Open-World Semi-Supervised Learning under Compound Distribution Shifts

Shijia Xu[1,2]
shijia@njust.edu.cn
Lin Zhao[1,3 *]
linzhao@njust.edu.cn
Jialiang Tang[1,2]
tangjialiang@njust.edu.cn
Guangyu Li[1,3]
guangyu.li2017@njust.edu.cn
Chen Gong[1,2,3 *]
chen.gong@njust.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[2] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, China
[3] Jiangsu Key Laboratory of Image and Video Understanding for Social Security, China

Abstract

Open-world Semi-Supervised Learning (OSSL) has drawn significant attention recently which assumes that the scarce labeled data and abundant unlabeled data for classifier training are sampled from different distributions. Existing methods typically assume that all unlabeled examples are drawn from the same domain following the same distribution. Nevertheless, this assumption may be violated as the unlabeled data are often collected from multiple unknown domains practically. Therefore, this paper tries to solve the OSSL problem under compound distribution shifts, in which the unlabeled data are from multiple unknown domains which may deviate from the distribution of labeled data. Specifically, we propose a novel Adversarial Mutual Information Disentanglement (AMID) framework to capture domain-invariant features for classifier training without the knowledge of domains. Particularly, we find that the class tokens of the pre-trained Vision Transformer (ViT) carry critical cues reflecting the styles of unlabeled data which can be deployed to attribute unlabeled data into different discovered domains. Subsequently, we train a feature encoder which captures the domain-invariant features shared among the attributed domains via designed adversarial confusion loss, so that the trained feature encoder can accurately represent the semantic information of unlabeled examples regardless of their domains. To further enhance feature disentanglement and enlarge the gap between useful domain-invariant features and interfered domain-specific features, we minimize the mutual information between the outputs of the encoders corresponding to domain-invariant features and domain-specific features. Comprehensive experiments conducted on various benchmark datasets demonstrate the effectiveness and generalizability of our approach in resolving the issue of compound distribution shifts in OSSL.

* Corresponding authors.
© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

1 Introduction

The success of deep neural networks largely depends on large-scale labeled datasets, which are quite difficult to acquire in many cases due to the unaffordable human labor and monetary cost. Therefore, Semi-Supervised Learning (SSL) [12, 18, 22, 32, 37] has emerged as an effective learning paradigm to mitigate the reliance on labeled data, which attempts to leverage scarce labeled data and abundant unlabeled data to train an accurate model.

Classical SSL methods [2, 6, 29, 36, 46] assume that the labeled and unlabeled data are drawn from the same distribution (see Figure 1(a)). However, in open-world scenarios, the class distribution or feature distribution of unlabeled data may differ from that of labeled data, as shown in Figure 1(b) and Figure 1(c). Therefore, various methods [7, 13, 14, 16, 25, 35, 43, 44] have been proposed to deal with the class mismatch problem. However, research on the problem of feature distribution mismatch among labeled data and unlabeled data is still in its early stages. CAFA [15] applies an adversarial feature adaptation strategy to eliminate the feature distribution mismatch between labeled and unlabeled data. BDA [17] proposes a weighted pseudo-labeling method to align the distributions of labeled and unlabeled data. Nevertheless, these methods mentioned above primarily focus on a simpler scenario where all unlabeled data are drawn from a single distribution, so the potential distribution inconsistency within the unlabeled data may be ignored (see Figure 1(c)). We call this case as compound feature distribution shifts. When faced with this more challenging compound feature distribution shifts, directly aligning the feature distributions of labeled and unlabeled data will result in negative transfer and suffer from performance degradation.

We think that the key to solve OSSL under compound distribution shifts is to explore domain-invariant features shared among different domains. Because domain-invariant features reveal semantic invariance across domains, which can promote transfer and help classifier training in OSSL. Therefore, a novel Adversarial Mutual Information Disentanglement (AMID) approach is proposed in this paper. First, considering that unlabeled data are from multiple unknown domains, and also inspired by the previous works [6, 8, 28, 38, 48] assuming the latent domain of images is reflected in their style, we utilize the class tokens extracted from the pre-trained DINO-ViT [9] model that is capable of capturing style information to discover the optimal latent domains of unlabeled data. Then we design an adversarial confusion loss to train a feature encoder which captures the domain-invariant features. In addition, to remove the interference of domain-specific features existing in domain-invariant features, a domain classifier is used to identify domain-specific features. After that, we utilize a variational formulation to estimate the upper bound of mutual information and minimize it between the domain-invariant and domain-specific

