbone network. As shown in Figure 6, both the APE and UODA methods could not discriminate the representations among Bus, Car, Train, and Truck classes well because these classes share common features. On the contrary, our approach could alleviate this problem, as shown in Figure 6 (c).

3) ATTENTION MAP VISUALIZATION

Grad-CAM results [45] of APE, UODA, and our method are displayed in Figure 7, which shows the results extracted by the last convolutional layer in ResNet-34 using an input randomly selected from the *Bicycle* and *Truck* classes of the *VisDa2017* dataset. The results of both classes showed that the model of our method could be enforced to focus on the main object while other models were quite sensitive to the background or noise. As shown in Figure 7, the model of our method performed better than the UODA model on the *Bicycle* class and was more robust to noise than the APE and UODA models on the *Truck* class. These results are concordant with the confusion matrix visualization in Figure 6 and the classification accuracy of the target domain in Table 3.

V. CONCLUSION

In this paper, we successfully integrated the multiple views strategy and collaborative training into a framework for SSDA. Specifically, the multiple views strategy is responsible for extracting unlabeled target data from the different aspects of labeled target samples. Collaborative learning encourages the different models to exchange their knowledge to alleviate the shortage in each model. We conducted extensive experiments on four visual benchmark domain adaptation datasets. The experimental results have shown that MVCL is better than other state-of-the-art SSDA approaches. The success of our method indicates the importance of preserving the discriminative information of each class for learning domain-invariant representations in domain adaptation.

ACKNOWLEDGMENT

(Ba Hung Ngo and Ju Hyun Kim are co-first authors.)

REFERENCES

[1] S. Yang, K. Yu, F. Cao, H. Wang, and X. Wu, "Dual-representation-based autoencoder for domain adaptation," *IEEE Trans. Cybern.*, early access, Jan. 5, 2021, doi: 10.1109/TCYB.2020.3040763.

[2] S. Zang, Y. Cheng, X. Wang, Q. Yu, and G.-S. Xie, "Cross domain mean approximation for unsupervised domain adaptation," *IEEE Access*, vol. 8, pp. 139052–139069, 2020.

[3] W. Ren, S. Zhou, Z. Yang, Q. Shi, X. Sun, and L. Yang, "Metric information matrix for maximum mean discrepancy for domain adaptation," *IEEE Access*, vol. 9, pp. 148017–148023, 2021.

[4] Y. Wang, Z. Zhang, W. Hao, and C. Song, "Attention guided multiple source and target domain adaptation," *IEEE Trans. Image Process.*, vol. 30, pp. 892–906, 2021.

[5] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. CVPR*, Jun. 2019, pp. 6936–6945.

[6] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. CVPR*, Jun. 2019, pp. 2517–2526.

[7] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11008–11017.

[8] W. Li, M. Wang, H. Wang, and Y. Zhang, "Object detection based on semi-supervised domain adaptation for imbalanced domain resources," *Mach. Vis. Appl.*, vol. 31, no. 3, p. 18, Mar. 2020.

[9] S. Yang, L. Wu, A. Wiliem, and B. C. Lovell, "Unsupervised domain adaptive object detection using forward-backward cyclic adaptation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 1–17.

[10] D. Guan, J. Huang, A. Xiao, S. Lu, and Y. Cao, "Uncertainty-aware unsupervised domain adaptation in object detection," *IEEE Trans. Multimedia*, early access, May 25, 2021, doi: 10.1109/TMM.2021.3082687.

[11] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.

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

[13] M. Long, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.

[14] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 1, no. 2, pp. 2962–2971.

[15] J. Hoffman, E. Tzeng, T. Park, J. Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, J. Dy and A. Krause, Eds. Stockholm, Sweden: Stockholmsmässan, Jul. 2018, pp. 1989–1998.

[16] A. Chadha and Y. Andreopoulos, "Improved techniques for adversarial discriminative domain adaptation," *IEEE Trans. Image Process.*, vol. 29, pp. 2622–2637, 2020.

[17] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8050–8058.

[18] T. Kim and C. Kim, "Attract, perturb and explore: Learning a feature alignment network for semi-supervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 591–607.

