Shift-Driven Learning for Unsupervised Domain Adaptation

Wentang Chen[1], Yibin Wen[2], Juepeng Zheng[2] (✉)
[1]College of Computer Science and Electronic Engineering, Hunan University
[2]School of Artificial Intelligence, Sun Yat-Sen University
B241000657@hnu.edu.cn, zhengjp8@mail.sysu.edu.cn, wenyb5@mail2.sysu.edu.cn

2025 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3315-9495-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICME59968.2025.11210226

***Abstract*—Self-training is widely used in unsupervised domain adaptation (UDA) by assigning pseudo labels to unlabeled samples. However, existing self-training strategies bring bias, while potentially inaccurate pseudo labels may accumulate errors during self-training (self-training shift) and the inability to accurately distinguish features may bring prediction bias (class shift). To address these issues, we propose Shift-Driven Learning (SDL). First, we decouple the generation and utilization of pseudo labels to mitigate the direct error accumulation. Second, we measure the maximum training shift of data, where the classifier achieves high accuracy on labeled data while making as many mistakes as possible on unlabeled data. Then we adversarially optimize the feature representations generation to indirectly decrease the self-training shift. Third, we minimize the class shift by data rearrangement strategy and joint contrastive learning, which find class-level discriminative feature representations. Extensive experiments justify that SDL outperforms SOTA methods on three UDA datasets with considerable gains.**

***Index Terms*—Unsupervised Domain Adaptation, Self-training, Contrastive Learning**

I. Introduction

The field of deep learning has seen considerable growth in various fields. In order to train state-of-the-art (SOTA) neural networks, large-scale well-annotated datasets are necessary. However, the collection and annotation progress can be labor-exhaustive and time-consuming. Therefore, it would be beneficial to use simulated or existing datasets, which are easier to annotate. However, a network trained on such a source dateset performs worse while applied to actual target dataset, as the data is not independent and identically distributed. To address this problem, unsupervised domain adaptation (UDA) methods [1]–[5], which adapt the network from labeled source domain to unlabeled target domain, are deeply explored.

UDA methods either reduce domain discrepancy via high-order moment matching [6] or adversarial training [7], or they employ round-based self-training using pseudo labels generated from model predictions on unlabeled data. Although self-training methods have achieved promising performance in UDA, the shifts in the iterative training process have not been well addressed. As shown in Fig. 1, the shifts come from two aspects: self-training shift and class shift. Self-training shift occurs when a model that generates pseudo labels is trained with the pseudo labels themselves. Thus the error learned from pseudo labels in the previous training round accumulates in following rounds, where the classification head learns accumulated bias. As shown in the top-right of Fig. 1, class shift

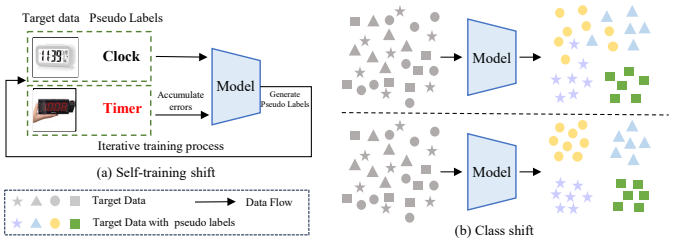

Fig. 1. The shifts derived from two aspects: (a) self-training shift. (b) class shift. The pseudo label in black is correct while the pseudo label in red is wrong. The self-training shift is referred as the errors accumulated when generating and utilizing inaccurate pseudo labels in the iterative training process on the target data. The class shift means that the model fails to guarantee the learning of class-wise discriminative features.

means that the model fails to distinguish class-wise discriminative features, which brings prediction bias while training the model. Recent domain adaptation (DA) methods primarily employed first-order statistics of conditional distributions [8] or pairwise loss [9] to address imbalanced class distributions. However, these methods face two challenges. Firstly, although they attempt to align the class-conditional distributions by minimizing the distances between the source and target class centers, this approach only achieves coarse alignment and does not effectively generate discriminative features. Secondly, pairwise loss or triplet loss fails to accurately estimate and maximize the mutual information (MI) between the learned representation and its corresponding label [10].

To address the above problems, we propose Shift-Driven Learning (SDL). Specifically, to reduce the self-training shift, we first decouple the generation and utilization of pseudo labels. Then the classifier head is only trained with the corrected-labeled samples to avoid negative impact from unreliable pseudo-labeled samples. Furthermore, we found that the data distribution shift drives the pseudo-labeling function to generate incorrect labels, which eventually leads to self-training shift. Since the data distribution shift cannot be measured directly, we turn to estimate the maximum training shift which is highly relevant with data distribution shift. Then we optimize the feature representations to adversarially decrease the maximum shift, thereby reducing the the data distribution shift and self-training shift. To reduce the class shift, we introduce a data rearrangement strategy, which combines the data from source domain and target domain and jointly exploit the mutual information between a feature and its label to reduce the joint error. The contributions of our work can be

summarized as:

1) We propose SDL to mitigate the self-training shift and class shift in UDA. We innovatively mitigate the self-training shift by decoupling the generation and utilization of pseudo labels and adversarially optimizing the feature representations.
2) By introducing a data rearrangement strategy, we maximize the MI between a feature and its label to enhance the feature discriminability and reduce class shift.
3) We conduct extensive experiments on three common domain adaptation benchmarks, indicating that SDL outperforms SOTA methods with considerable gains.