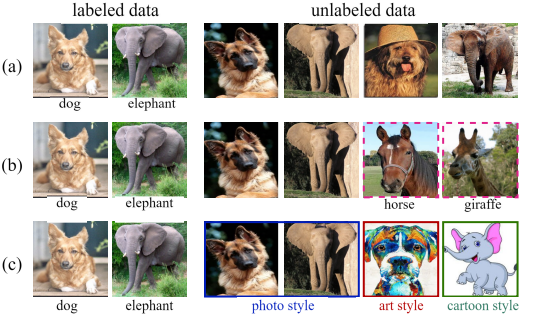

Figure 1: Problem illustration. (a) Traditional SSL setting. (b) Class mismatch: unlabeled data contain unknown classes not appeared in labeled data. (c) Compound distribution shifts: multiple domains are contained by unlabeled data. In this figure, the dashed pink box denotes images of mismatched classes, and the solid red, green and blue boxes denote the images belonging to different domains.

features. This strategy enhances the feature disentanglement and enlarges the gap between these features. We conduct experiments thoroughly on various datasets and demonstrate the superior performance of the proposed AMID to other typical SSL methods.

2 Related Work

In this part, we briefly introduce some studies that are closely relate to our work, including traditional semi-supervised learning and open-world semi-supervised learning.

Traditional Semi-Supervised Learning: Traditional SSL algorithms utilize both labeled and unlabeled data for training. There are mainly three classic strategies to train deep semi-supervised learning classifiers, namely: entropy minimization, consistency regularization, and generic regularization. Entropy minimization methods [12, 22, 32] minimize the label prediction entropy and enforce the networks to make confident predictions on unlabeled data. Consistency regularization methods [20, 29, 37] encourage consistent outputs for the same sample in temporally and spatially different models. For generic regularization methods [2, 3, 6, 36, 41, 46], semi-supervised learning algorithms utilize data augmentation strategies, combined with entropy minimization and consistency regularization to improve the model generalization performance.

Open-World Semi-Supervised Learning: Traditional SSL relies on the assumption that the labeled and unlabeled data are drawn from the same distribution. However, in open-world scenarios, the class distribution mismatch and feature distribution mismatch problems are common, which may lead to serious performance degradation in traditional SSL methods [31]. The major techniques to deal with the class distribution mismatch problem are example re-weighting [7, 14, 16, 35, 44] and open-set detection scoring [7, 14, 16, 35, 44]. The feature distribution mismatch problem is another more challenging problem in OSSL, but it has not been thoroughly studied. CAFA [15] adopts an adversarial feature adaptation strategy to align the distribution of unlabeled data to that of labeled data. BDA [17] designs a weighted pseudo-labeling method to solve the problem. However, CAFA [15] and BDA [17] only consider a single domain in unlabeled data, leading to significant performance decline when faced with unlabeled data from multiple unknown domains. Glocal [26] proposes to enhance the traditional pseudo-labeling mechanism by leveraging the cluster structure of unlabeled data to solve SSL under compound distribution shifts. However, it assigns pseudo-labels without accounting for variations across multiple domains within the unlabeled data, resulting in a failure to guarantee global robustness.

3 Methodology

3.1 Problem Description

In our semi-supervised learning under the setting of compound feature distribution shifts, we use $\mathcal{D}^L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ to denote a labeled set containing l labeled examples, where $\mathbf{x}_i \in \mathcal{R}^v$ (v represents the feature dimension) is the i-th training example and $y_i \in \{1, ..., c\}$ indicates the corresponding one-hot label. Let $\mathcal{D}^U = \{\mathcal{D}_m^U\}_{m=1}^{K} = \{\mathbf{x}_j\}_{j=1}^{u}$ be an unlabeled set composed of multiple unknown sub-domains, where $\mathcal{D}_m^U$ contains the unlabeled examples from the m-th domain. Here u denotes the number of unlabeled examples, and the number of sub-domains K is unknown. The labeled set and unlabeled set share the same label space containing c

classes. Like traditional SSL setting, here we assume $l \ll u$. The feature distributions of labeled data and unlabeled data are denoted as $p^L(\mathbf{x})$ and $p^U(\mathbf{x}) = \{p_m^U(\mathbf{x})\}_{m=1}^K$, respectively. Compound feature distribution shifts indicates that $p^L(\mathbf{x}) = p_m^U(\mathbf{x}), \exists m \in [1,\ldots,K]$ and $p_m^U(\mathbf{x}) \neq p_n^U(\mathbf{x}), \forall m,n \in [1,\ldots,K], m \neq n$. Then our target is to train a reliable classifier f: $\mathcal{R}^v \rightarrow \{1,...,c\}$ in classifying the test data under compound distribution shifts.

