PRIVACY PROTECTED MULTI-DOMAIN COLLABORATIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised domain adaptation (UDA) aims to transfer knowledge from one or more well-labeled source domains to improve model performance on the different-yet-related target domain without any annotations. However, existing UDA algorithms fail to bring any benefits to source domains and neglect privacy protection during data sharing. With these considerations, we define Privacy Protected Multi-Domain Collaborative Learning (**P^{2}MDCL**) and propose a novel Mask-Driven Federated Network (**MDFNet**) to reach a "win-win" deal for multiple domains with data protected. First, each domain is armed with individual local model via a mask disentangled mechanism to learn domain-invariant semantics. Second, the centralized server refines the global invariant model by integrating and exchanging local knowledge across all domains. Moreover, adaptive self-supervised optimization is deployed to learn discriminative features for unlabeled domains. Finally, theoretical studies and experimental results illustrate rationality and effectiveness of our method on solving P^{2}MDCL.

1 INTRODUCTION

Unsupervised domain adaptation (UDA) (Tang et al., 2020; Jiang et al., 2020; Zhang et al., 2020) attempts to transfer knowledge from well-labeled source domains to annotate unlabeled target samples, which have significant domain discrepancy with source domains due to the various data collection manners and devices. Recent explorations (Na et al., 2021; Dong et al., 2020) suppose the model to be trained has access to both source and target data during the training stage. With such basic assumption, it becomes possible to measure the domain discrepancy and adopt metric-based solutions (Kang et al., 2020) or domain confusion (Cui et al., 2020; Tang & Jia, 2020) to generate domain-invariant features. However, the hypothesis violates the concerns of practical application on privacy protection, and fails to be deployed to small devices with limited storage.

This requirement motivates source-free domain adaptation (SFDA), where the source-supervised model is available to assist the target domain without any source data (Liang et al., 2020; Li et al., 2020; Kundu et al., 2020). Generally, SFDA either adapts target samples to source-like ones (Liang et al., 2020) or generates fake source samples from source-model by subsequently taking UDA strategies (Kurmi et al., 2021). To improve the training efficiency, FADA (Peng et al., 2020) employs a federated learning paradigm (Karimireddy et al., 2020; Chen et al., 2020) by allocating the target domain on a centralized server while keeping multiple source ones as clients. However, this approach is vulnerable to attacks as the source features transition to target domain. Further, these domain adaptation works ignore the improvement of model generalization on source domain, which is inconsistent with requirement of reality. *For example*, the long-standing hospitals already have well-annotated patients' data, while other newly-built hospitals just collected data without annotation, which need help from long-standing hospitals with well-annotated data due to the huge labeling cost. Besides, with geographical restriction, different hospitals only record their local patients' data resulting in various population statistics, causing model bias for long-standing hospitals.

Inspired by the above observation, we introduce a more practical scenario called *Privacy Protected Multi-Domain Collaborative Learning* (**P^{2}MDCL**) (shown in Figure 1). Specifically, P^{2}MDCL assumes that the well-annotated source and unlabeled target domains are distributed across different clients and there exists a global server merely communicating with each client and integrating the received model parameters from clients. Finally, the server broadcasts the consensus model to all

clients for their use to reach the **win-win** deal. The key challenge for P^{2}MDCL is to learn a more generic model by solving two core issues: 1) how to achieve domain alignment during iterative communication; and 2) how to enhance discriminative feature learning.

In this paper, we propose a novel Mask-Driven Federated Network (MDFNet) to address **P^{2}MDCL**. First, our MDFNet introduces two orthogonal masks following high-level features in each client to activate domain-invariant and domain-specific semantics respectively. In practice, we minimize the confusion of these two masks to achieve high-quality feature separation and semantic complementary. Second, the unlabeled target client adopts adaptive self-supervised optimization to learn more discriminative representations via pseudo labels generation. Finally, MDFNet adopts a progressive weighting scheme to balance the effect of each client in model integration on the server, which discoveries more knowledge of the labeled client to adjust the model of unlabeled client during the initial communication rounds, then the mature unlabeled client model also yields positive effect on the feature learning of labeled client. The main contributions of our work are summarized as:

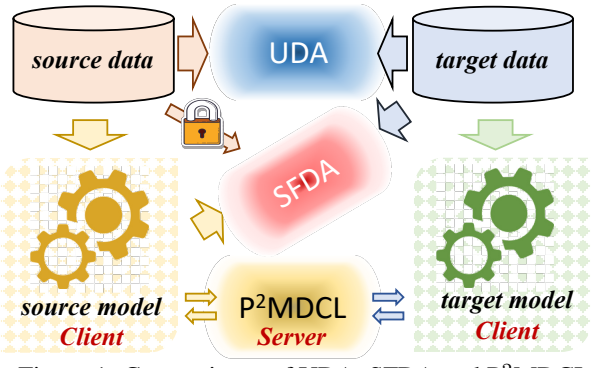


Figure 1: Comparisons of UDA, SFDA and P^{2}MDCL.

- First, we are the pioneers to take into account the "win-win" and privacy requirements under unsupervised domain adaptation scenarios by introducing Privacy Protected Multi-Domain Collaborative Learning (P^{2}MDCL).
- Second, we propose an effective algorithm MDFNet to fight off the domain shift in a federated training mechanism, which reaches the win-win deal for all involved domains.
- Finally, we derive the generalized error bound for our method, which theoretically testifies the rationality of MDFNet. Moreover, extensive experimental results and analysis empirically illustrate the effectiveness of our method on solving P^{2}MDCL.

