Label-Efficient Domain Generalization via Collaborative Exploration and Generalization

Junkun Yuan*
Xu Ma*
Zhejiang University
Hangzhou, China
yuanjk@zju.edu.cn
maxu@zju.edu.cn

Defang Chen
Zhejiang University
Hangzhou, China
defchern@zju.edu.cn

Kun Kuang†
Zhejiang University
Hangzhou, China
Shanghai AI Laboratory
Shanghai, China
kunkuang@zju.edu.cn

Fei Wu
Zhejiang University
Hangzhou, China
Shanghai Institute for Advanced
Study of Zhejiang University
Shanghai, China
wufei@zju.edu.cn

Lanfen Lin
Zhejiang University
Hangzhou, China
llf@zju.edu.cn

ABSTRACT

Considerable progress has been made in domain generalization (DG) which aims to learn a generalizable model from multiple well-annotated source domains to unknown target domains. However, it can be prohibitively expensive to obtain sufficient annotation for source datasets in many real scenarios. To escape from the dilemma between domain generalization and annotation costs, in this paper, we introduce a novel task named label-efficient domain generalization (LEDG) to enable model generalization with label-limited source domains. To address this challenging task, we propose a novel framework called Collaborative Exploration and Generalization (CEG) which jointly optimizes active exploration and semi-supervised generalization. Specifically, in active exploration, to explore class and domain discriminability while avoiding information divergence and redundancy, we query the labels of the samples with the highest overall ranking of class uncertainty, domain representativeness, and information diversity. In semi-supervised generalization, we design MixUp-based intra- and inter-domain knowledge augmentation to expand domain knowledge and generalize domain invariance. We unify active exploration and semi-supervised generalization in a collaborative way and promote mutual enhancement between them, boosting model generalization with limited annotation. Extensive experiments show that CEG yields superior generalization performance. In particular, CEG can even use only 5%

data annotation budget to achieve competitive results compared to the previous DG methods with fully labeled data on PACS dataset.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

domain generalization; image classification; label-efficient learning

1 INTRODUCTION

Despite the remarkable success achieved by modern machine learning algorithms in visual recognition [17, 51, 56], it heavily relies on the i.i.d. assumption [53] that training and test datasets should have a consistent statistical pattern. Since machine learning systems are usually deployed in a wide range of scenarios where the test data are unknown in advance, it may inevitably result in serious model performance degradation when there exists a distinct *distribution/domain shift* [44] between the training and test data.

With awareness of this problem, *domain generalization* (DG) [5] is introduced to extract domain invariance from multiple well-annotated source datasets/domains and train a generalizable model to unknown target domains. Lots of favorable DG algorithms [6, 15, 42, 46, 61, 71, 73] have been proposed recently, however, these methods may need to be fed with a large amount of labeled multi-source data for identifying domain invariance and improving model generalization. It might impede the realization of the DG approaches in many real-world applications where labeling massive data could be expensive or even infeasible. For example, a highly accurate and robust system for the detection of lung lesion images of COVID-19 patients may demand a large number of labeled medical images

*Both authors contributed equally to this research.
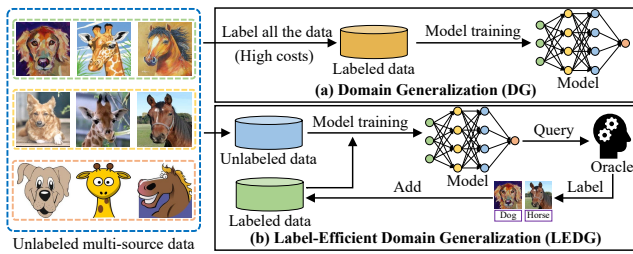†Corresponding author

Figure 1: Comparison between the conventional DG (a) and the proposed LEDG (b) tasks. DG may need to label all source data. In comparison, LEDG queries the labels of a small quota of data with limited annotation budget and boosts domain generalization by exploiting both labeled and unlabeled data.

from different hospitals as the source data for training [16], but it could be impractical to require numerous experienced clinicians to complete the annotation. Therefore, a *dilemma* is encountered: The requirements of obtaining massive labeled source data for training a generalizable model may hard to be met in realistic scenarios due to the limited annotation budget. Meanwhile, without sufficient labeled data to provide adequate information of multi-source distribution, improving model generalization by identifying and learning the domain invariance is at serious risk of being misled.