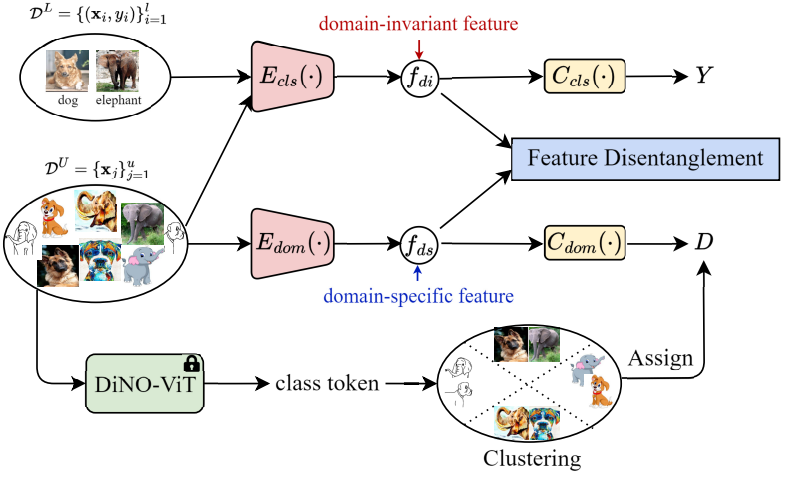

Figure 2: The overall framework of our proposed method, which assigns pseudo domain labels by clustering class token features extracted from the pre-trained DINO-ViT, and trains the domain-invariant feature extractor via feature disentanglement.

3.2 Overall Framework

The overall framework of our AMID approach is shown in Figure 2, in which $E_{cls}(\cdot)$, $C_{cls}(\cdot)$, $E_{dom}(\cdot)$, $C_{dom}(\cdot)$, and DiNO-ViT denote the domain-invariant feature extractor, class classifier, domain-specific feature extractor, domain classifier, and pre-trained DiNO-ViT. Our AMID includes three key components: Style-based cluster, adversarial confusion strategy, and mutual information minimization. The general framework of AMID is formulated as:

$$\begin{aligned}\min_{\theta_{E_{cls}},\theta_{E_{dom}},\theta_{C_{cls}},\theta_{C_{dom}}} & L_{ce}(\mathbf{x}_l;\theta_{E_{cls}},\theta_{C_{cls}}) + L_{ssl}(\mathbf{x}_u;\theta_{E_{cls}},\theta_{C_{cls}}) + L_{dom}(\mathbf{x}_u;\theta_{E_{dom}},\theta_{C_{dom}}) \\ & + L_{adv_c}(\mathbf{x}_u;\theta_{E_{cls}},\theta^*_{C_{dom}}) + L_{adv_d}(\mathbf{x}_u;\theta_{E_{dom}},\theta^*_{C_{cls}}) + L_{mi}(\mathbf{x}_u;\theta_{E_{cls}},\theta_{E_{dom}}),\end{aligned} \tag{1}$$

in which $\theta_{E_{cls}}, \theta_{E_{dom}}, \theta_{C_{cls}}, \theta_{C_{dom}}$ are the parameters of E_{cls}, E_{dom}, C_{cls} and C_{dom}, respectively. Here θ^* represents the parameter that does not update during gradient feedback, $\mathbf{x}_l$ means labeled examples, and $\mathbf{x}_u$ means unlabeled examples. We will detail each component below.

3.3 Style-based Clustering for Compound Unlabeled Data

When unlabeled data are drawn from compound domains, directly aligning the feature distributions of labeled and unlabeled data will result in negative transfer as the intrinsic inter-domain relationships are not considered. Inspired by works [6, 28] in domain adaption and generalization, we propose to cluster the compound unlabeled data using style information

to uncover latent domains. Specifically, we discover that the class tokens of the pre-trained DiNO-ViT [4] hold essential cues reflecting the styles of unlabeled data, as they capture global information through stretching, deforming, or flipping the objects in images. Hence, the class tokens are explored to assign unlabeled data pseudo domain labels.

Unlike methods [8, 48] that assume the number of latent domains is either known or predefined, we introduce an adaptive clustering method to predict the optimal number of latent domains using silhouette coefficient [34]. Assume the unlabeled samples are clustered into K categories $\{\mathcal{D}_m^U\}_{m=1}^K$. For each unlabeled data $\mathbf{x}_j$, we use $a(\mathbf{x}_j)$ to represent the average distance between $\mathbf{x}_j$ and all other samples in the cluster, and $b(\mathbf{x}_j)$ to denote the minimum average distance between $\mathbf{x}_j$ and all other samples in clusters it does not belong to. Suppose $\mathbf{x}_j$ belongs to $\{\mathcal{D}_m^U\}$, $a(\mathbf{x}_j)$ and $b(\mathbf{x}_j)$ can be formalized as:

$$\begin{aligned} a(\mathbf{x}_j) &= \frac{\sum_{\mathbf{x}_i \in \{\mathcal{D}_m^U\}, \mathbf{x}_i \neq \mathbf{x}_j} dist(s(\mathbf{x}_i), s(\mathbf{x}_j))}{|D_m^U| - 1}, \\ b(\mathbf{x}_j) &= \min_{1 \leq t \leq k, t \neq m} \frac{\sum_{\mathbf{x}_i \in D_t^U} dist(s(\mathbf{x}_i), s(\mathbf{x}_j))}{|D_t^U|}, \end{aligned} \tag{2}$$

where $dist(\cdot,\cdot)$ denotes the Euclidean distance, $s(\cdot)$ represents the global class token feature capturing style information, and $|\cdot|$ represents the number of samples in the cluster. The silhouette coefficient for a cluster can be formalized as follows:

$$S(K) = \sum_{\mathbf{x}_j \in D^U} \frac{b(\mathbf{x}_j) - a(\mathbf{x}_j)}{\max\{a(\mathbf{x}_j), b(\mathbf{x}_j)\}}. \tag{3}$$