2 Related Work

Domain Adaptation. Unsupervised domain adaptation (Cui et al., 2020) attempts to build a model with well-labeled source and unlabeled target data at hand, by mitigating the domain mismatch. Along this line, the recent explorations mainly adopt discrepancy metric-based method (Yan et al., 2017; Tzeng et al., 2014) and adversarial training scheme (Zhang et al., 2019; Tzeng et al., 2017) to learn domain-invariant features. Although these solutions effectively reduce the influence of domain discrepancy, the practical applications difficultly permit the co-existence of source and target data due to the limited storage of small device and data privacy. The demand stimulates the development of source-free domain adaptation (Liang et al., 2020; Kurmi et al., 2021), which merely provides the well-trained source model for knowledge adaption on target domain. In addition, Peng et al. (2020) respectively considers target domain and multiple source domains as the centralized server and clients and adopts federated learning fashion to achieve domain adaptation with multiple discriminators, which is vulnerable to the attack due to the source and target features transmission into the discriminators in the centralized target domain. Even though these strategies actually achieve the comparable transferring ability with the UDA solutions, empirical studies illustrate the current domain adaptation techniques fail to learn a generalized model for source and target domains. Alternatively, they only focus on the improvement of target performance, yet neglecting any benefit to source domain. To this end, this paper posts a novel and practical scenario privacy protected multi-domain collaborative learning (P^{2}MDCL), where source and target domains are both regarded as clients independently communicating with the server which produces and broadcasts the consensus model to clients for their use.

Federated Learning (FL). FL allows multi-clients collaboratively to complete the same task without data currency across clients (Yang et al., 2019). Along with this concept, recent works mainly focus on semi-supervised scenario (FSSL) where FedMatch (Jeong et al., 2021) allocates unlabeled

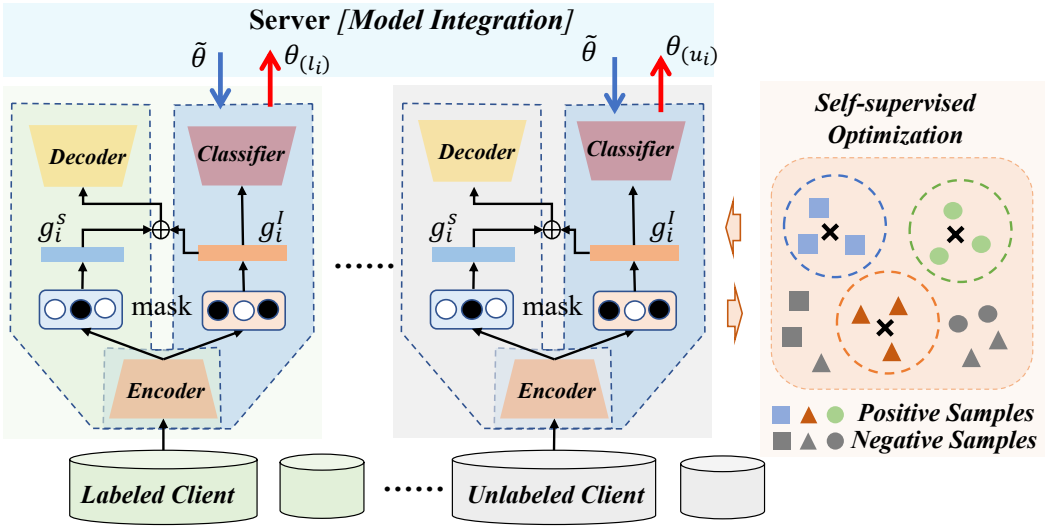


Figure 2: Framework of our proposed MDFNet including multiple local clients and one global server. For each client, there are three components (encoder, decoder, classifier). The encoder extracts high-level features to achieve domain-specific/invariant feature separation with two orthogonal masks, while the decoder takes the combination of separated features to reconstruct the original data. The server adopts progressive weight to execute the model integration.

data on client side and labeled data in the server while FedIRM (Liu et al., 2021) only deploys them on various clients. But they both assume the instances across all client are sampled from the identical distribution. Moreover, Smith et al. (2017); Liu et al. (2020) explore FSSL with non-i.i.d by supposing each client contains several well-annotated instances for training. Differently, our considered P^{2}MDCL closely approximates the reality, which involves several clients without any annotations and exists significant domain discrepancy across all clients.

3 The Proposed Method

3.1 Problem Definition and Motivation

The P^{2}MDCL scenario assumes there are L well-annotated source clients $\mathcal{D}_{l_i} = \{(x^l_{(i)j}, y^l_{(i)j})\}_{j=1}^{n_{l_i}}$ $(i \in \{1, \cdots, L\})$ and U unlabeled target clients $\mathcal{D}_{u_k} = \{x^u_{(k)j}\}_{j=1}^{n_{u_k}}$ $(k \in \{L+1, \cdots, L+U\})$, where x and y denote an input sample and its ground-truth label, respectively. The instances of these clients come from different distributions but share the identical category space and clients are not allowed to exchange private data with each other. Akin to federated learning, the additional global server in P^{2}MDCL collects and assembles all clients' network parameters to form the consensus model. The main motivation of P^{2}MDCL is addressing the negative effect of insufficient training samples in $\mathcal{D}_{l_i}$ and label shortage in $\mathcal{D}_{u_k}$ to reach a "win-win" deal across all clients. We face two challenges to solve P^{2}MDCL: 1) how to reduce the significant distribution discrepancy while protecting data privacy and 2) how to learn more generic and discriminative representations from unlabeled target clients. To this end, this work proposes an effective Mask-Driven Federated Network (MDFNet), which deploys mask-driven disentanglement to locally seek domain-specific/invariant features, and explores the adaptive self-supervised optimization to promote the discriminative ability of unlabeled target clients.