To escape from the dilemma, we introduce a more practical task named *label-efficient domain generalization* (LEDG) to enable model generalization with label-limited source domains as shown in Figure 1. Instead of getting fully labeled data, the LEDG task unleashes the potential of budget-limited annotation by querying the labels of a small quota of informative data, and leverages both the labeled and unlabeled data to improve domain generalization. LEDG permits the learning of generalizable models in real scenarios, but it could be much more challenging. The first challenge is from the clear domain distinctions that could exist in the multi-source data, which constitutes enormous obstacles for selecting the most informative samples and learning adequate information of multi-source distribution. Meanwhile, the second challenge is from the discrepant distributions that labeled and unlabeled data may be subject to, making the simultaneous utilization of them for extracting domain invariance and promoting model generalization extremely difficult.

Active learning (AL) [1, 19, 24, 25, 45, 57] and *semi-supervised learning* (SSL) [3, 4, 50, 52] provide possible solutions to the introduced LEDG task. AL aims to query the labels of high-quality samples and SSL leverages the unlabeled data to improve performance with limited labeled data. However, the existing AL and SSL methods mostly depend on the i.i.d. assumption, hence may not be favorably extended to the generalization scenarios under distinct domain shifts. *Semi-supervised domain generalization (SSDG)* [36, 47, 59, 65, 70] tackles domain shift under the SSL setting. But some data directly assumed to be labeled in this task might not be helpful for generalization improvement but could increase the annotation costs. Thus, it is imperative to find a solution to the challenging LEDG task for getting rid of the raised dilemma between domain generalization and annotation costs, realizing more practical training of generalizable models in real-world scenarios.

To address the LEDG task, in this paper, we propose a novel framework called Collaborative Exploration and Generalization (CEG) which jointly optimizes active exploration and semi-supervised generalization. In active exploration, to unleash the power of the limited annotation, we query the labels of the samples with the highest overall ranking of class uncertainty, domain representativeness, and information diversity, exploring class and domain discriminability while avoiding information divergence and redundancy. In semi-supervised generalization, we augment intra- and inter-domain knowledge with MixUp [67] to expand domain knowledge and generalize domain invariance. An augmentation consistency constraint for unlabeled data and a prediction supervision for labeled data are further included to improve performance. We unify active exploration and semi-supervised generalization in a collaborative way by repeating them alternately, promoting closed-loop mutual enhancement between them for effective learning of domain invariance and label-efficient training of generalizable models.

Our contributions are listed in the following. (1) We introduce a more practical task named label-efficient domain generalization to permit generalization learning in real-world scenarios by tackling the dilemma between domain generalization and annotation costs. (2) To solve this challenging task, we propose a semi-supervised active learning-based framework CEG to unify active query-based distribution exploration and semi-supervised training-based model generalization in a collaborative way, achieving closed-loop mutual enhancement between them. (3) Extensive experiments show the superior generalization performance of CEG, which can even achieve competitive results using 5% annotation budget compared to the previous DG methods with full annotation on PACS dataset.

2 RELATED WORK

Domain Generalization (DG). Different from domain adaptation (DA) [7–10, 12, 20, 34, 38, 39, 58, 62, 63] which adapts models from the source domain to the target, DG [5] assumes that the target domain is unknown during training and aims to train a generalizable model from the source domains. Increasing DG methods [15, 21, 26–29, 41, 42, 46, 48, 55, 64, 66, 72] are proposed recently, they popularize various strategies via invariant representation learning [13, 32, 33, 43, 69], meta-learning [2, 13, 31, 35, 49], data augmentation [6, 18, 22, 61, 68, 71, 73], and others [14, 37, 60]. But they mostly need the fully labeled data to learn generalization.