Given that a larger silhouette coefficient indicates better clustering effectiveness, we can determine the optimal number of clusters as: $K^* = \arg\max_K S(K)$.

After automatically dividing the unlabeled data into K^* clusters, we assign a domain label d_j to the unlabeled data $\mathbf{x}_j$ based on the clustering result. Consequently, the unlabeled dataset can be represented as $\mathcal{D}^U = \{(\mathbf{x}_j, d_j)\}_{j=1}^u$.

3.4 Feature Disentanglement

We first train $E_{cls}(\cdot)$ and $C_{cls}(\cdot)$ based on the classical semi-supervised learning algorithm FixMatch [36] with the following classification losses:

$$\begin{aligned} L_{ce} &= \frac{1}{l} \sum_{\mathbf{x}_i \in \mathcal{D}^L} H(y_i, C_{cls}(E_{cls}(\mathbf{x}_i))), \\ L_{ssl} &= \frac{1}{u} \sum_{\mathbf{x}_j \in \mathcal{D}^U} 1(\max(q_j) \geq \tau) H(\hat{y}_j, C_{cls}(E_{cls}(A(\mathbf{x}_j)))), \end{aligned} \tag{4}$$

where $H(\cdot,\cdot)$ is the cross-entropy loss, $q_j = C_{cls}(E_{cls}(\alpha(\mathbf{x}_j))$, and $\hat{y}_j = argmax(q_j)$. Notation $\alpha(\cdot)$ and $A(\cdot)$ represent the weak and strong augmentation respectively, τ is the confidence threshold. Here $E_{cls}(\cdot)$ primarily captures the class-discriminative features (f_{di}) of the input. Because of the feature distribution mismatch between labeled and unlabeled data, these features are inevitably interfered by multiple domains information, resulting in negative transfer. Thus, we propose to recognize and exclude domain-specific features, making f_{di} only contain domain-invariant features. We try to extract domain-specific features (f_{ds})

by the domain classifier $C_{dom}(\cdot)$ using the unlabeled data and their assigned domain labels. The domain classification loss L_{dom} is expressed as follows:

$$L_{dom} = \frac{1}{u} \sum_{\mathbf{x}_j \in \mathcal{D}^U} H(d_j, C_{dom}(E_{dom}(\mathbf{x}_j))). \tag{5}$$

Domain-invariant Feature Learning via Adversarial Confusion strategy: Motivated by the adversarial learning technique in domain adaptation [10, 11, 27, 39, 42, 47], we design an adversarial class confusion loss function to reduce the domain-specific information present in f_{di}, so that f_{di} can be transformed into domain-invariant features. The class confusion loss is expressed as follows:

$$L_{adv_c} = -\frac{1}{u} \sum_{\mathbf{x}_j \in \mathcal{D}^U} H(d_j, C_{dom}(E_{cls}(\mathbf{x}_j))), \tag{6}$$

where L_{adv_c} is only used to optimize $E_{cls}(\cdot)$, which tries to confuse $C_{dom}(\cdot)$ but enables $C_{cls}(\cdot)$ to classify correctly.

In addition, to make $E_{dom}(\cdot)$ only focus on extracting the domain-specific features, we introduce a domain confusion loss to optimize $E_{dom}(\cdot)$, defined as follows:

$$L_{adv_d} = \sum_{\mathbf{x}_j \in \mathcal{D}^U} C_{cls}(E_{dom}(\mathbf{x}_j)) \log C_{cls}(E_{dom}(\mathbf{x}_j)). \tag{7}$$

Likewise, the domain confusion loss aims to confuse $C_{cls}(\cdot)$ while enabling $C_{dom}(\cdot)$ to make accurate classifications.

Disentangling via Mutual Information Minimization: To make the domain-invariant feature f_{di} more robust, we further exclude domain-specific information from f_{di} by minimizing the mutual information between f_{di} and f_{ds}. However, it is challenging to estimate the mutual information of high-dimensional vectors. Therefore, we leverage variational approximation of the mutual information to estimate the upper bound of mutual information, and minimize it to further enhance feature disentanglement.

