Confidence-based Visual Dispersal for Few-shot Unsupervised Domain Adaptation

Yizhe Xiong[1,2,3] Hui Chen[2*] Zijia Lin[1] Sicheng Zhao[2] Guiguang Ding[1,2*]
[1]School of Software, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology (BNRist)
[3]Hangzhou Zhuoxi Institute of Brain and Intelligence
{xiongyizhe2001, jichenhui2012, schzhao}@gmail.com
linzijia07@tsinghua.org.cn dinggg@tsinghua.edu.cn

Abstract

Unsupervised domain adaptation aims to transfer knowledge from a fully-labeled source domain to an unlabeled target domain. However, in real-world scenarios, providing abundant labeled data even in the source domain can be infeasible due to the difficulty and high expense of annotation. To address this issue, recent works consider the Few-shot Unsupervised Domain Adaptation (FUDA) where only a few source samples are labeled, and conduct knowledge transfer via self-supervised learning methods. Yet existing methods generally overlook that the sparse label setting hinders learning reliable source knowledge for transfer. Additionally, the learning difficulty difference in target samples is different but ignored, leaving hard target samples poorly classified. To tackle both deficiencies, in this paper, we propose a novel Confidence-based Visual Dispersal Transfer learning method (C-VisDiT) for FUDA. Specifically, C-VisDiT consists of a cross-domain visual dispersal strategy that transfers only high-confidence source knowledge for model adaptation and an intra-domain visual dispersal strategy that guides the learning of hard target samples with easy ones. We conduct extensive experiments on Office-31, Office-Home, VisDA-C, and DomainNet benchmark datasets and the results demonstrate that the proposed C-VisDiT significantly outperforms state-of-the-art FUDA methods. Our code is available at `https://github.com/Bostoncake/C-VisDiT`.

1. Introduction

Deep learning has achieved considerably high performance in various computer vision tasks, such as instance recognition [13, 51], semantic segmentation [23, 21, 2], and object detection [32, 6, 22]. Despite the remarkable success, well-tuned deep models usually suffer from dramatic performance drops when being applied in real-world scenarios because of the domain gap issue [38, 39, 54, 44], *i.e*. the shift between the learnt distribution and the testing distribution. To cope with such performance degradation, researchers have been dedicated to unsupervised domain adaptation (UDA) [23, 9, 39, 25] aiming at transferring the learned knowledge from a labeled source domain to another unlabeled target domain.

Though being promising, UDA assumes that a fully-annotated source domain can be conveniently accessed for training. However, in some real-world scenarios, such an assumption is violated, ending up with sparsely labeled source data. For example, in the field of chip defect detection, it is unrealistic to perfectly annotate massive defect data in the source domain, due to the strict quality regulation and high expense of annotation [7, 15]. In the medical industry, the high cost and difficulty of labeling also prohibit providing sufficient labeled source data [42, 53]. Therefore, how to realize UDA with sparsely labeled source data becomes appealing and has attracted increasing attention [17, 50, 1, 31].

Recently, researchers address this issue by considering the Few-shot Unsupervised Domain Adaptation (FUDA), where only a few source samples are labeled. Compared with UDA, FUDA is more challenging due to the insufficiency of the source knowledge. With only a few labeled source samples, deep models are difficult to learn a discriminative feature space to represent the complicated semantic information, thus largely affecting the knowledge transfer. And thus to compensate for the shortage of labeled source data, existing methods [17, 50] generally employ self-supervised learning strategies to mine the latent correspondence between the source distribution and the target one. Specifically, Kim *et al*. [17] proposed a cross-

*Corresponding author.

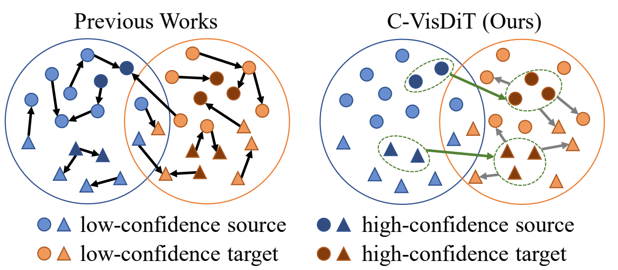

Figure 1: We address the task of few-shot unsupervised domain adaptation. Left: Existing FUDA methods utilize self-supervised learning methods to implicitly align source and target domains without considering the confidence of samples. Right: Our C-VisDiT transfers knowledge from high-confidence source samples and high-confidence target samples to enhance model adaptation.

domain self-supervised learning method (CDS) to derive features that are not only domain-invariant but also class-discriminative for the target domain. Yue *et al*. [50] further enhanced the adaptation with an end-to-end prototypical cross-domain self-supervised learning method (PCS) which achieves the state-of-the-art performance.

Despite great progress, existing self-supervised FUDA methods transfer knowledge from all visible samples without considering the discrimination of their individual confidence towards model adaptation, which exhibits two inherent weaknesses. First, existing methods generally align target features with all source features including unlabeled ones. Although having performance gains, such a strategy discards that the sparse label setting of FUDA inevitably introduces large noise, resulting in the confidence inconsistency among source samples and thus transfering unreliable knowledge. Second, existing methods generally feed all target data into the self-supervised learning modules, assuming that the model optimizes equally on each target sample. However, difficult target samples, often presented as samples with low prediction confidence, are not well-learned during the adaptation.

To tackle those issues, in this paper, we propose a novel **C**onfidence-based **Vis**ual **Di**spersal **T**ransfer learning method (**C-VisDiT**) for FUDA, aiming to comprehensively consider the different importance of each sample based on its confidence, as illustrated in Fig. 1. Considering the huge difficulty of learning a neat representation with very few labeled source samples, we propose to disperse the knowledge directly at the visual[1] level. Specifically, for the source domain, we introduce a cross-domain visual dispersal to enhance the knowledge transfer from high-confidence source samples to unlabeled target ones. Without the disturbance of unreliable low-confidence source knowledge, our cross-domain visual dispersal can achieve more reliable cross-domain knowledge transfer. Besides, we perform an intra-domain visual dispersal between samples of different learning difficulties for the target domain, aiming to guide the learning of hard target samples by the easy ones. In contrast to existing methods, our intra-domain visual dispersal can diminish the learning disparity between target samples of different learning difficulties, thus boosting the model adaptation.