Semi-Supervised Domain Generalization (SSDG) [36, 47, 59, 65, 70] aims to reduce the reliance of DG on annotation via pseudo-labeling [59], consistency learning [70], or bias filtering [65]. For example, StyleMatch [70] combines consistency learning, model uncertainty learning, and style augmentation to utilize the annotation for improving model robustness. However, partial samples assumed to be labeled in the SSDG task may not be informative for boosting model generalization but increase the annotation costs.

Semi-Supervised Learning (SSL). SSL [3, 4, 23, 50, 52] is a practicable way to use both labeled and unlabeled data. For example, MeanTeacher [52] achieves significant performance by using the labeled data to optimize a student model, the prediction of which is constrained to be consistent with the prediction of a teacher model. But most of the SSL methods rely on the i.i.d. assumption, which can impair their generalization performance under domain shift.
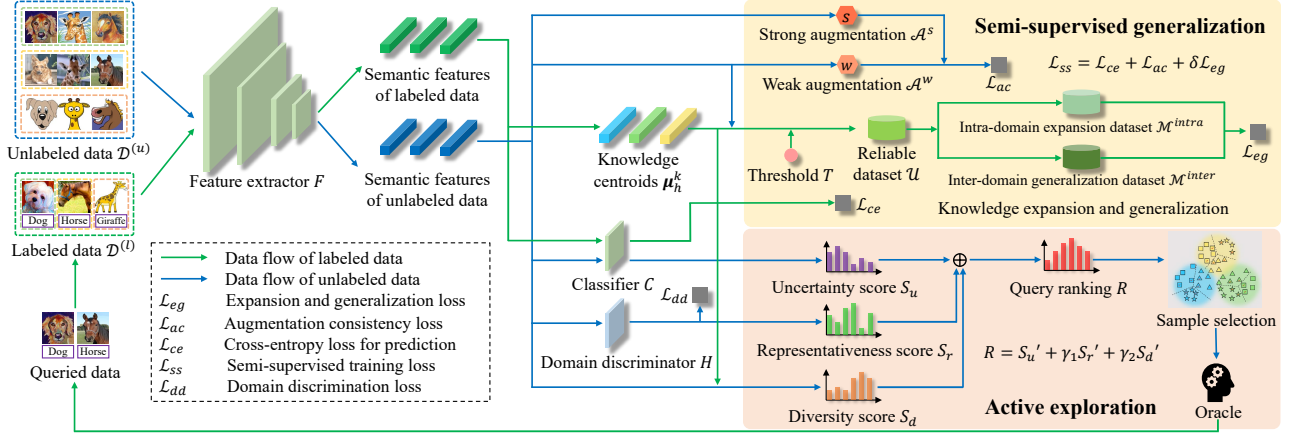
Figure 2: Overview of the Collaborative Exploration and Generalization (CEG) framework. In active exploration, samples with the highest overall ranking (R) of class uncertainty (S_u), domain representativeness (S_r), and information diversity (S_d) are selected to query for learning multi-source distribution. In semi-supervised generalization, expansion and generalization ($\mathcal{L}_{eg}$) of the known knowledge, as well as augmentation consistency ($\mathcal{L}_{ac}$) for the unlabeled data and prediction supervision ($\mathcal{L}_{ce}$) for the labeled data are devised to extract domain invariance and improve model generalization via semi-supervised training ($\mathcal{L}_{ss}$).