Because techniques for estimating the mutual information between data and features have become relatively mature [1], we introduce input $\mathbf{x}$, from which both f_{di} and f_{ds} are extracted. Moreover, similar to IIB [23], we leverage variational approximation $r(f_{di})$ to be the approximation of the true marginal $p(f_{di})$, and variational distribution $q(f_{di}|\mathbf{x})$ to be the approximation of the conditional distribution $p(f_{di}|\mathbf{x})$. Meanwhile, $p(f_{di}|\mathbf{x})$ can be expressed by the class feature extractor $E_{cls}(\cdot)$. Likewise, $r(f_{ds})$ is the approximation to the true marginal $p(f_{ds})$, and $q(f_{ds}|\mathbf{x})$ approximates to $p(f_{ds}|\mathbf{x})$. $p(f_{ds}|\mathbf{x})$ is expressed by the domain feature extractor $E_{dom}(\cdot)$. Now the upper bound of mutual information can be written as (we present the detailed derivation in the appendices):

$$\begin{aligned} I(f_{di}; f_{ds}) \leq\ & E_{p(\mathbf{x})}[KL(q(f_{di}|\mathbf{x})||r(f_{di})) + KL(q(f_{ds}|\mathbf{x})||r(f_{ds}))] && \text{(8a)} \\ & + E_{p(\mathbf{x})}[E_{q(f_{di}|\mathbf{x})q(f_{ds}|\mathbf{x})}[logq(\mathbf{x}|f_{di}, f_{ds})]], && \text{(8b)} \end{aligned}$$

where Equation (8b) represents the reconstruction loss incurred in reconstructing input $\mathbf{x}$ using f_{di} and f_{ds}. So the mutual information loss can be expressed as:

$$\begin{aligned} L_{mi} =\ & E_{p(\mathbf{x})}[KL(q(f_{di}|\mathbf{x})||r(f_{di})) + KL(q(f_{ds}|\mathbf{x})||r(f_{ds}))] \\ & + ||Decoder(f_{di}, f_{ds}), \mathbf{x}||^2. \end{aligned} \tag{9}$$

4 Experiments

4.1 Experimental Setup

Dataset. To evaluate our proposed method, we perform experiments on three public multi-domain datasets. PACS [24] consists of four domains (*i.e.*, Photo, Art Painting, Cartoon, and Sketch). The four domains have the same label set of 7 classes. DIGITS is built upon four classic digits datasets (SVHN [30], MNIST [21], MNIST-M [10], SYNNUM [10]) of 10 classes. CIFAR-STL [49] is created by combining low-resolution images from CIFAR-10 [19] with high-resolution images from STL-10 [9] of 9 classes.

To evaluate the effectiveness and generalizability of our method in handling OSSL tasks under compound distribution shifts, we consider the classification accuracy across multiple testing sets (*i.e.* in-domain testing set where testing samples are collected from the same domain as the labeled data, out-of-domain testing set where testing samples are drawn from the domain(s) different from the labeled data, and overall testing set where testing set contains all in-domain and out-of-domain samples) following Glocal [26]. In the testing phase, we compute the classification accuracy of the model that exhibits the best performance in the testing sets.

Compared Methods. We compare our method with the following traditional SSL methods: FixMatch [36], FlexMatch [46], AdaMatch [33], and SoftMatch [6], and the following OSSL under feature distribution mismatch methods: CAFA [15], BDA [17], and Glocal [26].

Implementation Details. Our experiments are conducted under the uniform codebase USB [40] for fair comparison, with experimental setups mirroring those utilized in Glocal [26]. The selected backbone network is the Wide ResNet-37-2 [45]. We employ an SGD optimizer with a learning rate of 0.03 and a weight decay of 5e-4. The optimizer operates over 200 training epochs, each comprising 1024 iterations. The batch size is set to 64.

4.2 Performance Comparison

For PACS dataset, we randomly choose 5 or 10 samples per class from the training set of each domain as labeled data and use the rest as unlabeled data. Table 1 shows in-domain, out-of-domain, and overall classification accuracies of different methods on PACS. It can be observed that CAFA [15] and BDA [17] perform worse than traditional SSL methods, while our method achieves the best performance, which certifies that by learning domain-invariant features, our feature disentanglement overcomes negative transfer and is beneficial for boosting the learning performance of SSL. Moreover, compared to Glocal [26], our method explicitly extracts domain-specific features and focuses on domain-invariant features to guarantee global robustness. For instance, when the labeled domain draws from Photo, our method outperforms the previous SOTA method by 10.6% and 8.12% in overall accuracy for the 5 and 10 labels per class, respectively.

Meanwhile, we conduct extensive experiments on DIGITS and CIFAR-STL to demonstrate the effectiveness and generalizability of our approach. The experimental results are reported in Table 2 and Table 3, respectively, showing the same trend as those for PACS. It is observed that our method consistently achieves the best performance. On DIGITS, the superiority of our method is relatively less pronounced compared to the other two datasets. This could be attributed to the simpler and more recognizable texture of digital numbers.