Due to limit space, **Related Work** is included in **supplementary material**.

II. Shift-Driven Learning (SDL)

A. Problem Statement

This paper focuses on the classification task under the setting of UDA. UDA constitutes a labeled source distribution $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$and an unlabeled target domain distribution $\mathcal{D}_T = \{x_i^t\}_{i=1}^{n_t}$, where x_i^s and x_i^t denote the samples from source domain and target domain, and y_i^s denotes the true labels. n_s and n_t denote the number of instances from the source domain and the target domain respectively. A domain is defined as a pair comprised of a distribution $\mathcal{D}$ in the input space $\mathcal{X}$ and its labeling function $f : \mathcal{X} \to \mathcal{Y}$, where the output space $\mathcal{Y}$ is constrained to be [0, 1] theoretically. We denote the source domain and target domain as $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$, respectively. The goal of unsupervised domain adaptation is to learn a function $h : \mathcal{X} \mapsto \mathcal{Y}$ that provides good generalization on the target domain.

B. Reduce self-training shift

1) Independent generation and utilization for pseudo labels : The self-training shift of Pseudo Label [11] relies on pseudo labels generated by the same model head. To enhance label reliability, some methods incorporate teacher models. As illustrated in Fig. 2(a) and (b), Noisy Student uses a previously trained model [12], while Mean Teacher employs an exponential moving average of the model for label generation. Despite these strategies, both teacher and student models remain closely tied, causing students to learn from inaccurate pseudo labels. As shown in Fig. 2(c), PLC [13] utilizes separate models for label generation and usage, but error accumulation from inaccurate labels remains.

To further mitigate the adverse effects of inaccurate pseudo labels, we optimize the classification head solely using the correct labels from the source domain $\mathcal{D}_S$, without utilizing any pseudo labels from the target domain $\mathcal{D}_T$. Moreover, to prevent overfitting of the model on the limited labeled samples from the source domain, we incorporate an additional head that utilizes pseudo labels from the target domain only for learning better feature representations. As shown in Fig. 2(d), we introduce an independent head h_{inde} connected to the feature generator G_F utilizes pseudo labels independently on the target domain. Therefore the objective is formulated as:

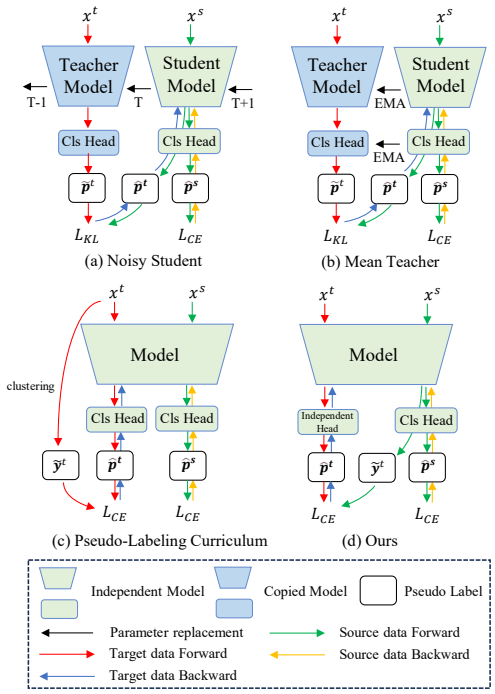

Fig. 2. Comparison of different models on how to generate and use pseudo labels. (a) Noisy Student generates pseudo labels from the teacher model learned from previous round of training. (b) Mean Teacher generate pseudo labels from the Exponential Moving Average of current model. (c) Pseudo-Labeling Curriculum generates pseudo labels from clustering. (d) Ours generates pseudo labels from head h_c and utilizes pseudo labels on independent head h_{inde}. $\mathbf{p}^s$ and $\mathbf{p}^t$ denotes predictions from classification (Cls) head on the source samples and target samples respectively; y^t denotes pseudo labels on the target samples;ˆdenotes labels or predictions from teacher model, while ˜ denotes that from student model or single model. L_{KL} and L_{CE} denote Kullback-Leibler (KL) divergence loss and Cross-Entropy loss.

$$\min_{G_F, h_c, h_{\text{inde}}} \mathcal{L}_{\langle \mathcal{D}_S, f_S \rangle}(G_F, h_c) + \lambda \mathcal{L}_{\langle \mathcal{D}_T, f_T \rangle}(G_F, h_{\text{inde}}) \quad (1)$$

Where $\mathcal{L}$ denotes the standard cross-entropy loss function and h_c is the classification head. λ is the trade-off between the loss on source domain and that on the target domain. h_c generates pseudo labels and h_{inde} utilizes them independently. Although they are connected to the same feature extractor, their parameters are independent. Therefore, the errors from inaccurate pseudo labels will not affect or be accumulated on h_c. h_{inde} is to propagate gradients during training to the feature generator. During inference, we only use h_c to ensure that h_{inde} does not impact the inference process.