We emphasize that the core contribution of our C-VisDiT is the confidence-based strategies, motivated by the idea that samples with low prediction confidence greatly harm the training in the label-scarse scenario of FUDA. Although we conduct visual dispersal via MixUp [52], we show that our C-VisDiT and its variants can greatly outperform their vanilla MixUp-based alternatives which completely discard the confidence of samples, well confirming the importance of the proposed confidence-based transfer learning for FUDA.

To summarize, our contributions are three-fold:

- We propose a confidence-based visual dispersal transfer learning method (C-VisDiT) for FUDA, which simultaneously takes the reliability of the source samples and the learning difficulty of the target samples into consideration.
- We introduce a cross-domain visual dispersal to avoid the negative impact of source knowledge with low confidence. An intra-domain visual dispersal is also employed to boost the learning of hard unlabeled target samples with the guidance of easy ones.
- We conduct extensive experiments on four popular benchmark datasets for FUDA, *i.e.*, Office-31 [33], Office-home [41], VisDA-C [28], and DomainNet [27]. Experiment results show that our C-VisDiT consistently outperforms existing methods with a significant margin, establishing new state-of-the-art results for all datasets. That well demonstrate the effectiveness and superiority of our method.

2. Related Work

Unsupervised Domain Adaptation (UDA). UDA aims to transfer the knowledge from a fully labeled source domain to an unlabeled target domain. Most UDA methods tend to learn domain-invariant feature representations. Earlier works in UDA aim to optimize feature extractors explicitly by domain divergence minimization [11, 24, 26, 56, 37], adversarial domain classifier confusion [39, 14, 25], or cross-domain data reconstruction [10, 5]. Recent works focus on the domain mapping [36, 19, 4], which translates the source domain images to the target domain via conditional generative adversarial networks [55, 49, 18]. Existing UDA methods utilize the abundant labeled source data, which is

[1] Here we refer to "visual" as the raw RGB image information.

not well satisfied in some practical scenarios, like chip defect detection [7] and medical image segmentation [29].

Although designed for FUDA, our method can be adopted to the UDA problem. Empirical analysis show that the proposed confidence-based strategy can achieve slight improvements over baseline models. This is reasonable because the abundant source annotations can substantially guarantee the reliability of knowledge transfer, which, however, does not hold in FUDA. Therefore, we advocate thoughtfully managing each sample for both the source domain and the target domain, making our C-VisDiT possess a distinctive edge for FUDA.

Few-shot Unsupervised Domain Adaptation (FUDA). FUDA transfers knowledge from a sparsely labeled source domain to an unlabeled target domain. Existing methods implicitly align the source and the target domains with self-supervised learning. Kim *et al*. [17] and Rakshit *et al*. [31] utilized an instance-level self-supervised contrastive learning for representation learning. As the instance-level training is sensitive to outliers, Yue *et al*. [50] utilized the self-supervised learning via class prototypes to enhance model performance and robustness. Despite the effectiveness, existing methods generally discard the confidence of different samples, leading to sub-optimal adaptation. Therefore, we propose a novel confidence-based transfer learning method to mitigate such weakness.

MixUp in UDA. MixUp [52] is an augmentation strategy where linearly interpolated samples are introduced for training. MixUp has been extensively explored in UDA. For example, Wu *et al*. [45] and Xu *et al*. [46] utilize MixUp to create complex samples for adversarial learning. Yan *et al*. [48] and Ding *et al*. [8] directly conduct the MixUp regularization along with the model training process to further boost model performance. In the FUDA task, to the best of our knowledge, no previous work has applied MixUp, mainly because the severely sparse annotation (*e.g*., 1-shot) prohibits providing reliable samples for adaptation, thus harming the capability of MixUp. In our work, we show that with the proposed confidence-based strategy, we can conveniently regain the power of MixUp for FUDA, establishing new state-of-the-art performance on all standard benchmarks for FUDA.

3. Methodology

3.1. Preliminary

We explore the problem of FUDA in which there are two data domains with different data distributions. As for the source domain, we have sparsely labeled source data $\mathcal{D}_{ls} = \left\{(\mathbf{x}_i^{ls}, y_i^{ls})\right\}_{i=1}^{N_{ls}}$ and the remaining unlabeled source data $\mathcal{D}_{us} = \{(\mathbf{x}_i^{us})\}_{i=1}^{N_{us}}$, where $N_{us} \gg N_{ls}$. For the target domain, only unlabeled data $\mathcal{D}_{ut} = \left\{(\mathbf{x}_i^{ut})\right\}_{i=1}^{N_{ut}}$ are provided. Following previous works [17, 50], we focus on the classification task, where $\mathcal{D}_{ls}$, $\mathcal{D}_{us}$ and $\mathcal{D}_{ut}$ are *all visible* during the model training process.

3.2. Baseline Model

We first introduce a generic solution as our baseline model. Specifically, following [17, 50], for a given image $\mathbf{x}$, we leverage a convolutional neural network (CNN), denoted as $F(\cdot)$, to extract its feature. A softmax layer, denoted as $\phi(\cdot)$, is employed as the classifier to predict the semantic distribution $p(y|\mathbf{x})$. The forward process can be formulated as follows:

$$p(y|\mathbf{x}) = \phi(F(\mathbf{x})) = [p_1, p_2, \cdots, p_c] \tag{1}$$

where c is the number of different categories.

During training, for the labeled source data, we employ the standard cross-entropy objective $\mathcal{L}_{CE}$:

$$\mathcal{L}_{cls} = \sum_{j}^{N_{ls}} \mathcal{L}_{CE}(p(y_j^{ls}|\mathbf{x}_j^{ls}), y_j^{ls}) \tag{2}$$

For the unlabeled target data, following [20], we utilize the mutual information maximization objective to encourage individually certain and globally diverse predictions:

$$\mathcal{L}_{MI} = -\mathcal{H}\left(\sum_{i}^{N_{ut}}(p(y|\mathbf{x}_i^{ut}))\right) + \frac{1}{N_{ut}}\sum_{i}^{N_{ut}}\mathcal{H}(p(y|\mathbf{x}_i^{ut})) \tag{3}$$

where the entropy metric $\mathcal{H}(p(y|\mathbf{x})) = \sum_{i=1}^{c} p_i \log p_i$.