3.2 Mask-Driven Disentanglement

Feature separation is a commonly-used strategy in domain adaptation to disentangle latent representation into domain-specific features and domain-invariant ones (Bousmalis et al., 2016; Peng et al., 2019). However, they typically develop two separated networks to extract the corresponding features, which increase storage burden for each local device with insufficient computational resources. Peng et al. (2019) points out the high-level neurons from feature extractor actually involve domain-

specific and invariant knowledge. Inspired by (Chattopadhyay et al., 2020), we explore the binary mask to achieve feature disentanglement by activating the interested neurons.

For the brevity, we omit the symbols l/u and (k) in the following illustration. As Figure 2 shows, each client of our MDFNet contains a basic feature encoder parameterized θ_e mapping the raw input into the hidden space via $g_i = \theta_e(x_i) \in \mathbb{R}^d$. Subsequently, two additional parameters $\hat{\mathbf{m}}^s, \hat{\mathbf{m}}^I \in \mathbb{R}^d$ are introduced into the local network and activated to form the mask probabilities by using the sigmoid function $\sigma(\cdot)$ to get $\mathbf{m}^s = \sigma(\hat{\mathbf{m}}^s)$ and $\mathbf{m}^I = \sigma(\hat{\mathbf{m}}^I)$. For each feature g_i, based on the mask probabilities, we sample the binary domain-specific and invariant masks ($m_i^s, m_i^I \in \{0,1\}^d$) from the Bernoulli distributions. To this end, we obtain the domain specific and invariant features with the element-wise multiplication $\otimes$ over binary masks and features, i.e., $g_i^s = m_i^s \otimes g_i$ and $g_i^I = m_i^I \otimes g_i$. Moreover, we adopt three strategies to achieve high-quality feature separation. Concretely, each client firstly minimizes the semantic overlap between g_i^I and g_i^s to store complementary information in them. Motivated by Rahman & Wang (2016), we design the following soft-interactive loss as:

$$\mathcal{L}_s = \sum\nolimits_i \frac{\langle g_i^s, g_i^I \rangle}{\mathbf{sum}(g_i^s + g_i^I - g_i^s \otimes g_i^I)}, \tag{1}$$

where $\langle , \rangle$ means the inner product of two feature vectors, and $\mathbf{sum}(\cdot)$ represents the sum of all elements for a vector. This approximately reflects the information overlap of two mask distributions. The minimization of soft-interactive loss gradually increases the difference between m_i^s and m_i^I which activate different neurons. Similar to DSN (Bousmalis et al. (2016)), each client also develops the individual classifier $\theta_c(\cdot)$ taking domain-invariant features as input to the category probability distribution $\theta_c(g_j^I)$. The cross-entropy loss between ground-truth and prediction intensifies the discriminnative ability of domain-invariant features. On the other hand, we also feed the combination of g_i^s and g_i^I into decoder $\theta_d(\cdot)$ to reconstruct the original input with $\mathcal{L}_r = \sum_i \|\theta_d(g_i^s, g_i^I) - x_i\|_2^2$. Thus, the overall loss function of mask-driven disentanglement for labeled clients is formulated as:

$$\min_{\theta_c, \theta_e, \theta_d, \hat{\mathbf{m}}^s, \hat{\mathbf{m}}^I} \mathcal{L}_o^l = \sum\nolimits_i -y_i \log\big(\theta_c(g_i^I)\big) + \mathcal{L}_r + \mathcal{L}_s, \tag{2}$$

where we actually adopt straight-through estimator (Bengio et al., 2013) to progressively optimize $\hat{\mathbf{m}}^s$ and $\hat{\mathbf{m}}^I$ instead of the discrete binary masks leading to the invalid back-propagation.

3.3 Adaptive Self-Supervised Optimization

Due to the availability of annotation in the well-labeled clients, we can easily calibrate the predicted category distribution to generate discriminative features by using the supervision of ground-truth. However, we cannot directly adopt the supervised learning manner to optimize the model in unlabeled clients with the absence of annotation. Inspired by the successful application of pseudo-label on solving UDA issue (Xie et al., 2018; Gu et al., 2020; Liang et al., 2019; Morerio et al., 2020), we thus propose the adaptive clustering optimization module to gradually produce the pseudo-label as "ground-truth" supervision.

Specifically, after each round of communication, the unlabeled client first receives the model broadcast from the server and uses it to initialize the parameters of $\theta_e, \theta_d, \theta_c, \hat{\mathbf{m}}^s, \hat{\mathbf{m}}^I$. Before further optimizing, the client annotates its local data with the received global model, i.e., $\hat{y}_j = \arg\max_k \theta_c(g_j^I)_k$. With the predictions, the initial centroid of each category is computed as $\mathcal{O}_k = \frac{\sum_j \mathbf{1}(\hat{y}_j = k) g_j^I}{\sum_j 1(\hat{y}_j = k)}$, where $\mathbf{1}(\cdot)$ is the indicator function. Since the server model integrates knowledge from multiple clients, the domain shift negatively affects the accuracy of inference $\hat{y}_j$. To decrease the influence, we adopt an iterative approach to further update the class centers and pseudo-label with the local data points. The proposed adaptive clustering optimization mainly includes two operations. The first step is to reassign the label for each instance with the spherical K-means (Buchta et al., 2012):

$$\hat{y}_j = \arg\min_k \tilde{\mathbf{d}}\big(g_j^I, \mathcal{O}_k\big) = \frac{1}{2}\big(1 - \frac{\langle g_j^I, \mathcal{O}_k \rangle}{\|g_j^I\| \cdot \|\mathcal{O}_k\|}\big). \tag{3}$$

With the reattached annotations, the second step is to update the class prototype with $\mathcal{O}_k = \sum_j \frac{\mathbf{1}(\hat{y}_j = k) g_j^I}{\|g_j^I\|}$. The above two steps are repeated till convergence.