2) Reducing wrong labels by the worst head: Section II-B1 provides a resolution of self-training shift for pseudo-labeling. However, data distribution shift also affects the pseudo-labeling function f and then generates self-training shift. As illustrated in Fig. **??**(a), due to data distribution shift, samples from different classes have different distances from the feature decision hyperplane, leading to bias between the learned hyperplane and the real decision hyperplane. Consequently, the pseudo-labeling function f is prone to generating in accurate pseudo labels on samples that are close to the biased learned hyperplanes. Therefore, our objective is to optimize the spatial

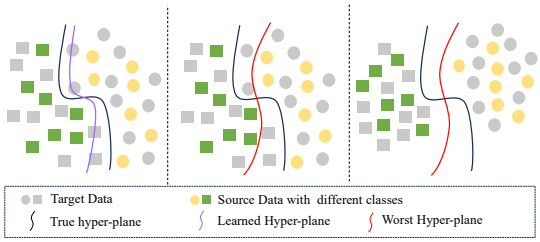

Fig. 3. The worst head explanation. (a) The learned hyperplanes exhibits a shift from the true hyperplanes. (b) The worst hyperplanes are as accurate as possible on the labeled source domain data, while making as many mistakes as possible on the unlabeled target domain. (c) The feature representations are optimized to improve the performance of the worst hyperplane.

distribution of feature representation to reduce data distribution shift and then indirectly reduce self-training shift.

Since no labels are available in the target domain, we are unable to directly reduce data distribution shift. As discussed before, inaccurate pseudo labels can impact optimization process, bringing biased learned hyperplane and self-training shift. Consequently, self-training shift can be viewed as the accumulation of data distribution shift with inappropriate utilization of inaccurate pseudo labels, which is training algorithm dependent. Therefore, the maximum self-training shift can serve as a good indicator of data distribution shift. The maximum self-training shift is from worst-case head learned from pseudo-labeling function f, which predicts as accurately as possible on labeled source domain data, while making as many mistakes as possible on unlabeled target domain data.

$$h_{\text{worst}}\left(G_F\right) = \arg\max_{h'} \mathcal{L}_{\langle \mathcal{D}_T, f_t \rangle}\left(G_F, h', f_{G_F,h_c}\right) - \mathcal{L}_{\langle \mathcal{D}_S, f_s \rangle}\left(G_F, h'\right) \tag{2}$$

where the mistakes of the worst-case head h' on unlabeled data are measured by its divergence from the current pseudo labeling function f. Eq. 2 is designed to identify the maximum self-training shift for the classification head h_c that could be learned in future iterations when trained using pseudo labeling on the current feature generator G_F and data. This scenario corresponds to the worst hyperplanes, as shown in Fig. **??**(b), which deviate from the currently learned hyperplanes while ensuring accurate classification of all labeled samples. Eq. 2 measures the degree of self-training shift, which depends on the feature representations generated by G_F. Consequently, we can adversarially optimize it to indirectly minimize the self-training shift by optimizing the feature generator G_F.

$$\min_{G_F} \mathcal{L}_{\langle \mathcal{D}_T, f_t \rangle}\left(G_F, h_{\text{worst}}\left(G_F\right), f_{G_F,h_c}\right) - \mathcal{L}_{\langle \mathcal{D}_S, f_s \rangle}\left(G_F, h_{\text{worst}}\left(G_F\right)\right) \tag{3}$$

As shown in Fig. 3(c), Eq. (3) promotes accurate discrimination of features belonging to unlabeled samples, even with the worst hyperplanes. Specifically, it encourages the generation of features that are significantly distant from current hyperplanes, thereby reducing data distribution shift and self-training shift.

C. Reduce class shift

1) Data rearrangement strategy: To reduce the marginal label distribution discrepancy and a small class shift [14],

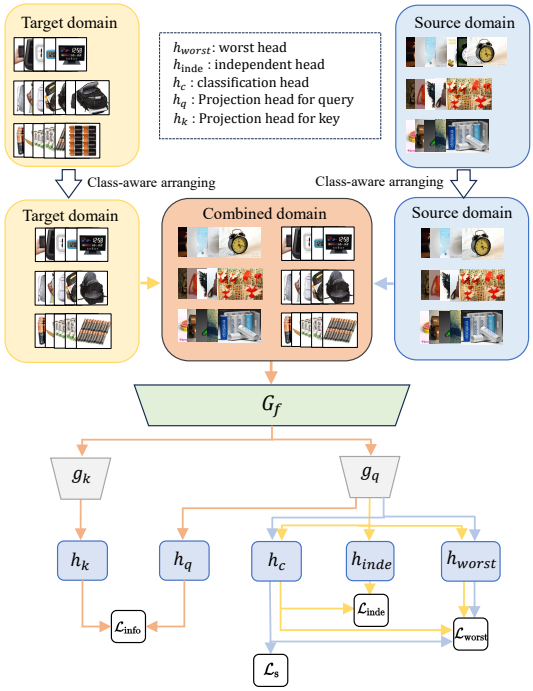

Fig. 4. An overview of SDL. We extract features from source domain, target domain and combined domain. For joint contrastive learning, we adopt Momentum Contrast [15], which uses an on-the-fly maintained queue and a moving-averaged encoder to establish a substantial and consistent dictionary. After the encoders, a worst head h_{worst} is utilized for reducing self-training shift indirectly, an independent head h_{inde} is adopted for utilizing pseudo for learning better feature representations, a fully connected head h_c is utilized for classification, while h_q, h_k is employed for contrastive loss.