In order to enhance the cross-domain feature alignment, we further adopt the prototypical self-supervised learning method [50] to optimize the whole network. We denote it as $\mathcal{L}_{self}$ and leave its details to [50] due to the page limit.

Finally, the overall learning objective of the baseline model can be derived as:

$$\mathcal{L}_{\text{Baseline}} = \mathcal{L}_{cls} + \lambda_{MI} \cdot \mathcal{L}_{MI} + \lambda_{self} \cdot \mathcal{L}_{self} \tag{4}$$

where λ_{MI} and λ_{self} are two hyper-parameters.

3.3. Cross-domain Visual Dispersal (X-VD)

Transferring knowledge from the source domain to the unlabeled target domain in FUDA is notably difficult, due to the sparsely labeled source data. As we observe, in the Clipart-to-Art task with 3% labeled source on the Office-Home [41] dataset, existing state-of-the-art methods, *i.e*., CDS [17] and PCS [50], only achieve 38.7% and 45.5% classification accuracy on the *unlabeled source dataset*, *i.e*., $\mathcal{D}_{us}$, respectively. This indicates that involving $\mathcal{D}_{us}$ for knowledge transfer inevitably introduces noise to the model adaptation. As a result, simply conducting the self-supervised contrastive learning with these unreliable source samples as in [17, 50] unavoidably leads target samples to

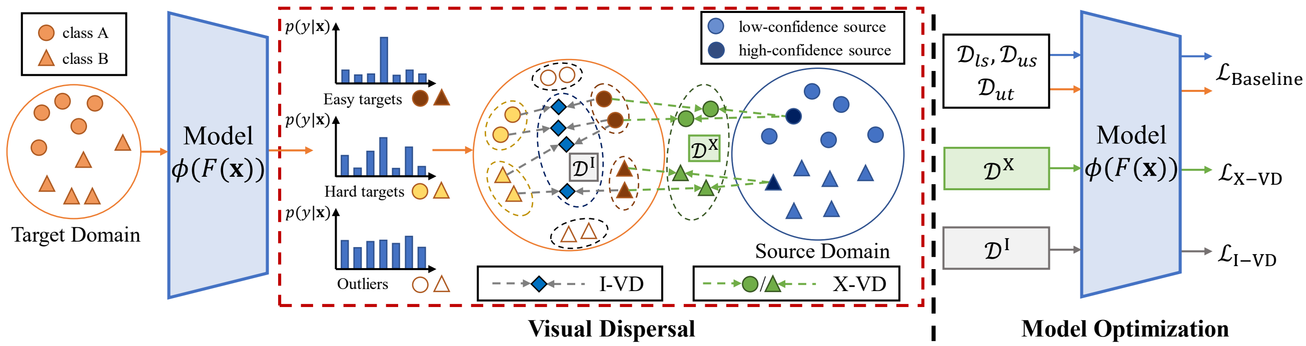

Figure 2: An overview of the proposed C-VisDiT method. X-VD and I-VD strategies are performed across and within domains, respectively, via visual dispersal. We carefully consider the confidence of samples, ending up with a cross-blending dataset $\mathcal{D}^{\mathrm{X}}$ and an intra-mixing dataset $\mathcal{D}^{\mathrm{I}}$. Best viewed in color.

be aligned to inappropriate source semantics, thus affecting target classification performance. Therefore, to remedy such an issue, we propose a cross-domain visual dispersal (**X-VD**) strategy to enhance the knowledge transfer by concentrating on the high-confidence source samples.

Nearest Source Matching. We aim to transfer knowledge from source samples that are highly correlated with the target samples. For each target sample, we perform a nearest sample matching to associate it with its most similar labeled source sample. Specifically, we first calculate a score matrix $M = \{m_{ij}\}_{N_{ut} \times N_{ls}}$ in each training epoch. m_{ij} is the confidence score for a pair of samples $(\mathbf{x}_i^{ut}, \mathbf{x}_j^{ls})$, which is defined by the Euclidean distance between the corresponding features:

$$m_{ij} = \|F(\mathbf{x}_i^{ut}) - F(\mathbf{x}_j^{ls})\|_2 \tag{5}$$

Then we derive the nearest source sample for each target sample by finding the column index $\bar{j}_i$ of the minimum m_{ij} in each row of M:

$$\bar{j}_i = \arg\min_j (m_{ij}) \tag{6}$$

With the index $\bar{j}_i$, we can then locate the nearest labeled source sample for unlabeled target samples, forming a high-confidence source dataset, denoted by $\mathcal{D}_{ut\leftarrow ls} = \{(\mathbf{x}_i^{ut\leftarrow ls}, y_i^{ut\leftarrow ls})\}_{i=1}^{N_{ut}}$:

$$(\mathbf{x}_i^{ut\leftarrow ls}, y_i^{ut\leftarrow ls}) = (\mathbf{x}_{\bar{j}_i}^{ls}, y_{\bar{j}_i}^{ls}) \tag{7}$$

High-confidence Target Sampling. To reduce the impact of mismatched sample pairs which have different semantics, we leave out those unreliable target samples and focus on those with high-confidence. Specifically, we set r_h^{X} as the fixed ratio of high-confidence target samples and find the corresponding indices of $N_{\mathrm{X}} = r_h^{\mathrm{X}} N_{ut}$ target samples with the lowest entropy, formulated as:

$$O_{conf} = \arg \mathrm{TopK}(-\mathcal{H}(\mathbf{x}_i^{ut}), N_{\mathrm{X}}) \tag{8}$$

where the function $\arg \mathrm{TopK}(\cdot,\cdot)$ returns the indices of N_{X} largest elements of $-\mathcal{H}(\mathbf{x}_i^{ut})$. Then, the high-confidence target set $\mathcal{D}_{ut}^{conf}$ and its corresponding nearest labeled source sample set $\mathcal{D}_{ut\leftarrow ls}^{conf}$ are:

$$\begin{aligned} \mathcal{D}_{ut}^{conf} &= \mathbb{I}(i \in O_{conf})\mathcal{D}_{ut} \\ \mathcal{D}_{ut\leftarrow ls}^{conf} &= \mathbb{I}(i \in O_{conf})\mathcal{D}_{ut\leftarrow ls} \end{aligned} \tag{9}$$