After the adaptive self-supervised optimization, we attain the final pseudo-label for each sample and use them as supervision to optimize the local models. However, not all samples of unlabeled clients would contribute to the parameter sharing due to the domain mismatch, which in turn causes the low reliability and uncertainty for some samples. In this way, we consider samples to be positive and negative samples, identifying their potential benefit for the labeled clients. Therefore, it is crucial to distinguish positive and negative samples by identifying their potential benefit to the labeled clients. To this end, we add additional entropy-minimization (EM) to further improve the certainty of category prediction, reformulate the Eq. (2) for the unlabeled clients as:

$$\min_{\theta_c,\theta_e,\theta_d,\hat{\mathbf{m}}^s,\hat{\mathbf{m}}^I} \mathcal{L}_o^u = \sum\nolimits_j \Big(-\mathbb{I}(\max(\theta_c(g_j^I)) \geq \sigma) y_j \log(\theta_c(g_j^I)) - \theta_c(g_j^I) \log(\theta_c(g_j^I)) \Big) + \mathcal{L}_r + \mathcal{L}_s, \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator to filter out samples with $\theta_c(g_j^I)$ less than a threshold σ, which is set as 0.1 by default throughout our experiments.

3.3.1 Federated Training

The overall training of our MDFNet involves two important procedures: a) local clients training, and b) global sever model integration. The clients and server collaboratively update these steps per communication round and repeat the process until model convergence or reaching the maximum communication rounds.

Independent Client Training. In each round, the server broadcasts the consensus model integrated from the last round to all available clients for the initialization of local models. The well-annotated clients then employ their local data to optimize all the modules for one epoch via Eq. (2), while the clients without labels rely on the adaptive clustering optimization to generate pseudo-labels for their samples and update their models with Eq. (4). The clients will locally store the parameters of domain-specific mask and decoder and use them to initialize the network in the next round.

Model Integration. After the local training, the clients will send their local models (excluding the parameters of domain-specific mask and decoder) to the server, where the models are integrated to achieve consensus. However, adopting pseudo-labels as supervision significantly reduce the reliability of the models, especially in the initial training stages. To avoid the negative effect of pseudo-labels, the server considers providing different weights to labeled and unlabeled clients as: $\tilde{\theta} = \frac{1-\eta_r}{L} \sum_{i=1}^{L} \theta_{(l_i)} + \frac{\eta_r}{U} \sum_{i=L+1}^{L+U} \theta_{(u_i)}$, where $\theta \in \{\theta_e, \theta_d, \theta_c, \hat{\mathbf{m}}^I, \hat{\mathbf{m}}^s\}$ and $\eta_r = \frac{1-\exp(-\rho r)}{2(1+\exp(-\rho r))}$, r is the round of communication and each round means one epoch, and ρ is set as 10 in experiments.

3.4 Generalized Error Bound Analysis

We firstly define the basic notation and employ them to derive the generalization error bound for P^{2}MDCL from the high-level interpretation and the specific proofs are shown in the supplementary.

Notation. Given the distributions of labeled and unlabeled clients $\mathcal{D}_{l_i}$ and $\mathcal{D}_{u_j}$ on the input space $\mathcal{X}$, we have access to the ground-truth labeling function $f_{l_i} : \mathcal{X} \to \{0, 1\}$ for the clients with the annotation, while also have the pseudo labeling function $f_{u_j} : \mathcal{X} \to \{0, 1\}$ available for the unlabeled clients. A hypothesis is corresponding to a function $h : \mathcal{X} \to \{0, 1\}$ with the error, i.e., $\epsilon_{l_i}(h, f_{l_i}) := \mathrm{E}_{\mathbf{x}\sim D_{l_i}}[|h(\mathbf{x} - f_{l_i}(\mathbf{x})|]$ and $\epsilon_{u_j}(h, f_{u_j}) := \mathrm{E}_{\mathbf{x}\sim D_{u_j}}[|h(\mathbf{x} - f_{u_j}(\mathbf{x})|]$. Thus, the risk and the empirical risk of hypothesis h on $\mathcal{D}_{l_i}$ and $\mathcal{D}_{u_j}$ are respectively represented as $\epsilon_{l_i}(h)$, $\hat{\epsilon}_{l_i}(h)$, and $\epsilon_{u_j}(h)$, $\hat{\epsilon}_{u_j}(h)$. Moreover, we define the $\mathcal{H}$-divergence between two arbitrary distributions $\mathcal{D}$ and $\mathcal{D}^{'}$ as $d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}^{'}) = 2 \sup_{A\in\mathcal{A}_{\mathcal{H}}} |Pr_{\mathcal{D}}(A) - Pr_{\mathcal{D}'}(A)|$, where $\mathcal{H}$ means the hypothesis class for input space $\mathcal{X}$ and $\mathcal{A}_{\mathcal{H}}$ is the collection of subsets of $\mathcal{X}$ that are the support of some hypothesis in $\mathcal{H}$. The symmetric difference space with the hypothesis class is formulated as $\mathcal{H}\Delta\mathcal{H} = \{h(\mathbf{x}) * h^{'}(\mathbf{x}) | h, h^{'} \in \mathcal{H}\}$, where $*$ denotes the XOR operation.

Our model aims to learn a consensus model through the communication between the server and all available clients. Such learning strategy actually attempts to minimize a convex combination of empirical risks over all clients with parameters α_i ($\sum_{i=1}^{L+U} \alpha_i = 1$) as $\hat{\epsilon}_\alpha = \sum_{i=1}^{L} \alpha_i \hat{\epsilon}_{l_i}(h) + \sum_{i=L+1}^{L+U} \alpha_i \epsilon_{u_i}(h)$. Similarly, we obtain the weighted combination of the risks over all clients as $\epsilon_\alpha(h)$. In addition, since each client independently trains the model with its specific data, we denote the optimal hypothesises achieving the minimum risk on the labeled and unlabeled clients as $h_{l_i}^* :=$

$\arg\min_{h\in\mathcal{H}} \epsilon_{l_i}(h)$ and $h^*_{u_j} := \arg\min_{h\in\mathcal{H}} \epsilon_{u_j}(h)$. With these definitions, it still is intractable to directly deduce the generalized error bound under this scenario. Therefore, we alternatively divide the entire problem into multiple sub-problems and solve them in each client.