we propose the reformation of datasets through a data rearrangement strategy, which equalizes the number of data instances per class as illustrated in Fig. 4. We first need to access the labels of the target data to determine whether two samples in the combined domain have the same label. So we generates pseudo labels for the target data by spherical K-means clustering in the feature space $\mathcal{Z}$ and assign labels to them. After clustering at each beginning of each epoch, each target sample is assigned as its affiliated clusters. A target sample, which is far away from its affiliated clustering center than distance d, is excluded from the combined dataset.

2) Contrastive Learning: In the field of unsupervised learning, previous works [16], [17] have adopted discriminative feature learning to reduce the class shift in various task. These methods focus on minimizing a $\mathcal{H}\Delta\mathcal{H}$ distance between source domain and target domain in feature space $\mathcal{Z}$, where $\mathcal{H}\Delta\mathcal{H}$ [10] measures the dissimilarity between source and target domains in feature space. However, as demonstrated in [18], the ideal joint hypothesis error λ in the theoretical basis of domain adaptation [19] which was previously considered to be minor, has been proved to be substantial as feature transferability is optimized. As a result, the severe joint hypothesis error leads to extreme class shift. Our objective is to minimize the joint hypothesis error through jointly learning discriminative feature representation and reduce class shift.

Inspired by the idea in [10], we aim to maximize the mutual information between a feature and its label through maximizing the Jensen-Shannon divergence D_{JS}. According

TABLE I
ACCURACY(%) ON **VISDA-2017** DATASET FOR UNSUPERVISED DOMAIN ADAPTATION (RESNET-101).

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
ResNet-101 [22] (CVPR'16)	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [23] (JMLR'16)	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD [7] (CVPR'18)	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
MSTN+DSBN [24] (CVPR'19)	94.7	86.7	76.0	72.0	95.2	75.1	87.9	81.3	91.1	68.9	88.3	45.5	80.2
BNM [25] (CVPR'20)	89.6	61.5	76.9	55.0	89.3	69.1	81.3	65.5	90.0	47.3	89.1	30.1	70.4
CGDM [26] (CVPR'21)	93.4	82.7	73.2	68.4	92.9	94.5	88.7	**82.1**	93.4	82.5	86.8	49.2	82.3
SDAT [27] (ICML'22)	94.8	77.1	82.8	60.9	92.3	95.2	**91.7**	79.9	89.9	91.2	88.5	41.2	82.1
MIC [28] (CVPR'24)	96.7	88.5	**84.2**	**74.3**	**96.0**	**96.3**	90.2	81.2	94.3	95.4	88.9	56.6	86.9
SDL (Ours)	**97.5**	**92.2**	83.6	65.0	95.4	95.7	87.9	81.4	**96.7**	**96.4**	**89.7**	**71.1**	**87.7**

TABLE II
ACCURACY (%) ON **IMAGECLEF-DA** FOR UNSUPERVISED DOMAIN ADAPTATION (RESNET-50).

Method	I → P	P → I	I → C	C → I	C → P	P → C	Average
ResNet-50 [22] (CVPR'16)	74.8 ± 0.3	83.9 ± 0.1	91.5 ± 0.3	78.0 ± 0.2	65.5 ± 0.3	91.2 ± 0.3	80.7
DANN [23] (JMLR'16)	75.0 ± 0.6	86.0 ± 0.3	96.2 ± 0.4	87.0 ± 0.5	74.3 ± 0.5	91.5 ± 0.6	85.0
MADA [29] (AAAI'18)	75.0 ± 0.3	87.9 ± 0.2	96.0 ± 0.3	88.8 ± 0.3	75.2 ± 0.2	92.2 ± 0.3	85.8
CAT [30] (ICCV'19)	77.2 ± 0.2	91.0 ± 0.3	95.5 ± 0.3	91.3 ± 0.3	75.3 ± 0.6	93.6 ± 0.5	87.3
CDAN [31] (CVPR'21)	77.7 ± 0.3	90.7 ± 0.2	**97.7** ± 0.3	91.3 ± 0.3	74.2 ± 0.2	94.3 ± 0.3	87.7
SDAT [27] (ICML'22)	78.0 ± 0.3	92.5 ± 0.2	96.0 ± 0.2	90.7 ± 0.3	77.0 ± 0.2	93.8 ± 0.2	88.0
MIC [28] (CVPR'24)	78.1 ± 0.2	93.2 ± 0.1	96.8 ± 0.2	91.7 ± 0.1	**77.3** ± 0.4	94.7 ± 0.1	88.6
SDL (Ours)	**78.2** ± 0.1	**93.5** ± 0.2	97.5 ± 0.3	**92.7** ± 0.5	76.5 ± 0.2	**98.3** ± 0.2	**89.5**