Cross-domain Knowledge Transferring via Visual Dispersal. We utilize the MixUp augmentation [52] to create a cross-blending dataset $\mathcal{D}^{\mathrm{X}} = \{(\mathbf{x}_i^{\mathrm{X}}, y_i^{\mathrm{X}})\}_{i=1}^{N_{\mathrm{X}}}$:

$$\begin{aligned} \mathbf{x}_i^{\mathrm{X}} &= \beta \bar{\mathbf{x}}_i^{ut} + (1-\beta)\bar{\mathbf{x}}_i^{ut\leftarrow ls} \\ y_i^{\mathrm{X}} &= \beta y_i^{ut} + (1-\beta) y_i^{ut\leftarrow ls} \end{aligned} \tag{10}$$

where $y_i^{ut} = \arg\max_j p(y|\bar{\mathbf{x}}_i^{ut})_j$ is the one-hot pseudo label for $\bar{\mathbf{x}}_i^{ut} \in \mathcal{D}_{ut}^{conf}$ and $y_i^{ut\leftarrow ls}$ is the one-hot ground-truth of $\bar{\mathbf{x}}_i^{ut\leftarrow ls} \in \mathcal{D}_{ut\leftarrow ls}^{conf}$. $\beta \sim \mathrm{Beta}(\alpha,\alpha)$ is sampled from a beta distribution. As a result, the source knowledge can be well dispersed to the target domain at the visual level, rather than from the latent representation level in CDS [17] and PCS [50].

Then, our cross-domain visual dispersal loss can be formulated with the cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{\text{X-VD}} = \sum_{(\mathbf{x},y)\in\mathcal{D}^{\mathrm{X}}} \mathcal{L}_{CE}(p(y|\mathbf{x}), y) \tag{11}$$

Discussion. Cross-domain visual dispersal transfers only high-confidence source knowledge to target domain, reducing the disturbance of noisy supervisions. During the transfer process, the nearest sample matching strategy intends to match similar samples as pairs. Therefore, we can regard each sample in $\mathcal{D}^{\mathrm{X}}$ as a fusion of two independent samples from two similar semantic distributions, although they belong to different domains. Benefited from the visual dispersal, *i.e.*, Eq. (10), the adaptation model can be encouraged to pull two samples with similar semantics but from

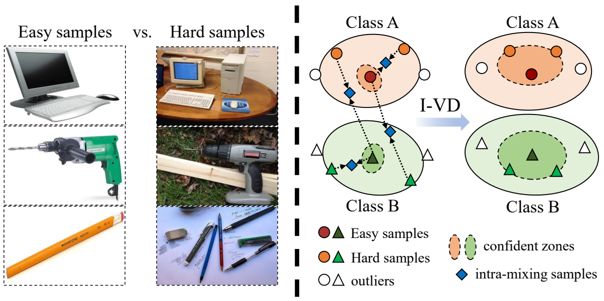

Figure 3: Different target samples present different difficulties to the model learning because of various differences, *e.g.*, object size and background information. Due to the guidance of intra-mixing samples, I-VD gradually fixes the decision boundaries for the target domain, improving the adaptation on hard target samples.

different domains close to each other, realizing domain-invariant feature learning. As a result, the domain gap can be mitigated with a more reliable knowledge transfer.

3.4. Intra-domain Visual Dispersal (I-VD)

Hard samples widely exist during the model learning process [30, 35, 47]. For example, in the Real World domain on the Office-Home dataset [41], samples with complicated backgrounds and small objects are more likely to have low-confidence probability distributions compared with samples with only one object brimming the image, as shown in Fig. 3. Therefore, applying the same learning metrics to all target samples without differentiation makes the model overfit easy target samples, which results in decision boundaries misclassifying hard samples. However, existing methods [17, 50] usually ignore this phenomenon. To enhance the classification of hard target samples, we propose an intra-domain visual dispersal (**I-VD**) strategy to guide them with easy ones.

Confidence-based Target Domain Splitting. The learning difficulty of a target sample is correlated to the confidence of its predicted probability distribution, which can be measured via entropy, as large prediction entropy leads to uncertain model prediction [43]. To differentiate these samples, we split the target samples into three groups based on their prediction entropy. The number of samples in each group is denoted as $N_1 = r^{\mathrm{I}}_E N_{ut}$, $N_2 = r^{\mathrm{I}}_H N_{ut}$, and $N_3 = N_{ut} - N_1 - N_2$, respectively, controlled by hyper-parameters r^{I}_E and r^{I}_H. Specifically, the indices of samples in three groups are derived as:

$$\begin{gathered} O = \arg\mathrm{Sort}(\mathcal{H}(p(y|\mathbf{x}^{ut}_i))) \\ O_1 = O[0{:}N_1],\ O_2 = O[N_1{:}N^{'}],\ O_3 = O[N^{'}{:}N_{ut}] \end{gathered} \tag{12}$$

where $\arg\mathrm{Sort}(\cdot) = [i_1, i_2, \cdots, c_n]$ returns the indices that would sort an array in ascending order and $N^{'} = N_1 + N_2$. Therefore, we define the easy target sample set $\mathcal{D}^E_{ut}$ and the hard target sample set $\mathcal{D}^H_{ut}$ as:

$$\mathcal{D}^E_{ut} = \mathbb{I}(i \in O_1)\mathcal{D}_{ut},\ \ \mathcal{D}^H_{ut} = \mathbb{I}(i \in O_2)\mathcal{D}_{ut} \tag{13}$$

We exclude the O_3 outlier samples $\mathcal{D}^O_{ut}$ for training because they are of high uncertainty and thus not trustworthy.