Concretely, we first explore the relationship between ϵ_α and ϵ_{l_i} or ϵ_{u_j} with **Lemma 1**. Second, we drive the upper bound of the difference between ϵ_α and $\hat{\epsilon}_\alpha$ via **Lemma 2**. Thus, we easily deduce the generalized error bound of a hypothesis per client in **Theorem 1**.

Lemma 1. Suppose the h is a hypothesis of class $\mathcal{H}$, for each unlabeled client, we then achieve:

$$|\epsilon_\alpha(h) - \epsilon_{u_j}(h)| \leq \sum_{i=1}^{L} \alpha_i (\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{l_i}, \mathcal{D}_{u_j}) + \lambda_{l_i}) + \sum_{i=L+1, i\neq j}^{L+U} \alpha_i (\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{u_i}, \mathcal{D}_{u_j}) + \lambda_{u_i}),$$

where $\lambda_{l_i} := \epsilon_{l_i}(h^*) + \epsilon_{u_j}(h^*)$ and h^* is the hypothesis which achieves the minimum risk on $\mathcal{D}_{l_i}$ and $\mathcal{D}_{u_j}$, and λ_{u_i} similarly means the risk of optimal hypothesis on the mixture of $\mathcal{D}_{u_i}$ and $\mathcal{D}_{u_j}$. Akin to unlabeled clients, we also derive the analogous inequality in clients with ground-truth as:

$$|\epsilon_\alpha(h) - \epsilon_{l_j}(h)| \leq \sum_{i=1, i\neq j}^{L} \alpha_i (\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{l_i}, \mathcal{D}_{l_j}) + \lambda_{l_i}) + \sum_{i=L+1}^{L+U} \alpha_i (\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{u_i}, \mathcal{D}_{l_j}) + \lambda_{u_i}),$$

where λ_{l_i} is the risk of optimal hypothesis of $\mathcal{D}_{l_i}$ and $\mathcal{D}_{l_j}$, and $\lambda_{u_i} := \epsilon_{u_i}(h^*) + \epsilon_{l_j}(h^*)$.

Lemma 2. Given a hypothesis space $\mathcal{H}$ of VC-dimension d, if a random sample of size n is generated by selecting $n\beta_j$ data points from $\mathcal{D}_{l_j}$ or $\mathcal{D}_{u_j}$, and annotating them through f_{l_j} and f_{u_j}, then with probability at least $1-\delta$, $\forall h \in \mathcal{H}$, we have:

$$|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)| \leq \sqrt{\sum_{j=1}^{L+U} \frac{\alpha_j^2}{\beta_j}} \sqrt{\frac{d\log(2n) - \log\delta}{2n}}.$$

Theorem 1. Suppose given $n\beta_i$ labeled instances from client $\mathcal{D}_{l_i}$ for $i = 1, \cdots, L$, and $n\beta_j$ unlabeled instances from client $\mathcal{D}_{u_j}$ in a federated learning system, we define $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\epsilon}_\alpha(h)$, $h^*_{l_i} := \arg\min_{h\in\mathcal{H}} \epsilon_{l_i}(h)$ and $h^*_{u_j} := \arg\min_{h\in\mathcal{H}} \epsilon_{u_j}(h)$. Then, $\forall \alpha_i \in \mathbb{R}^+, \sum_{i=1}^{L+U} \alpha_i = 1$, with probability at least $1-\delta$ over the choice of samples from each client:

$$\begin{aligned}\epsilon_{u_j}(\hat{h}) \leq & \epsilon_{u_j}(h^*_{u_j}) + 2\sqrt{\sum_{j=1}^{L+U} \frac{\alpha_j^2}{\beta_j}} \sqrt{\frac{d\log(2n) - \log\delta}{2n}} \\ & + 2\Big(\sum_{i=1}^{L} \alpha_i (\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{l_i}, \mathcal{D}_{u_j}) + \lambda_{l_i}) + \sum_{i=L+1, i\neq j}^{L+U} \alpha_i (\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{u_i}, \mathcal{D}_{u_j}) + \lambda_{u_i})\Big).\end{aligned}$$

For the annotated client, we can achieve the similar inequality. From **Theorem 1**, we explicitly observe the risk of a hypothesis with the federated training manner on the client is determined by three components: the error of the optimal hypothesis $h^*_{u_j}$ on its own samples, the VC-dimension constraint and the distribution discrepancy across various clients. To effectively reduce the risk of a hypothesis on all clients, we should not only learn the discriminative features by the independent client training, but also attempt to solve the domain shift with the constraint of data privacy.

4 Experiments

4.1 Experimental Settings

Datasets. Image-CLEF collects visual signals from three domains: Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**) and Pascal VOC 2012 (**P**) with the same number of samples. Concretely, arbitrary subset includes 600 images evenly distributed in 12 categories. **Office-Home** (Venkateswara et al., 2017) consists of four domains: Artistic images (**Ar**, 2,183), Clip Art (**Cl**, 4,365), Product images (**Pr**, 4,439) and Real-World images (**Rw**, 4,357), which share the identical 65 object categories. To verify the "win-win" deal, we randomly split the original data of labeled client into the training and test sets evenly, and repeat this operation for ten times[1].

[1]We further report the comparison under the original protocol of SHOT in supplemental materials, where all source samples are used for training and the evaluation is only on target domain.