to [10], let $\mathcal{D}_U^{\mathcal{Z}}$ be the combined domain of source domain and target domain $\mathcal{D}_U$ over the representation space $\mathcal{Z}$, and its class-conditional distribution $\mathcal{D}_{U|y}^{\mathcal{Z}}$, where y is the class label. We formalize our objective to maximize the JS divergence D_{JS} between different class-conditional distributions as:

$$\max_{\theta_g} D_{\mathrm{JS}}\left(\mathcal{D}_{U|0}^{\mathcal{Z}} \| \mathcal{D}_{U|1}^{\mathcal{Z}}\right) \tag{4}$$

where θ_{G_F} denotes the parameters of the feature generator G_F. The value 0 and 1 are the class labels. Since the label distribution is not uniform, we can reformulate a dataset to be class-wise uniform. Consider Y as a uniform random variable, which takes value in $\{0,\ 1\}$. Based on Y, the distribution $\mathcal{D}_{U|Y}^{\mathcal{Z}}$ is the mixture of $\mathcal{D}_{U|0}^{\mathcal{Z}}$ and $\mathcal{D}_{U|1}^{\mathcal{Z}}$. We denote the feature random variable with the distribution $\mathcal{D}_{U|Y}^{\mathcal{Z}}$ as $Z_{U|Y}$. Consequently, the relation between JS divergence D_{JS} and mutual information (MI) can be formulated as: $D_{\mathrm{JS}}\left(\mathcal{D}_{U|0}^{\mathcal{Z}} \| \mathcal{D}_{U|1}^{\mathcal{Z}}\right) = I\left(Y; Z_{U|Y}\right)$, where I denotes mutual information between the label distribution and feature distribution. Thus the objective can be transformed as:

$$\max_{\theta_g} I\left(Y; Z_{U|Y}\right) \tag{5}$$

In application, we maximize the mutual information between feature and its label indirectly. We adopt a strategy where features extracted from the same conditional distribution are paired together, and we adopt InfoNCE loss [16] to maximize the mutual information between them:

$$I(X;Y) \geq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} \log \frac{e^{c(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K} e^{c(x_i,y_j)}}\right] \triangleq I_{\mathrm{NCE}}(X;Y) \tag{6}$$

where $\mathbb{E}$ denotes the expectation of K samples from the joint distribution [20]. $c(x, y)$ denotes the critic function that predicts whether x and y originate from the same jointly drawn pair by setting a high threshold [21]. Consider a given variable Y. Let $Z_{U|Y}^{(1)}$ and $Z_{U|Y}^{(2)}$ represent two distinct features sampled from the same conditional distribution $\mathcal{D}_{U|Y}^{\mathcal{Z}}$. Then, $I\left(Z_{U|Y}^{(1)}; Z_{U|Y}^{(2)}\right) \leq I\left(Y; Z_{U|Y}^{(1)}, Z_{U|Y}^{(2)}\right)$holds due to data processing inequality [10] Consequently, $\max_{\theta_{G_F}} I\left(Z_{U|Y}^{(1)}; Z_{U|Y}^{(2)}\right)$ serves as a lower bound for objective $\max_{\theta_{G_F}} I\left(Y; Z_{U|Y}\right)$ which will be optimized the following InfoNCE loss $\mathcal{L}_{\mathrm{info}}$.

D. Optimization

The overview of our framework is illustrated in Fig. 4. We adopt Momentum Contrast [15] for our joint contrastive learning structure. It consists of an encoder g_q with parameter θ_q and a momentum-updated encoder g_k with parameter θ_k, and both of them are to transfer features from space $\mathcal{X} \mapsto \mathcal{Z}$. The parameter θ_k are updated by $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$, where $m \in [0, 1)$ is momentum coefficient. Test results [17] suggest that it is advantageous to define the InfoNCE loss in the projected space $\mathcal{W}$ rather than $\mathcal{Z}$. Thus, to map the encoded feature representations to a space $\mathcal{W}$ where the InfoNCE loss is applied, we use two fully connected layer projection head h_q and h_k for projection: $\mathcal{Z} \mapsto \mathcal{W}$. Note that h_q and h_k are defined in a similar manner to encoder g_q and g_k, respectively. For Eq.6, the feature pairs consist of an encoded query $w_q = h_q\left(g_q(x)\right)$ and a encoded key $w_k = h_k\left(g_k(x)\right)$. We use cosine similarity function $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$ as the critic function c . τ are the temperature hyper-parameter [33]. Then the InfoNCE loss is formulated as:

$$\mathcal{L}_{\text{info}} = \mathbb{E}_{w_q \sim \mathcal{D}_U^W} \left[\mathbb{E}_{w_k^+}\left[-\log \frac{\exp\left(\mathrm{sim}\left(w_q, w_k^+\right)/\tau\right)}{\sum_{w_k \in N_k \cup \left\{w_k^+\right\}} \exp\left(\mathrm{sim}\left(w_q, w_k\right)/\tau\right)}\right]\right] \tag{7}$$

where w_q and w_k^+ are the features that have the same label, and N_k denotes a set of features whose labels is different from w_q. For the classification head h_c, we simply adopt another fully connected layer. To minimize the classification loss on the source domain, we employ the cross-entropy loss:

$$\mathcal{L}_{\text{s}} = \mathbb{E}_{(x^s, y^s) \sim \mathcal{D}_S}\left[-\log h_c\left(g_q\left(G_F(x^s)\right)\right)_{y^s}\right] \tag{8}$$