Intra-domain Sample Guidance via Visual Dispersal. To guide the learning of hard target samples, we disperse the easy target data at the visual level by leveraging the MixUp strategy [52]. We find it helpful to add a small set of high-confidence labeled source samples to the easy target samples which enlarges the easy sample set. For ease of explanation, we denote the final training easy sample set by $\mathcal{T}^E_{ut} = \mathcal{D}_{ls} \cup \mathcal{D}^E_{ut}$, and denote the training hard sample set by $\mathcal{T}^H_{ut} = \mathcal{D}^H_{ut}$. For each $\bar{\mathbf{x}}^H_i \in \mathcal{T}^H_{ut}$, we choose a random $\bar{\mathbf{x}}^E_j \in \mathcal{T}^E_{ut}$ to guide its learning by conducting image-level MixUp:

$$\begin{aligned} \mathbf{x}^{\mathrm{I}}_i &= \beta\bar{\mathbf{x}}^E_j + (1-\beta)\bar{\mathbf{x}}^H_i \\ y^{\mathrm{I}}_i &= \beta y^E_j + (1-\beta) y^H_i \end{aligned} \tag{14}$$

where one-hot pseudo label $y^H_i = \arg\max_k p(y|\bar{\mathbf{x}}^H_i)_k$, and $\beta \sim \mathrm{Beta}(\alpha, \alpha)$ sampled from a beta distribution. For y^E_j, we have:

$$y^E_j = \begin{cases} y^s & \text{if } \bar{\mathbf{x}}^E_j \in \mathcal{D}_{ls} \\ \arg\max_k p(y|\bar{\mathbf{x}}^E_j)_k & \text{if } \bar{\mathbf{x}}^E_j \in \mathcal{D}^E_{ut} \end{cases} \tag{15}$$

With the transformations above, we construct an intra-mixing dataset $\mathcal{D}^{\mathrm{I}} = \{(\mathbf{x}^{\mathrm{I}}_i, y^{\mathrm{I}}_i)\}^{N_{\mathrm{I}}}_{i=1}$, where $N_{\mathrm{I}} = |\mathcal{T}^H_{ut}|$. We directly employ the cross-entropy loss to optimize the intra-mixing samples:

$$\mathcal{L}_{\text{I-VD}} = \sum_{(\mathbf{x},y)\in\mathcal{D}^{\mathrm{I}}} \mathcal{L}_{CE}(p(y|\mathbf{x}), y) \tag{16}$$

Discussion. The constructed intra-mixing samples act as guidance towards the learning of hard samples. As illustrated in Fig. 3, our intra-mixing samples are located between the easy target samples and the hard ones. Through learning on the probability distributions of the intra-mixing samples, our model can reduce the undesirable prediction oscillations between the easy samples and the hard ones, as revealed in [52]. Due to the fact that model predictions on easy samples are more consistent towards model optimization [3], our model in turn enhances the learning of hard samples, thus diminishing the learning disparity between easy samples and hard ones. Therefore, our I-VD incisively leads to better performance on hard target samples.

3.5. C-VisDiT for the FUDA Problem

Our C-VisDiT learning method mainly introduces cross-domain and intra-domain visual dispersal strategies for the FUDA problem. Together with the baseline training loss, our final training objective is expressed by:

$$\mathcal{L} = \mathcal{L}_{\text{Baseline}} + \lambda_{\text{X-VD}}\mathcal{L}_{\text{X-VD}} + \lambda_{\text{I-VD}}\mathcal{L}_{\text{I-VD}} \tag{17}$$

Table 1: Adaptation accuracy (%) comparison on 1-shot / 3-shots labeled source per class on the Office-31 dataset.

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg
Source Only	31.3 / 49.0	19.6 / 43.3	41.3 / 55.7	58.7 / 81.8	39.9 / 49.8	61.9 / 81.7	42.1 / 60.2
MME [34]	21.5 / 51.0	12.2 / 54.6	23.1 / 60.2	60.9 / 89.7	14.0 / 52.3	62.4 / 91.4	32.3 / 66.5
CDAN [25]	11.2 / 43.7	6.2 / 50.1	9.1 / 65.1	54.8 / 91.6	10.4 / 57.0	41.6 / 89.8	22.2 / 66.2
MDDIA [16]	45.0 / 62.9	54.5 / 65.4	55.6 / 67.9	84.4 / 93.3	53.4 / 70.3	79.5 / 93.2	62.1 / 75.5
CDS [17]	52.6 / 65.1	55.2 / 68.8	65.7 / 71.2	76.6 / 88.1	59.7 / 71.0	73.3 / 87.3	63.9 / 75.3
PCS [50]	60.2 / 78.2	69.8 / 82.9	**76.1** / 76.4	90.6 / 94.1	71.2 / 76.3	91.8 / 96.0	76.6 / 84.0
C-VisDiT (Ours)	**74.1 / 82.7**	**72.3 / 86.0**	75.7 / **76.5**	**93.2 / 95.0**	**76.4 / 76.9**	**94.2 / 97.0**	**81.0 / 85.7**

Table 2: Adaptation accuracy (%) comparison on 3% and 6% labeled source samples per class on the Office-Home dataset.

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
3% labeled source													
Source Only	22.5	36.5	41.1	18.5	29.7	28.6	27.2	25.9	38.4	33.5	20.3	41.4	30.3
MME [34]	4.5	15.4	25.0	28.7	34.1	37.0	25.6	25.4	44.9	39.3	29.0	52.0	30.1
CDAN [25]	5.0	8.4	11.8	20.6	26.1	27.5	26.6	27.0	40.3	38.7	25.5	44.9	25.2
MDDIA [16]	21.7	37.3	42.8	29.4	43.9	44.2	37.7	29.5	51.0	47.1	29.2	56.4	39.1
CDS [17]	43.8	55.5	60.2	51.5	56.4	59.6	51.3	46.4	64.5	62.2	52.4	70.2	56.2
PCS [50]	42.1	61.5	63.9	52.3	61.5	61.4	58.0	47.6	73.9	66.0	**52.5**	75.6	59.7
C-VisDiT (Ours)	**44.1**	**66.8**	**67.0**	**54.9**	**66.4**	**66.8**	**60.5**	**47.9**	**75.7**	**67.2**	51.6	**78.8**	**62.3**
6% labeled source													
Source Only	26.5	41.3	46.7	29.3	40.4	37.9	35.5	31.6	57.2	46.2	32.7	59.2	40.4
MME [34]	27.6	43.2	49.5	41.1	46.6	49.5	43.7	30.5	61.3	54.9	37.3	66.8	46.0
CDAN [25]	26.2	33.7	44.5	34.8	42.9	44.7	42.9	36.0	59.3	54.9	40.1	63.6	43.6
MDDIA [16]	25.1	44.5	51.9	35.6	46.7	50.3	48.3	37.1	64.5	58.2	36.9	68.4	50.3
CDS [17]	45.4	60.4	65.5	54.9	59.2	63.8	55.4	**49.0**	71.6	66.6	**54.1**	75.4	60.1
PCS [50]	46.1	65.7	69.2	57.1	64.7	66.2	61.4	47.9	75.2	67.0	53.9	76.6	62.6
C-VisDiT (Ours)	**46.5**	**69.8**	**72.5**	**60.2**	**71.2**	**71.2**	**64.1**	**49.0**	**78.5**	**69.1**	52.8	**80.1**	**65.4**