Baselines. To the best of our knowledge, we are the pioneers to consider P^{2}MDCL scenario, and we aim to assess if our algorithm can learn a server model with higher generalization ability to enhance the performance across all clients via federated training. To explicitly testify the generalization of model on each client, this section focuses on P^{2}MDCL with a labeled client and an unlabeled one. Noted that we report the P^{2}MDCL with more clients in supplementary material. Since this scenario is similar with UDA and source-free DA, we not only select CDAN (Long et al., 2017) and SRDC (Tang et al., 2020) achieving the state-of-the-art results on UDA problem as the benchmarks and also regard the SHOT (Liang et al., 2020) as one important competitor. In addition, we also consider the source-only method merely training the model on the labeled client. For the mentioned baselines, we use their released code with the suggested parameters to carry out each task for ten times

Evaluation Metric. In terms of the data organization, the labeled client includes the training and testing sets without any overlap, while all samples of the unlabeled client participant the model training and evaluation. For UDA and source-free solutions, the training set of labeled client is considered as the source domain and the unlabeled client servers as the target domain. During the test stage, the final model learned by each method is not only evaluated on the test set of labeled client with the corresponding source accuracy $\mathbf{ACC}_s$ but also tested on the unlabeled client with the target accuracy $\mathbf{ACC}_t$. Moreover, to comprehensively reflect the generalization of model, we adopt the Harmonic Mean (HM) (Dixon & Chapman, 1980) defined as $\mathbf{HM} = \frac{2\times\mathbf{ACC}_s\times\mathbf{ACC}_t}{\mathbf{ACC}_s+\mathbf{ACC}_t}$.

Implementation Details. We implement our MDFNet with Pytorch as platform. The encoder of each client includes ResNet-50 pre-trianed on ImageNet dataset (Krizhevsky et al., 2012) without the last FC layer and two new additional FC layers (2,048→512→128) followed the ResNet-50. The decoder consists of two FC layers (256→512→2,048) and the classifier only includes one FC layer. The dimensions of $\hat{\mathbf{m}}^s$ and $\hat{\mathbf{m}}^I$ are both 2,048. For the training period, we fix the parameters of the pre-trained ResNet-50 and adopt the stochastic gradient descent (SGD) as the optimizer with momentum 0.9. Following (Zhang et al., 2019), the learning rate is adjusted by $\zeta_r = \frac{\zeta_0}{(1+10r)^{0.75}}$ where $\zeta_0 = 0.01$ and r is the communication round. The code is available in the supplementary.

4.2 Result Analysis

Table 1 and Table 2 report the average image recognition results in terms of random data split and various model training. According to the results, we achieve four meaningful and interesting conclusions as below.

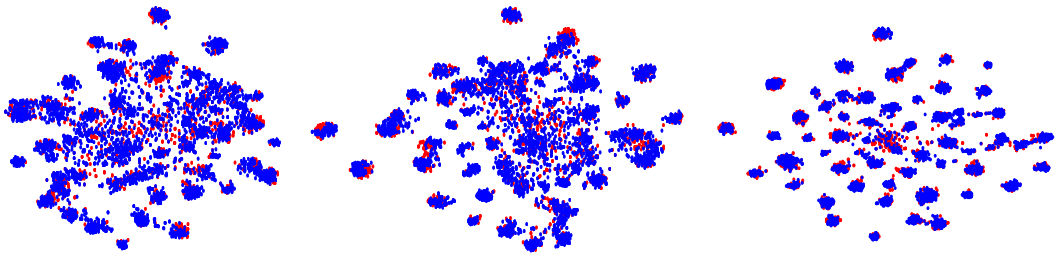
Table 1: Comparisons of Object Recognition Rate (%) for P^{2}MDCL on Image-CLEF benchmark. We adopt **bold** to highlight the best performance and underline to emphasis the second highest result.

	Method	C-I	C-P	I-C	I-P	P-C	P-I	Avg.
$\mathbf{ACC}_s$	Src-Only	97.98	97.98	95.33	95.33	73.70	73.70	89.00
	CDAN	96.67	93.33	90.00	91.00	68.00	68.00	84.50
	SRDC	95.00	95.33	93.00	94.33	72.67	76.33	87.78
	SHOT	97.33	96.00	94.33	94.67	73.19	77.99	88.92
	Ours	**98.50**	**98.33**	**97.00**	**97.00**	**77.30**	**80.67**	**91.47**
$\mathbf{ACC}_t$	Src-Only	82.00	69.83	92.00	76.17	87.90	85.46	82.23
	CDAN	86.17	73.66	**96.83**	76.50	93.33	85.67	85.36
	SRDC	91.50	75.16	94.16	76.83	93.33	90.33	86.89
	SHOT	90.00	74.67	94.83	78.17	**94.83**	90.86	87.23
	Ours	**93.00**	**77.33**	96.00	**79.33**	94.58	**92.83**	**88.85**
HM	Src-Only	89.28	81.54	93.64	84.68	80.18	79.15	84.74
	CDAN	91.12	82.34	93.29	83.12	78.68	75.82	84.93
	SRDC	93.22	84.05	93.58	84.69	81.71	82.74	86.67
	SHOT	93.52	84.00	94.58	85.63	82.62	83.93	87.38
	Ours	**95.67**	**86.57**	**96.49**	**87.28**	**85.09**	**86.32**	**90.13**

First, compared with others, the model learned by our training strategy achieves the best classification accuracy, when evaluated on the test set of labeled and unlabeled clients. In terms of the average harmonic metric, our MDFNet performs better than the second best SHOT by 3.4% on Office-Home. It illustrates that even if the data privacy hinders the currency of knowledge between these two clients, our method still employs the federated training paradigm to gradually eliminate the domain shift across different clients to improve the generalization of model. Second, although the UDA based solutions and SHOT effectively facilitates the well-trained source models to adapt the data distribution of unlabeled client, the progressive adaptation discards considerable source knowledge and results in performance degradation on the test set of labeled client. For instance, CDAN achieves better average accuracy on

Table 2: Comparisons of Object Recognition Rate (%) for P^{2}MDCL on Office-Home benchmark. We adopt **bold** to highlight the best performance and underline to emphasis the second highest result.