For both the independent head h_{inde} and worst-case head h_{worst} , the loss for self-training shift can be fomulated as:

$$\mathcal{L}_{\text{inde}} = \mathbb{E}_{(x^t, y^t) \sim \mathcal{D}_T}\left[-\log h_{\text{inde}}\left(g_q\left(G_F(x^t)\right)\right)_{y^t}\right] \tag{9}$$

where the pseudo labels generated by h_c from target domain are regarded as true labels, as described in SectionII-B1. And the loss for worst-case head h_{worst} is:

$$\begin{aligned}\mathcal{L}_{\text{worst}} = &\mathbb{E}_{(x^s, y^s) \sim \mathcal{D}_S}\left[-\log h_{\text{worst}}\left(g_q\left(G_F\left(x^s\right)\right)\right)_{y^s}\right] \\ &- \mathbb{E}_{(x^t, y^t) \sim \mathcal{D}_T}\left[-\log h_{\text{worst}}\left(g_q\left(G_F(x^t)\right)\right)_{y^t}\right]\end{aligned} \tag{10}$$

where the predictions generated by h_c from source domain are regarded as true labels for $\mathbb{E}_{(\boldsymbol{x}_s, y_s) \sim \mathcal{D}_S}$, and pseudo labels generated by h_c from target domain are regarded as true labels for $\mathbb{E}_{(\boldsymbol{x}_t, y_t) \sim \mathcal{D}_T}$.

1) Overall loss: The final objective of the Shift-Driven Learning is to reduce self-training shift and class shift. Combining the $\mathcal{L}_{\text{s}}$, $\mathcal{L}_{\text{info}}$, $\mathcal{L}_{\text{inde}}$ and $\mathcal{L}_{\text{worst}}$ with hyper-parameter α, γ and β , the overall objective is then formulated as follows:

$$\min_{\theta, G_F, h_c, h_{\text{inde}}} \max_{h_{\text{worst}}} \mathcal{L}_{\text{s}} + \alpha\mathcal{L}_{\text{info}} + \gamma\mathcal{L}_{\text{inde}} + \beta\mathcal{L}_{\text{worst}} \tag{11}$$

TABLE III
ACCURACY (%) ON **OFFICE-HOME** DATASET FOR UNSUPERVISED DOMAIN ADAPTATION (RESNET-50).

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
ResNet-50 [22] (CVPR'16)	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
JAN [6] (ICML'17)	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
DSR [32] (IJCAI'19)	53.4	71.6	77.4	57.1	66.8	69.3	56.7	49.2	75.7	68.0	54.0	79.5	64.9
CDAN [31] (CVPR'21)	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
SDAT [27] (ICML'22)	**55.5**	72.4	78.0	61.9	71.1	72.1	61.9	**55.3**	80.9	**75.1**	**61.9**	83.7	69.1
MIC [28] (CVPR'24)	51.3	73.8	76.5	61.0	69.4	74.2	**68.0**	53.9	82.3	74.2	61.4	**84.7**	69.2
SDL (Ours)	55.4	**76.4**	**80.0**	**65.6**	**73.9**	**75.9**	65.1	52.5	**80.7**	69.2	57.4	84.1	**69.7**

TABLE IV
ABLATION STUDY ON **OFFICEHOME**, **VISDA-2017** AND **IMAGECLEF-DA** (P→C) WITH DIFFERENT LOSS FUNCTIONS.

$\mathcal{L}_{\text{info}}$	$\mathcal{L}_{\text{inde}}$	$\mathcal{L}_{\text{worst}}$	OfficeHome (Pr→Cl)	Visda-2017	ImageCLEF-DA (P→C)
			40.9	68.9	96.6
✓			50.3	86.9	98.2
	✓		40.7	40.0	96.8
		✓	40.0	67.9	95.8
✓	✓		51.2	75.0	96.0
✓		✓	51.1	86.4	97.5
	✓	✓	40.8	76.8	97.7
✓	✓	✓	**52.5**	**87.7**	**98.5**

III. EXPERIMENTS

A. Setup

Implementation Details. We use **VisDA-2017**, **Office-Home**, and **Image-CLEF** datasets for evaluation. We use mini-batch SGD with momentum of 0.9 and the learning rate η_p is adjusted as $\eta_p = \eta_0(1+\alpha p)^{-\beta}$, where p denotes the training progress ranging from 0 to 1. The initial learning rate η_0 is set to 0.001 for the pre-trained layers and 0.01 for the added FC layers. The values of α and β remain fixed at 10 and 0.75, respectively. Due to limit space, more details are in **supplementary material**.