Table 3: Adaptation accuracy (%) comparison on 1% labeled source samples per class on the VisDA-C dataset. "SO" denotes training with only labeled source samples (Source Only).

Method	SO	MME [34]	CDAN [25]	CDS [17]	PCS [50]	C-VisDiT
Acc.	42.6	69.4	61.5	69.4	79.0	**80.5**

Table 4: Adaptation accuracy (%) comparison on 1-shot and 3-shots labeled source per class on the DomainNet dataset. "SO" denotes training with only labeled source samples (Source Only).

Method	SO	MME [34]	CDAN [25]	CDS [17]
1-shot	13.1	17.5	14.6	21.5
3-shots	27.1	29.1	27.3	42.5
Method	PCS [50]	C-VisDiT	BrAD [12]	BrAD+C-VisDiT
1-shot	34.7	**35.9**	49.7	**51.0**
3-shots	46.1	**48.1**	60.9	**63.4**

4. Experiments

4.1. Experimental Setting

Datasets. We evaluate our method on four widely-used benchmark datasets for FUDA, *i.e.*, **Office-31** [33], **Office-Home** [41], **VisDA-C** [28], and **DomainNet** [27], following prior works [17, 50]. For Office-31, We conduct experiments with 1-shot and 3-shots source labels per class. For Office-home, we employ 3% and 6% labeled source images per class. For VisDA-C, we validate our model with 1% labeled source images per class. For DomainNet, we employ 1-shot and 3-shots source labels per class. Statistics of benchmarks can be found in the supplementary materials.

For fair comparison, we implement our model following PCS [50] and BrAD [12]. More implementation details are provided in the supplementary materials.

4.2. Comparison with State-of-the-art Methods

We choose the baseline methods as PCS [50], including MME [34], CDAN [25], MDDIA [16], and CDS [17]. Extensive experiments are conducted on the Office-31, Office-Home, VisDA-C, and DomainNet datasets, with the corresponding results in terms of classification accuracy reported in Tab. 1, Tab. 2, Tab. 3, and Tab. 4, respectively. On the DomainNet dataset, we also initialize the model backbone with generalization weights pre-trained on four unlabeled domains provided in BrAD [12], and then train the model with our C-VisDiT method, denoted as BrAD+C-VisDiT. We can observe that the proposed C-VisDiT can significantly outperform existing methods on all benchmarks. Specifically, compared to the existing state-of-the-art method PCS, C-VisDiT improves the classification accuracy on the Office-31 dataset by 4.4% (1-shot) and 1.7% (3-shots). On the Office-Home dataset, C-VisDiT achieves accuracy gains of 2.6% (3% labeled source) and 2.8% (6% labeled source). On the VisDA-C dataset, C-VisDiT achieves a performance improvement of 1.5%. On the DomainNet dataset, C-VisDiT can achieve 1.2%/2.0% and 1.3%/2.5% accuracy gain (1-shot/3-shots) compared to PCS and BrAD, respec-

Table 5: Performance contribution of each component on the Office-31 dataset in terms of adaptation accuracy (%).

Method	$\mathcal{L}_{\text{X-VD}}$	$\mathcal{L}_{\text{I-VD}}$	1-shot / 3-shots
PCS [50]	-	-	76.6 / 84.0
Baseline	×	×	77.6 / 83.8
C-VisDiT-X	√	×	79.2 / 84.9
C-VisDiT-I	×	√	78.8 / 85.0
C-VisDiT	√	√	**81.0 / 85.7**

Table 6: Adaptation accuracy for C-VisDiT-X with different confidence-based strategies on the Office-31 dataset (%).

Method	Unlabeled source	1-shot / 3-shots
Baseline	-	77.6 / 83.8
MixUp	100%	74.8 / 82.6
	0% (**Ours**)	78.6 / 84.5
C-VisDiT-X	100%	74.8 / 82.4
	10%	76.1 / 83.8
	1%	76.3 / 84.2
	0% (**Ours**)	**79.2 / 84.9**

tively. Looking into each adaptation setting inside both datasets, our method can accomplish a maximum performance improvement of 13.9% (1-shot setting in A→D on the Office-31 dataset). These results show that the proposed C-VisDiT establishes new state-of-the-art performance for FUDA, well demonstrating its effectiveness. More results and details can be found in the supplementary materials.

4.3. Model Analysis

In this section, we present the ablation studies of our proposed C-VisDiT. We also give more investigations of our X-VD and I-VD strategies. For the convenience of analysis, we construct two C-VisDiT variants, *i.e.*, C-VisDiT-X and C-VisDiT-I, trained with X-VD and I-VD only, respectively. More analysis, *e.g.* the hyper-parameter analysis, can be found in the supplementary materials. Following PCS [50], here we employ the Office-31 dataset to inspect our model.