Number of Labels	35 (5 labels per class)											
Labeled Domain	Photo			Art			Cartoon			Sketch		
Test Data	In	Out	All	In	Out	All	In	Out	All	In	Out	All
FixMatch [36]	70.58	15.24	23.86	32.69	19.95	21.49	58.65	25.49	32.57	70.63	21.63	37.33
FlexMatch [46]	75.88	19.76	27.33	57.69	26.81	30.99	71.72	30.01	38.61	56.96	29.27	34.95
AdaMatch [3]	62.94	23.33	29.11	54.32	29.18	34.06	61.60	27.04	33.17	68.86	30.08	43.66
SoftMatch [6]	70.00	23.69	30.00	33.65	21.07	23.56	72.15	32.21	40.10	62.53	28.62	41.19
CAFA [15]	17.65	20.24	19.50	18.75	23.32	22.18	33.76	18.24	20.79	21.27	18.21	19.01
BDA [17]	32.35	14.05	16.44	21.63	19.95	20.20	43.04	21.21	25.84	47.09	14.96	26.93
Glocal [26]	82.94	32.02	39.80	56.73	29.30	30.40	79.75	32.60	42.97	75.44	**43.09**	54.65
Ours	**85.88**	**45.00**	**50.40**	**69.23**	**43.02**	**49.21**	**82.70**	**42.95**	**52.77**	**86.33**	36.26	**55.25**
Number of Labels	70 (10 labels per class)											
Labeled Domain	Photo			Art			Cartoon			Sketch		
Test Data	In	Out	All	In	Out	All	In	Out	All	In	Out	All
FixMatch [36]	74.70	11.19	21.78	64.90	28.55	33.37	88.18	32.47	44.95	86.58	27.97	49.50
FlexMatch [46]	77.65	21.90	30.69	55.76	31.55	32.28	83.96	32.60	44.46	78.73	42.76	55.84
AdaMatch [3]	79.41	28.69	36.14	55.77	24.69	28.91	88.18	35.96	46.63	83.04	28.46	48.61
SoftMatch [6]	82.35	26.19	35.05	54.80	31.42	32.57	85.64	35.71	45.15	79.75	28.13	47.52
CAFA [15]	33.53	21.07	22.97	22.59	22.44	22.28	48.95	28.98	33.27	38.48	20.16	24.75
BDA [17]	54.12	15.48	21.98	26.44	11.10	13.76	45.15	25.87	29.80	51.14	15.12	29.21
Glocal [26]	**88.24**	39.05	46.93	75.48	**52.37**	**56.44**	88.19	44.37	53.56	84.05	38.05	54.75
Ours	**88.24**	**46.79**	**55.05**	**78.36**	45.76	51.09	**89.45**	**55.24**	**62.18**	**90.13**	**52.68**	**63.17**

Table 1: Classification accuracies(%) on PACS. The best results are highlighted in **bold**, while the second best result is highlighted with an underline. The notations "In", "Out", and "All" denote that the test data are from the in-domain, out-of-domain and overall domain.

Number of Labels	50 (5 labels per class)											
Labeled Domain	SVHN			MNIST			MNIST-M			SYNNUM		
Test Data	In	Out	All	In	Out	All	In	Out	All	In	Out	All
FixMatch [36]	32.40	10.07	15.65	94.20	11.40	32.10	**96.50**	10.03	31.65	96.70	10.03	31.70
FlexMatch [46]	35.60	10.03	16.43	**96.80**	10.03	31.72	78.30	76.43	76.90	78.00	16.53	31.90
AdaMatch [3]	32.20	13.17	17.92	95.80	53.33	63.95	72.60	73.57	73.32	95.30	56.83	66.45
SoftMatch [6]	35.20	10.13	16.40	95.70	45.50	58.05	95.70	42.77	45.20	96.80	**59.70**	**69.43**
CAFA [15]	12.60	14.30	12.60	44.00	16.60	23.13	14.50	16.27	14.50	12.50	16.53	15.32
BDA [17]	11.10	10.80	10.65	57.20	11.17	21.85	14.20	13.90	13.98	11.90	11.07	11.15
Glocal [26]	40.70	39.13	39.52	94.80	**77.23**	**81.63**	66.00	46.30	51.22	94.70	59.07	67.97
Ours	**78.06**	**49.65**	**52.92**	96.20	44.70	57.57	94.20	**79.90**	**83.25**	**98.20**	54.63	64.42
Number of Labels	100 (10 labels per class)											
Labeled Domain	SVHN			MNIST			MNIST-M			SYNNUM		
Test Data	In	Out	All	In	Out	All	In	Out	All	In	Out	All
FixMatch [36]	**90.00**	37.10	49.73	97.37	12.93	33.22	97.50	39.70	53.97	98.00	43.97	57.33
FlexMatch [46]	87.70	59.43	65.50	97.10	39.43	53.77	96.90	39.37	53.85	98.40	60.80	70.20
AdaMatch [3]	89.00	62.63	69.17	96.90	57.37	67.00	90.60	89.57	89.82	98.20	59.77	69.37
SoftMatch [6]	89.20	64.10	70.15	96.40	51.77	62.55	96.50	89.87	**91.70**	98.20	39.00	53.80
CAFA [15]	13.00	12.87	12.58	60.90	19.00	29.48	15.50	14.93	14.78	18.40	15.83	16.25
BDA [17]	12.80	13.07	12.78	68.12	12.76	26.53	11.60	12.83	12.38	14.20	12.73	12.93
Glocal [26]	88.90	64.57	70.47	96.40	89.07	90.62	96.70	82.37	85.75	98.10	63.20	71.90
Ours	89.30	**67.43**	**72.40**	**97.90**	**91.80**	**92.95**	**97.70**	**89.93**	90.60	**98.60**	**63.90**	**72.00**