B. Results

Results on VisDa-2017 are shown in Table I. SDL outperforms the other methods even if there is large domain gap between the synthetic and real images. In general, our method performs best in : plane, bicycle, plant, skateboard, train and truck. In the more challenging categories such as bicycle, skateboard, and truck, our proposed method exhibits superior performance compared to existing methods, achieving accuracy rates of 92.2%, 96.4%, and 71.1%, respectively. In particular, SDL boosts the accuracy of the truck class by a significant margin (14.5%) comparing to MIC. These results can be attributed to the maximization of MI between the feature distribution and label distribution, which is achieved by minimizing the InfoNCE loss.

Results on ImageCLEF-DA are shown in Table II. SDL outperforms all the SOTA approaches and achieves the highest accuracy (89.5%). In particular, our proposed method boosts the accuracy of the transfer task P → C by a considerable margin (3.6%) compared to MIC, this is because MIC accumulates error during training by directly generating and utilizing the pseudo labels while SDL does not. Furthermore, the methods incorporating conditional distribution considerations achieve higher accuracies compared to those focusing on matching marginal distributions. The results suggests the significance of jointly learning discriminative features to minimize class shift.

Results on Office-Home are shown in Table III. SDL demonstrates superior performance compared to all SOTA methods with the highest accuracy 69.7%. In particular, SDL outperforms MIC on transfer task on Cl → Ar and Cl → Pr by huge margins, validating the effectiveness of SDL to reduce self-training shift and class shift. Generally, SDL achieves the highest accuracy across 6 adaptation tasks, which demonstrates the exceptional generalization capability.

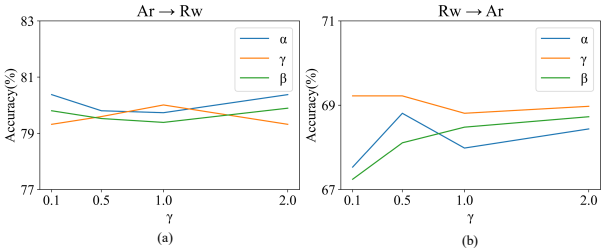

Fig. 5. The accuracy sensitivity of SDL to α, γ, β on **OfficeHome** for adaptation scenario: Ar→Rw (a) and Rw→Ar (b).

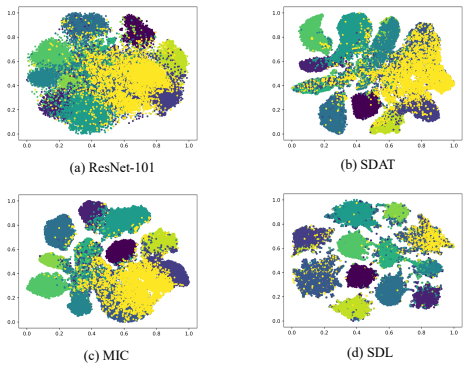

Fig. 6. The t-SNE visualization of ResNet-101, SDAT, MIC and SDL.

C. Analysis

Ablation study. We examine SDL on **OfficeHome** (Pr→Cl), **Visda-2017** and **ImageCLEF-DA** (P→C) in Table IV and have the following findings: (1) The loss $\mathcal{L}_{\text{info}}$ improves the performance by large margins on **OfficeHome** and **Visda-2017** dataset. This is attributed to the combined effect of the data rearrangement strategy and contrastive learning, which is highly effective in achieving our objective to enhance feature discriminability. But $\mathcal{L}_{\text{info}}$ demonstrates only marginal improvement on **ImageCLEF-DA** (P→C) dataset. This may be attributed to the disparate distributions of the two datasets, wherein marginal distributions by data rearrangement strategy inconsistently affect contrastive learning. (2) The accuracy with independent head h_{inde} and worst head h_{worst} is higher compared to the scenarios where they are not adopted. This can be attributed to the reduction of self-training shift. (3) The joint optimization of these losses enables the model to achieve superior performance in adapting to the target domain.

Sensitive Analysis. We investigate the sensitivity of SDL to the weight hyper-parameter α, γ, β ranging from 0.1 to 2.0 on **OfficeHome** for adaptation scenario: Ar→Rw and Rw→Ar, and the results are shown in Fig. 5. We could observe that SDL is not sensitive to the change in the value of α, γ and β.

Feature Visualization. We visualize the t-SNE [34] embeddings of learned target representations of **VisDa-2017** dataset by ResNet101, SDAT, MIC and SDL in Fig. 6. It is evident that though marginal distributions of source and target domains are aligned, the features are not well discriminated by SDAT and MIC. On the opposite, the target features are well discriminated by SDL, which can be attributed to two aspects: (1) The way that reduces self-training shift makes the pseudo labels more reliable, thus providing more accurate labeled target data and eventually enhancing the model's discriminating ability; (2) The data rearrangement strategy and the objective to maximize the JS divergence between different class-conditional distributions are achieved in enhancing feature discriminability.

Applicability of the data rearrangement strategy.The data rearrangement strategy aims to construct uniform label distributions. It generates pseudo labels by spherical K-means clustering for target domain samples and uniformly combines domains with source data and target data. Thus, the data rearrangement strategy is regardless of the network, and it is applicable to other methods.