Ablation Studies. Here we investigate the effect of the proposed X-VD and I-VD strategies in C-VisDiT. The results are reported in Tab. 5. We can see that all three C-VisDiT variants can consistently outperform the baseline model by a great margin. Meanwhile, compared with the previous state-of-the-art method, *i.e.*, PCS [50], both C-VisDiT-X and C-VisDiT-I can achieve better results. As a combination of X-VD and I-VD, C-VisDiT significantly exceeds the PCS with a maximum performance improvement of 4.4% (1-shot setting on the Office-31). These results show that our X-VD and I-VD strategies are effective and complementary.

Superiority of Confidence-based Knowledge Transfer in X-VD. The confidence-based strategies play a vital role in the proposed X-VD. First, we demonstrate the effectiveness of our confidence-based strategies by validating on the C-VisDiT-X variant. We directly conduct MixUp as a compared baseline, in which for a given target sample, a random source sample is selected for MixUp. As shown in Tab. 6, MixUp with 100% unlabeled source samples performs significantly worse than the baseline. This shows that plainly conducting MixUp severely harms the performance. However, when using only high-confidence knowledge with 0% unlabeled source, MixUp can easily outperform both the baseline and the 100% unlabeled source variant. This shows that MixUp is greatly restrained due to the massive noise in FUDA setting and eliminating the noise is crucial in the proposed X-VD. Furthermore, replacing MixUp (0% unlabeled source) with our C-VisDiT-X achieves the best result, well demonstrating the effectiveness of our confidence-based strategies. Therefore, the superiority of our X-VD can largely attribute to the confidence-based knowledge transfer, rather than the MixUp transformations.

Table 7: Adaptation accuracy of different target sample guidance strategies for C-VisDiT-I on the Office-31 dataset (%).

Method	$\mathcal{T}_{ut}^E$	$\mathcal{T}_{ut}^H$	1-shot / 3-shots
Baseline	-	-	77.6 / 83.8
MixUp	$\mathcal{D}_{ut}$	$\mathcal{D}_{ut}$	77.3 / 83.9
C-VisDiT-I	$\mathcal{D}_{ut}^E$	$\mathcal{D}_{ut}^H \cup \mathcal{D}_{ut}^O$	77.7 / 84.5
	$\mathcal{D}_{ut}^E$	$\mathcal{D}_{ut}^H$ (**Ours**)	78.2 / 84.4
	$\mathcal{D}_{ls} \cup \mathcal{D}_{ut}^E$ (**Ours**)	$\mathcal{D}_{ut}^H \cup \mathcal{D}_{ut}^O$	78.3 / 84.6
	$\mathcal{D}_{ls} \cup \mathcal{D}_{ut}^E$ (**Ours**)	$\mathcal{D}_{ut}^H$ (**Ours**)	**78.8 / 85.0**

Second, we investigate the impact of source knowledge purity. We compare C-VisDiT-X performance by further involving 100%, 10%, and 1% unlabeled source samples. As shown in Tab. 6, limiting the source samples to only labeled source samples (0% unlabeled source) gives the best performance, while including 100%, 10%, and 1% unlabeled source samples all lead to performance drops. This indicates that the high purity of source knowledge is very helpful for the cross-domain knowledge transfer.

Necessity of Confidence-based Target Sample Guidance in I-VD. To prove the necessity and superiority of our I-VD strategy for model adaptation, we compare a series of sample guidance strategies. First, we compare our proposed I-VD with applying random MixUp on the entire target domain. As shown in Tab. 7, plainly conducting MixUp barely changes target performance, while all our I-VD variants boosted performance. This shows that, although our I-VD is based on MixUp, our confidence-based target sample guidance greatly improves target performance.

Furthermore, we dive into the designs of our confidence-based strategies, and investigate how the choice of $\mathcal{T}_{ut}^E$ and $\mathcal{T}_{ut}^H$ affects target performance. Results show that both adding confident labeled source samples $\mathcal{D}_{ls}$ to $\mathcal{T}_{ut}^E$ and excluding low-confidence target outliers $\mathcal{D}_{ut}^O$ for $\mathcal{T}_{ut}^H$ lead

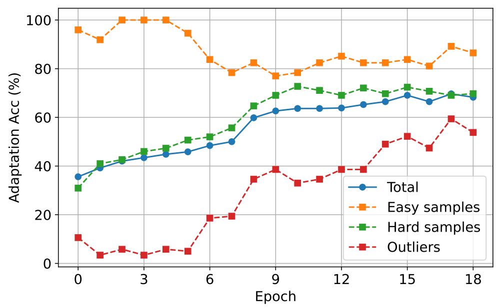

Figure 4: Accuracy change of target samples with different learning difficulties in A→D (1-shot) on the Office-31 dataset.

Table 8: Effect of the confidence-based strategies in C-VisDiT on the Office-31 dataset in terms of adaptation accuracy (%).

Implementation	1-shot / 3-shots
Baseline	77.6 / 83.8
Plain MixUp	75.1 / 83.0
C-VisDiT	**81.0 / 85.7**

to performance gains, and combining the strategies would boost the performance even more, further demonstrating the effectiveness of confidence target sample guidance.

Besides, we qualitatively show the guidance feature of our I-VD strategy towards target samples with different learning difficulties. As illustrated in Fig. 4, I-VD improves adaptation accuracy on hard target samples and outliers dramatically. The slight accuracy loss on easy target samples is due to the model overfitting these samples at the beginning. This proves that I-VD is capable of enhancing the model learning on target samples with consideration of their different learning difficulties.

Superiority of C-VisDiT over MixUp. Motivated by the fact that samples with low prediction confidence commonly exist in FUDA, we propose the confidence-based strategies to comprehensively consider sample confidence during training. To prove the effectiveness of confidence-based strategies more directly, we introduce plain MixUp based on C-VisDiT, in which X-VD is substituted by MixUp on a given target sample and a random source sample, and I-VD is substituted by random MixUp on all target samples. As shown in Tab. 8, plain MixUp underperforms the baseline model, while our C-VisDiT greatly outperforms the baseline model, indicating that the proposed confidence-based strategies are the dominant factor to the superior performance. Therefore, our C-VisDiT is more suitable for FUDA compared to vanilla MixUp.

Visualization results. To qualitatively analyze the effect of the proposed X-VD and I-VD, we visualize the image features extracted by C-VisDiT-X and C-VisDiT-I via t-SNE [40]. As shown in Fig. 5, compared to PCS [50], C-VisDiT-X matches the target samples to the source samples better and C-VisDiT-I produces better clusters. These observations support that our proposed method can construct a better feature space for the model adaptation.