Table 2: Classification accuracies(%) on DIGITS

4.3 Ablation Study

To investigate the effectiveness of different key components in our AMID, we conduct the following ablative experiments, including: 1) we remove the adversarial confusion strategy

Number of Labels	45 (5 labels per class)						90 (10 labels per class)					
Labeled Domain	CIFAR			STL			CIFAR			STL		
Test Data	In	Out	All	In	Out	All	In	Out	All	In	Out	All
FixMatch [36]	52.73	11.11	31.92	59.82	11.12	35.47	67.16	15.60	41.38	**73.37**	11.11	42.24
FlexMatch [46]	47.49	11.11	29.30	53.42	18.64	36.03	61.47	30.53	46.00	68.02	19.18	43.60
Adamatch [3]	56.84	20.16	38.50	63.80	20.46	42.13	67.00	30.44	48.72	70.64	17.68	44.16
SoftMatch [6]	55.96	12.40	34.18	63.18	20.60	41.89	67.56	16.10	41.83	70.26	19.94	45.17
CAFA [15]	24.31	22.15	23.23	26.87	22.71	24.79	26.94	24.60	25.77	28.73	22.15	25.44
BDA [17]	23.42	20.18	21.80	23.96	19.12	21.54	26.85	27.23	27.04	29.18	23.42	26.30
Glocal [26]	55.89	46.87	51.38	69.49	47.15	58.32	66.06	**60.54**	**63.30**	71.96	**61.42**	**66.69**
Ours	**68.71**	**56.75**	**62.73**	**69.67**	**56.31**	**62.99**	**72.18**	53.96	63.07	**73.55**	60.55	**67.05**

Table 3: Classification accuracies on CIFAR-STL

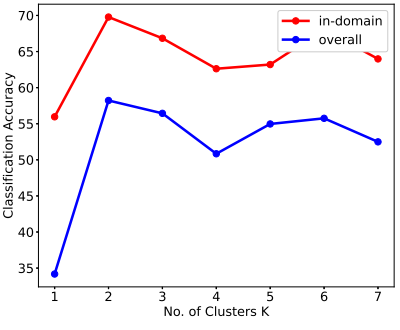

Figure 3: The ablation study on the number of clusters K on CIFAR-STL.

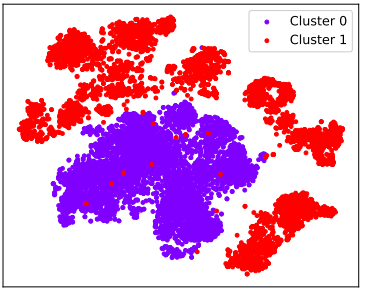

Figure 4: The t-SNE visualization of clustering result on CIFAR-STL with $K = 2$.

while keeping others fixed, denoted as "w/o adv"; 2) we remove the mutual information minimization term while keeping others fixed, denoted as "w/o mi". Table 4 shows the ablative results on the art domain of the PACS dataset. When either the adversarial confusion strategy or the mutual information minimization term is removed, we observe that AMID suffers considerable performance degradation in both in-domain and out-of-domain accuracy. Particularly, the absence of the adversarial confusion strategy can lead to a significant drop in out-of-domain accuracy. This indicates the adversarial confusion strategy helps capture the domain-invariant features shared in different domains to improve performance. The mutual information minimization term helps enlarge the gap between domain-invariant features and domain-specific features to improve in-domain accuracy.

Number of Labels	35 (5 labels per class)		
Labeled Domain	Art		
Test Data	In	Out	All
ours w/o adv	67.79	29.93	36.53
ours w/o mi	61.53	40.52	41.45
Ours	**69.23**	**43.02**	**49.21**

Table 4: Classification accuracies(%) of ablation experiments on the art domain of the PACS.