IV. Conclusion

We introduce Shift-Driven learning (SDL), a novel approach to minimize both self-training shift and class shift. For self-training shift, we decouple the generation and utilization of pseudo labels and reduce the self-training shift adversarially to optimize the feature representations. For class-shift, we introduce a data rearrangement strategy to equalize the number of data instances per class and jointly learn more discriminative features by contrastive learning. Experiments demonstrate the superiority of SDL compared with the strong baselines.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (42401415) and Jiangsu Innovation Capacity Building Program (Project BM2022028).

References

[1] Wentang Chen, Yibin Wen, Juepeng Zheng, Jianxi Huang, and Haohuan Fu, "Ban: A universal paradigm for cross-scene classification under noisy annotations from rgb and hyperspectral remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[2] Juepeng Zheng, Yibin Wen, Mengxuan Chen, Shuai Yuan, Weijia Li, Yi Zhao, Wenzhao Wu, Lixian Zhang, Runmin Dong, and Haohuan Fu, "Open-set domain adaptation for scene classification using multi-adversarial learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 245–260, 2024.

[3] Juepeng Zheng, Wenzhao Wu, Shuai Yuan, Yi Zhao, Weijia Li, Lixian Zhang, Runmin Dong, and Haohuan Fu, "A two-stage adaptation network (tsan) for remote sensing scene classification in single-source-mixed-multiple-target domain adaptation (s^2m^2t da) scenarios," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[4] Qingmei Li, Yibin Wen, Juepeng Zheng, Yuxiang Zhang, and Haohuan Fu, "Hyunida: Breaking label set constraints for universal domain adaptation in cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[5] Juepeng Zheng, Yi Zhao, Wenzhao Wu, Mengxuan Chen, Weijia Li, and Haohuan Fu, "Partial domain adaptation for scene classification from remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2022.

[6] Mingsheng Long, Han Zhu, et al., "Deep transfer learning with joint adaptation networks," in *ICML*. PMLR, 2017, pp. 2208–2217.

[7] Kuniaki Saito, Kohei Watanabe, et al., "Maximum classifier discrepancy for unsupervised domain adaptation," in *CVPR*, 2018, pp. 3723–3732.

[8] Guoliang Kang, Lu Jiang, Yi Yang, et al., "Contrastive adaptation network for unsupervised domain adaptation," in *CVPR*, June 2019.

[9] Yucen Luo, Jun Zhu, Mengxi Li, et al., "Smooth neighbors on teacher graphs for semi-supervised learning," in *CVPR*, 2018, pp. 8896–8905.

[10] Changhwa Park et al., "Joint contrastive learning for unsupervised domain adaptation," *arXiv preprint arXiv:2006.10297*, 2020.

[11] Dong-Hyun Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop*. Atlanta, 2013, vol. 3, p. 896.

[12] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, vol. 30, 2017.

[13] Jaehoon Choi et al., "Pseudo-labeling curriculum for unsupervised domain adaptation," *arXiv preprint arXiv:1908.00262*, 2019.

[14] Han Zhao et al., "On learning invariant representations for domain adaptation," in *ICML*. PMLR, 2019, pp. 7523–7532.

[15] Kaiming He, Haoqi Fan, et al., "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[16] Aaron van den Oord et al., "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[17] Ting Chen et al., "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.

[18] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang, "Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation," in *ICML*. PMLR, 2019, pp. 1081–1090.

[19] Shai Ben-David, John Blitzer, et al., "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.

[20] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, et al., "On variational bounds of mutual information," in *ICML*. PMLR, 2019, pp. 5171–5180.

[21] Michael Tschannen et al., "On mutual information maximization for representation learning," *arXiv preprint arXiv:1907.13625*, 2019.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[23] Yaroslav Ganin, Evgeniya Ustinova, et al., "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.

[24] Woong-Gi Chang et al., "Domain-specific batch normalization for unsupervised domain adaptation," in *CVPR*, 2019, pp. 7354–7362.

[25] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, et al., "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *CVPR*, 2020, pp. 3941–3950.

[26] Zhekai Du et al., "Cross-domain gradient discrepancy minimization for unsupervised domain adaptation," in *CVPR*, 2021, pp. 3937–3946.

[27] Harsh Rangwani et al., "A closer look at smoothness in domain adversarial training," in *ICML*. PMLR, 2022, pp. 18378–18399.

[28] Lukas Hoyer et al., "Mic: Masked image consistency for context-enhanced domain adaptation," in *CVPR*, 2023, pp. 11721–11732.

[29] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang, "Multi-adversarial domain adaptation," in *AAAI*, 2018, vol. 32.

[30] Zhijie Deng, Yucen Luo, et al., "Cluster alignment with a teacher for unsupervised domain adaptation," in *ICCV*, 2019, pp. 9944–9953.

[31] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, "Conditional adversarial domain adaptation," *NeurIPS*, vol. 31, 2018.

[32] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao, "Learning disentangled semantic representation for domain adaptation," in *IJCAI*. NIH Public Access, 2019, vol. 2019, p. 2060.

[33] Zhirong Wu et al., "Unsupervised feature learning via non-parametric instance discrimination," in *CVPR*, 2018, pp. 3733–3742.

[34] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.