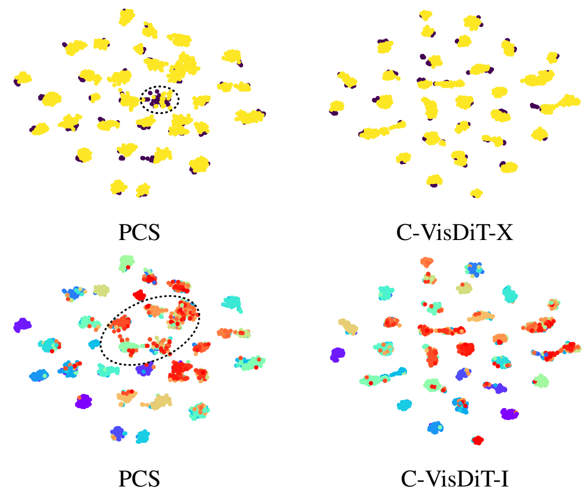

Figure 5: t-SNE [40] visualization results of C-VisDiT-X and C-VisDiT-I compared to PCS [50] in A→D (1-shot) setting on the Office-31 dataset. Top row: Yellow represents target features and **Purple** represents source features. Bottom row: different colors represent different sample classes. Best viewed in color.

5. Conclusion

In this paper, we cope with Few-shot Unsupervised Domain Adaptation (FUDA) in which only a small set of source domain samples are labeled, while all samples in the target domain are unlabeled. We propose a confidence-based novel visual dispersal transfer learning method (C-VisDiT) that transfers knowledge only from high-confidence samples to low-confidence samples via cross-domain visual dispersal (X-VD) and intra-domain visual dispersal (I-VD). The proposed X-VD enhances cross-domain knowledge transfer and mitigates the domain gap, while the proposed I-VD guides the learning of hard target samples and boosts the model adaptation. Extensive experiments conducted on various domain adaptation benchmarks show that our C-VisDiT outperforms existing methods, establishing new state-of-the-art performance for FUDA. In the future, we plan to extend our method to semantic segmentation, object detection, etc.

6. Acknowledgements

This research was supported by National Key R&D Program of China (2021ZD0114703), National Natural Science Foundation of China (Nos. 61925107, U1936202, 62271281, 62021002), Zhejiang Provincial Natural Science Foundation of China (No. LDT23F01013F01), and CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

References

[1] Christian Abbet, Linda Studer, Andreas Fischer, Heather Dawson, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-rule to adapt: Learning generalized features from sparsely-labeled data using unsupervised domain adaptation for colorectal cancer tissue phenotyping. In *Medical Imaging with Deep Learning*, 2021.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017.

[3] Robert Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. *Advances in Neural Information Processing Systems*, 34:10876–10889, 2021.

[4] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4243–4250. IEEE, 2018.

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[7] Ssu-Han Chen and Chia-Chun Tsai. SMD LED chips defect detection using a yolov3-dense model. *Adv. Eng. Informatics*, 47:101255, 2021.

[8] Yuhe Ding, Lijun Sheng, Jian Liang, Aihua Zheng, and Ran He. Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. *arXiv preprint arXiv:2205.14566*, 2022.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[10] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[11] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, 2012.

[12] Sivan Harary, Eli Schwartz, Assaf Arbelle, Peter Staar, Shady Abu-Hussein, Elad Amrani, Roei Herzig, Amit Alfassy, Raja Giryes, Hilde Kuehne, et al. Unsupervised domain generalization by learning a bridge across domains. In *Computer Vision and Pattern Recognition*, pages 5280–5290, 2022.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998, 2018.

[15] Haixin Huang, Xueduo Tang, Feng Wen, and Xin Jin. Small object detection method with shallow feature fusion network for chip surface defect detection. *Scientific reports*, 12(1):1–9, 2022.

[16] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International conference on machine learning*, pages 4816–4827, 2020.

[17] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels. *CoRR*, abs/2003.08264, 2020.

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017.

[19] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726*, 2018.

[20] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[21] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *CoRR*, abs/1506.04579, 2015.

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, pages 9992–10002, 2021.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, 2015.

[25] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source

domain adaptation. In *International Conference on Computer Vision*, pages 1406–1415, 2019.

[28] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[29] Christian S Perone, Pedro Ballester, Rodrigo C Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 2019.

[30] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.

[31] Sayan Rakshit, Hmrishav Bandyopadhyay, Piyush Bharambe, Sai Nandan Desetti, Biplab Banerjee, Subhasis Chaudhuri, et al. Open-set domain adaptation under few source-domain labeled samples. In *Computer Vision and Pattern Recognition*, pages 4029–4038, 2022.

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[33] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226, 2010.

[34] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *International Conference on Computer Vision*, pages 8050–8058, 2019.

[35] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. In *Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[36] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Computer Vision and Pattern Recognition*, pages 2107–2116, 2017.

[37] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[38] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE, 2011.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[41] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[42] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *Computer Vision and Pattern Recognition*, pages 3950–3959, 2020.

[43] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[44] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):51:1–51:46, 2020.

[45] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *European Conference on Computer Vision*, pages 540–555. Springer, 2020.

[46] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6502–6509, 2020.

[47] Hong Xuan, Abby Stylianou, Xiaotong Liu, and Robert Pless. Hard negative examples are hard, but useful. In *European Conference on Computer Vision*, pages 126–142. Springer, 2020.

[48] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.

[49] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *International Conference on Computer Vision*, pages 2849–2857, 2017.

[50] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto L. Sangiovanni-Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*, pages 13834–13844, 2021.

[51] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

[52] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[53] Yang Zhang, Liru Qiu, Yongkai Zhu, Long Wen, and Xiaoping Luo. A new childhood pneumonia diagnosis method based on fine-grained convolutional neural network. *Computer Modeling in Engineering & Sciences*, 133:873–894, 01 2022.

[54] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):473–493, 2022.

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017.

[56] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. Deep unsupervised convolutional domain adaptation. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 261–269, 2017.