Moreover, we conduct ablation study on the number of clusters K in the adaptive clustering method on the CIFAR-STL dataset, which is varied from 1 to 7. The result is shown in Figure 3. The optimal number of clusters selected by silhouette coefficien is 2, which is consistent with the actual potential domains of the CIFAR-STL dataset. Moreover, both too less and too many number of clusters would hurt the result. This is because a small cluster number is unable to learn the diversity of styles in potential domains, and a large cluster number

would increase the difficulty of domain classification, further making it hard to achieve the feature disentanglement. In Figure 4, we show the t-SNE visualization of clustering results using the class tokens on the CIFAR-STL dataset. It can be clearly seen that class tokens do indeed capture the style information of different domains.

5 Conclusion

In this paper, we propose a novel AMID method to tackle OSSL under compound distribution shifts. Specifically, we conduct style-based clustering to divide unlabeled data into different latent domains and assign pseudo domain labels, which helps to extract domain-specific features. Then, feature disentanglement is conducted using the adversarial confusion strategy and the mutual information minimization, which excludes domain-specific features and captures domain-invariant features, so that the domain-invariant features can represent the semantic information regardless of their domains. Comprehensive experiments show the effectiveness and robustness of our AMID framework in solving OSSL under compound distribution shifts problems.

Acknowledgement

This research is supported by NSF of China (Nos: 62336003, 12371510, 62172222, 62006119), NSF of Jiangsu Province (No: BZ2021013), NSF for Distinguished Young Scholar of Jiangsu Province (No: BK20220080), the Fundamental Research Funds for the Central Universities (Nos: 30920032202, 30921013114), the "111" Program (No: B13022), National Key R&D Program for Key International S&T Cooperation Projects (SQ2023YFE0102775).

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[5] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7119–7129, 2022.

[6] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

[7] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3569–3576, 2020.

[8] Ziliang Chen, Jingyu Zhuang, Xiaodan Liang, and Liang Lin. Blending-target domain adaptation by adversarial meta-adaptation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2248–2257, 2019.

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems*, 17, 2004.

[13] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020.

[14] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8310–8319, 2021.

[15] Zhuo Huang, Chao Xue, Bo Han, Jian Yang, and Chen Gong. Universal semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:26714–26725, 2021.

[16] Zhuo Huang, Jian Yang, and Chen Gong. They are not completely useless: Towards recycling transferable unlabeled data for class-mismatched semi-supervised learning. *IEEE Transactions on Multimedia*, 2022.

[17] Lin-Han Jia, Lan-Zhe Guo, Zhi Zhou, Jie-Jing Shao, Yuke Xiang, and Yu-Feng Li. Bidirectional adaptation for robust semi-supervised learning with inconsistent data distributions. In *International Conference on Machine Learning*, pages 14886–14901. PMLR, 2023.

[18] Tao Jiang, Wanqing Chen, Hangping Zhou, Jinyang He, and Peihan Qi. Towards semi-supervised classification of abnormal spectrum signals based on deep learning. *Chinese Journal of Electronics*, 33(3):721–731, 2024.

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning*, page 896, 2013.

[23] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7399–7407, 2022.

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5542–5550, 2017.

[25] Mingyu Li, Tao Zhou, Zhuo Huang, Jian Yang, Jie Yang, and Chen Gong. Dynamic weighted adversarial learning for semi-supervised classification under intersectional class mismatch. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(4):1–24, 2024.

[26] Zekun Li, Lei Qi, Yawen Li, Yinghuan Shi, and Yang Gao. Open-domain semi-supervised learning via glocal cluster structure exploitation. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–15, 2024.

[27] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11749–11756, 2020.

[29] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2011, page 7, 2011.

[31] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

[32] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

[33] Becca Roelofs, David Berthelot, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *International Conference on Learning Representations*, 2022.

[34] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[35] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34:25956–25967, 2021.

[36] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[38] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[40] Yidong Wang, Hao Chen, Yue Fan, Wang Sun, Ran Tao, Wenxin Hou, Renjie Wang, Linyi Yang, Zhi Zhou, and Lan-Zhe Guo. Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35: 3938–3961, 2022.

[41] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023.

[42] Wang Xuesong, Zhao Jijuan, Cheng Yuhu, and Yu Qiang. Joint feature representation and classifier learning based unsupervised domain adaption elm. *Chinese Journal of Electronics*, 30(1):109–118, 2021.

[43] Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14421–14430, 2022.

[44] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Proceedings of the European Conference on Computer Vision*, pages 438–454, 2020.

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[46] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

[47] Yun Zhang, Nianbin Wang, and Shaobin Cai. Learning domain-invariant and discriminative features for homogeneous unsupervised domain adaptation. *Chinese Journal of Electronics*, 29(6):1119–1125, 2020.

[48] 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*, 60:1–13, 2021.

[49] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.