[19] B. H. Ngo, J. H. Park, S. J. Park, and S. I. Cho, "Semi-supervised domain adaptation using explicit class-wise matching for domain-invariant and class-discriminative feature learning," *IEEE Access*, vol. 9, pp. 128467–128480, 2021.

[20] C. Qin, L. Wang, Q. Ma, Y. Yin, H. Wang, and Y. Fu, "Opposite structure learning for semi-supervised domain adaptation," in *Proc. SDM*, 2021, pp. 576–584.

[21] Z. Huang, K. Sheng, W. Dong, X. Mei, C. Ma, F. Huang, D. Zhou, and C. Xu, "Effective label propagation for discriminative semi-supervised domain adaptation," 2020, *arXiv:2012.02621*.

[22] J. Li, G. Li, Y. Shi, and Y. Yu, "Cross-domain adaptive clustering for semi-supervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2505–2514.

[23] S. Mishra, K. Saenko, and V. Saligrama, "Surprisingly simple semi-supervised domain adaptation with pretraining and consistency," 2021, *arXiv:2101.12727*.

[24] A. Singh, N. Doraiswamy, S. Takamuku, M. Bhalerao, T. Dutta, S. Biswas, A. Chepuri, B. Vengatesan, and N. Natori, "Improving semi-supervised domain adaptation using effective target selection and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2709–2718.

[25] P. Jiang, A. Wu, Y. Han, Y. Shao, M. Qi, and B. Li, "Bidirectional adversarial training for semi-supervised domain adaptation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 934–940.

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

[27] P. Yang and W. Gao, "Multi-view discriminant transfer learning," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1848–1854.

[28] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[29] Z. Ding, M. Shao, and Y. Fu, "Robust multi-view representation: A unified perspective from multi-view learning to domain adaption," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 5434–5440.

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

[31] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3008–3017.

[32] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)* 2012, pp. 1097–1105.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

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

[35] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. ECCV*, vol. 6314, 2010, pp. 213–226.

[36] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5018–5027.

[37] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "VisDA: The visual domain adaptation challenge," 2017, *arXiv:1710.06924*.

[38] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2019, pp. 1406–1415.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Workshop*, 2017, pp. 1–4.

[41] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)* 2019, pp. 1–24.

[42] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 529–536.

[43] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semisupervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1–21.

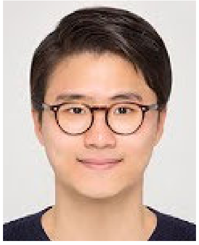
[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[45] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

BA HUNG NGO (Graduate Student Member, IEEE) received the B.S. degree in control engineering and automation from the Hanoi University of Mining and Geology, Hanoi, Vietnam, in 2014, and the M.S. degree in control engineering and automation from the Hanoi University of Science and Technology, in 2016. He is currently pursuing the Ph.D. degree with Dongguk University, Seoul, Republic of Korea. His current research interests include computer vision and deep learning, especially deep transfer learning, domain adaptation, and deep learning in medical imaging.

JU HYUN KIM received the B.S. degree in electronic and electrical engineering from Dongguk University, Seoul, Republic of Korea, in 2019, where he is currently pursuing the M.S. degree in multimedia engineering. From 2019 to 2021, he worked as an Engineer with LG Display. His current research interests include object detection and domain adaptation.

YEON JEONG CHAE received the B.S. degree in multimedia engineering from Dongguk University, Seoul, Republic of Korea, in 2021, where she is currently pursuing the M.S. degree. Her current research interests include object detection, semantic segmentation, and domain adaptation.



SUNG IN CHO (Member, IEEE) received the B.S. degree in electronic engineering from Sogang University, Seoul, Republic of Korea, in 2010, and the Ph.D. degree in electrical and computer engineering from the Pohang University of Science and Technology, Pohang, Republic of Korea, in 2015. From 2015 to 2017, he was a Senior Researcher with LG Display. From 2017 to 2019, he was an Assistant Professor of electronic engineering at Daegu University, Gyeongsan, Republic of Korea. He is currently an Assistant Professor of Multimedia Engineering with Dongguk University, Seoul. His current research interests include image analysis and enhancement, video processing, multimedia signal processing, and circuit design for LCD and OLED systems.

• • •