	Method	Ar-Cl	Ar-Pr	Ar-Rw	Cl-Ar	Cl-Pr	Cl-Rw	Pr-Ar	Pr-Cl	Pr-Rw	Rw-Ar	Rw-Cl	Rw-Pr	Avg.
ACC_s	Src-only	75.45	**75.45**	75.45	84.31	84.31	84.31	92.90	92.90	92.90	88.57	88.57	88.57	85.31
	CDAN	63.83	66.14	67.21	69.62	72.65	70.77	87.28	85.99	88.02	79.26	76.87	81.37	75.75
	SRDC	63.50	68.53	70.92	69.53	70.59	70.17	85.94	85.13	88.82	82.19	78.98	82.05	76.36
	SHOT	69.28	70.51	74.22	79.98	78.52	79.89	86.26	86.17	90.63	86.00	83.20	88.11	81.06
	Ours	**76.03**	75.12	**76.11**	**85.12**	**85.35**	**85.24**	**93.86**	**93.42**	**93.92**	**88.80**	**88.91**	**89.18**	**85.91**
ACC_t	Src-only	41.66	60.84	67.20	49.48	59.32	62.66	49.77	41.12	71.08	60.86	46.60	76.08	57.22
	CDAN	45.06	63.78	70.78	55.66	65.02	65.46	51.63	45.26	74.77	67.04	53.49	79.14	61.42
	SRDC	46.07	70.96	76.32	56.24	72.30	70.69	57.49	49.30	78.17	**68.34**	52.98	79.47	64.86
	SHOT	50.97	72.19	72.69	56.90	70.71	**72.18**	57.27	48.22	78.13	64.89	53.84	80.58	64.88
	Ours	**52.22**	**74.50**	**77.16**	**58.05**	**72.67**	72.11	**58.91**	**50.79**	**79.83**	66.84	**54.84**	**81.62**	**66.63**
HM	Src-only	53.68	67.36	71.09	62.36	69.64	71.89	64.82	57.01	80.54	72.15	61.07	81.85	67.79
	CDAN	52.83	64.94	68.95	61.86	68.62	68.01	64.88	59.31	80.86	72.64	63.08	80.24	67.18
	SRDC	53.40	69.72	73.52	62.18	71.43	70.43	68.89	62.44	83.16	74.63	63.42	80.74	69.50
	SHOT	58.73	71.34	73.45	66.49	74.41	75.84	68.84	61.84	83.92	73.97	65.37	84.18	71.53
	Ours	**61.91**	**74.80**	**76.63**	**69.03**	**78.50**	**78.12**	**72.38**	**65.80**	**86.30**	**76.27**	**67.78**	**85.19**	**75.00**



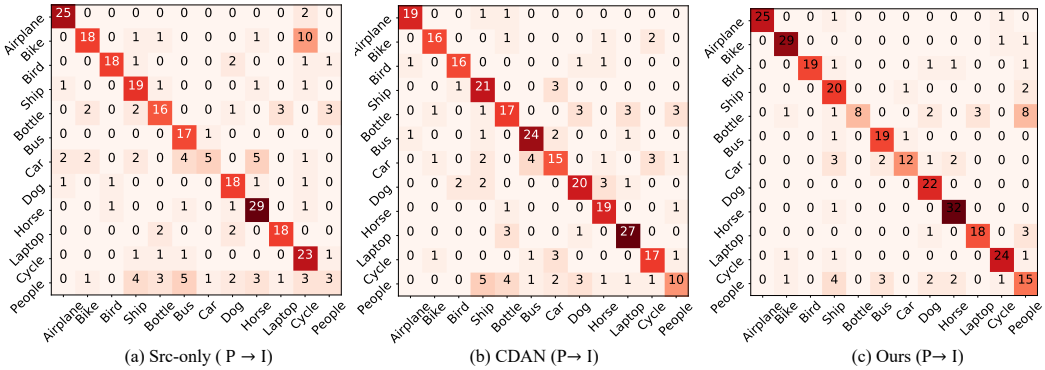
(a) Src-only (Ar→Rw) (b) SHOT (Ar→Rw) (c) Ours (Ar→Rw)

Figure 3: Comparison of feature visualization achieved by Src-only, SHOT and our MDFNet, where color red and blue represent the samples of labeled and unlabeled clients, respectively.

unlabeled client with office-home than source-only method, i.e., 61.42 v.s. 57.22%. However, such improvement heavily affects the generalization of CDAN on the labeled client, i.e., CDAN (75.75%) vs Src-only (85.31%). Different from CDAN, our method even attains the more improvement than the source-only method on the test set, especially for the task P→I of Image-CLEF dataset, where our MDFNet surpasses the source-only method by 6.9%. Thirdly, even though SHOT and MDFNet both protect the data privacy of source domains, our MDFNet learns a better hypothesis with lower error on the unlabeled client. Concretely, for the task Ar→Rw, when assessed on the unlabeled client, our method fights off SHOT by 4.5%, which means our training manner captures more knowledge from the labeled client via the frequent communication between server and clients to improve the discriminative ability of model on recognizing unlabeled instances. Finally, we find that all methods achieve higher classification accuracy on unlabeled clients than that of labeled clients for tasks P→C and P→I. Specifically, with P as the source domain, the well-trained Src-Only model achieves better performance on the target domain than that of source test set. The main reason lies in the fact that the samples of P domain lie in a more diverse distribution within class, which makes it difficult to learn a discriminative model to recognize its images. Concretely, P domain includes many images with multiple objects but one single label. For example, one image of "bird" class in P domain involves bird and dog, and another image includes bird and person. However, there are almost no such multi-object images in C and I domains. Moreover, the same class in P domain has more animals than that of C or I domain. For instance, besides several common birds as C and I domains, the bird class of P domain also consists chicken and ostrich, etc. Thus, the well-trained source model with P domain can easily classify the images of C and I domains, but difficultly recognize its source test set.

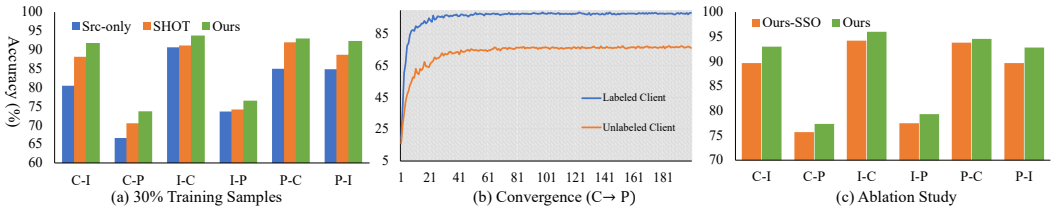
4.3 Empirical Analysis

Feature Visualization & Confusion Matrix. Our MDFNet aims to eliminate the distribution discrepancy across different clients and improve the generalization of model with the data privacy protection. Thus, we extract the hidden features from Src-only, SHOT and MDFNet on task Ar→Rw



(a) Src-only (P → I) (b) CDAN (P→ I) (c) Ours (P→ I)

Figure 4: Comparison of confusion matrix achieved by Src-only, CDAN and our MDFNet.



(a) 30% Training Samples (b) Convergence (C→ P) (c) Ablation Study

Figure 5: Ablation study and Convergence with experiments on Image-CLEF dataset.

and follow (Zhang et al., 2019) to draw the the feature embedding of the test samples of labeled client and those of unlabeled client in 2-D canvas. From Figure 3, we notice that our MDFNet achieves better alignment across these two clients and significantly promotes the discriminative ability of model as class boundaries are explicit among various categories. Moreover, we utilize the confusion matrix to analyse the model performance on the test set of labeled client (P domain of Image-CLEF). As Figure 4 shows, our method accurately distinguishes several similar objects as bike and motorcycle when compared with Src-only, which illustrates our MDFNet transfers the valuable semantic from unlabeled client to assist model recognizing the samples of labeled client.

Ablation Study & Convergence. To reveal the importance of adaptive self-supervised optimization module, we attempt to remove the supervisor of pseudo-label and consider it as a variant (Ours-SSO) of MDFNet and evaluate the model on the unlabeled client. Their comparison is reported in Figure 5 (c) where the variant suffers from the obvious performance degradation, which demonstrates the self-supervised learning based module effectively facilitates the model to learn more discriminative features. In addition, we further adjust the number of training samples in the labeled client (training set/test set = 3:7) and contrast the performances of three methods on the unlabeled client. The result in Figure 5 (a) reports our MDFNet still beats others in most tasks even with insufficient well-annotated samples. Finally, we record the relationship between classification accuracy and the communication round in Figure 5 (b) where the model is assessed on the test set of labeled client and unlabeled client. The performance means MDFNet rapidly achieves the convergence and has no negative effect on recognizing samples of labeled client during adaptation.

5 Conclusion

Although UDA based methods effectively avoid performance degradation when applying source knowledge to target domain, the UDA assumption ignores the improvement of model generalization on source domain and conflicts with privacy protection. Thus, this paper formulates these practical demands with domain adaptation as a novel scenario P^{2}MDCL and proposes mask-driven federated network (MDFNet) to address this challenge. Concretely, each individual domain explores mask disentangle mechanism to learn domain-invariant features, and the unlabeled clients exploits adaptive self-supervised optimization to generate high-quality pseudo labels facilitating discriminative feature learning. Moreover, the centralized server refines the global invariant by assembling local knowledge across all domains. Finally, theoretical and experimental analysis demonstrate the rationality and effectiveness of our MDFNet on solving P^{2}MDCL problem.

REFERENCES

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29:343–351, 2016.

Christian Buchta, Martin Kober, Ingo Feinerer, and Kurt Hornik. Spherical k-means clustering. *Journal of statistical software*, 50(10):1–22, 2012.

Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pp. 301–318. Springer, 2020.

Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. A joint learning and communications framework for federated learning over wireless networks. *IEEE Transactions on Wireless Communications*, 2020.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12455–12464, 2020.

Kenneth R Dixon and Joseph A Chapman. Harmonic mean measure of animal activity areas. *Ecology*, 61(5):1040–1044, 1980.

Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4023–4032, 2020.

Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9101–9110, 2020.

Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. *ICLR*, 2021.

Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 4816–4827. PMLR, 2020.

Guoliang Kang, Lu Jiang, Yunchao Wei, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4544–4553, 2020.

Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 615–625, 2021.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9650, 2020.

Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Exploring uncertainty in pseudo-label guided unsupervised domain adaptation. *Pattern Recognition*, 96:106996, 2019.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. Federated semi-supervised medical image classification via inter-client relation matching. *MICCAI*, 2021.

Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4):70–82, 2020.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.

Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3130–3139, 2020.

Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1094–1103, 2021.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning*, pp. 5102–5112. PMLR, 2019.

Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *ICLR 2020*, 2020.

Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pp. 234–244. Springer, 2016.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5940–5947, 2020.

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8725–8735, 2020.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pp. 5423–5432. PMLR, 2018.

Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2272–2281, 2017.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5031–5040, 2019.

Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020.