Active Learning (AL). AL [1, 19, 24, 25, 45, 57] aims to select high-quality data for querying the labels. Pool-based AL [1, 19, 25, 45] is the most popular that chooses samples from an unlabeled pool and hand them over to the oracle to label. The labeled samples are then added to a labeled pool as newly acquired knowledge. Some successful uncertainty [1, 24, 57] and diversity [1, 45] based methods select uncertain and diverse samples for learning task boundary and comprehensive information, respectively. However, the AL algorithms are mainly designed for single-domain data, thus may not be directly extended to the generalization scenarios.

3 METHOD

3.1 Label-Efficient Domain Generalization

We begin with the task setting of the introduced Label-Efficient Domain Generalization (LEDG). In the LEDG task, we have K unlabeled source datasets $\{\mathcal{D}^1, ..., \mathcal{D}^K\}$ sampled from different data distributions $\{P(X^1, Y^1), ..., P(X^K, Y^K)\}$, respectively. There are N^k unlabeled data points being sampled for each dataset $\mathcal{D}^k$, i.e., $\mathcal{D}^k = \{x_i^k\}_{i=1}^{N^k}$, for $k = 1, ..., K$. We further have an annotation budget B, i.e., the maximum number of samples that we are allowed to query their class labels. Each sample pair (x, y) is defined on the image and label joint space $X \times Y$. Besides, the domain label p_i^k for each sample x_i^k is given in our task. We consider a classification model G composed of a feature extractor F and a classifier head C, i.e., $G = C \circ F$. The goal of LEDG is to train the model G by utilizing the unlabeled multi-source data $\{\mathcal{D}^k\}_{k=1}^K$ as well as the limited annotation budget B for improving the generalization performance of the model on the target domains with unknown distributions. For convenience of the statement of our method, we denote the dataset consists of all the labeled (queried) samples $(x_i^{(l)}, y_i^{(l)})$ as $\mathcal{D}^{(l)} = \{(x_i^{(l)}, y_i^{(l)})\}_{i=1}^{N^{(l)}}$ and the dataset with all the unlabeled (not queried) samples $x_i^{(u)}$ as $\mathcal{D}^{(u)} = \{x_i^{(u)}\}_{i=1}^{N^{(u)}}$, where $N^{(l)}$ and $N^{(u)}$

are the data size of the labeled and unlabeled datasets, respectively. It is obvious that the whole data size $N^{(l)} + N^{(u)} = N^1 + N^2 + ... + N^K$.

Our insight for this challenging task is to consider the labeled and unlabeled samples as the *"known" and "unknown" regions* of the multi-source distribution, respectively. In view of this, the core idea of our solution is to: (1) explore the key knowledge hidden in the unknown regions via active query for adequate multi-source distribution learning, (2) extract and generalize the domain invariance contained in the obtained knowledge in both the known and unknown regions via semi-supervised training, and (3) make the active query-based exploration and semi-supervised training-based generalization complement and promote each other to train a generalizable model. An overview of our framework, i.e., Collaborative Exploration and Generalization (CEG), is shown in Figure 2.

3.2 Active Exploration

There might be distinct domain divergence among the data distributions of the source domains. Meanwhile, each source domain contains the discriminative information of class boundary which is essential for the prediction task. Thus, we take class and domain discriminability as the key knowledge for learning the multi-source distribution in the active exploration. In light of this, we present to select the samples with high class uncertainty and domain representativeness. To avoid information redundancy, we further take information diversity into consideration. Figure 3 provides revealing insights into the elaborate strategy for the active exploration.

To capture the key knowledge of class discriminability, we select the samples with high class uncertainty. Specifically, we adopt the margin of the top two model predictions to choose class-ambiguous samples to query. Let G_h be the h-th dimension of the class prediction of the model G (after softmax operation), then the class uncertainty score S_u for each unlabeled sample $x^{(u)}$ is defined as

$$S_u(x^{(u)}) = \left(1 - \left(\max_h G_h\left(x^{(u)}\right) - \max_{h'|h' \neq h} G_{h'}\left(x^{(u)}\right)\right)\right). \quad (1)$$
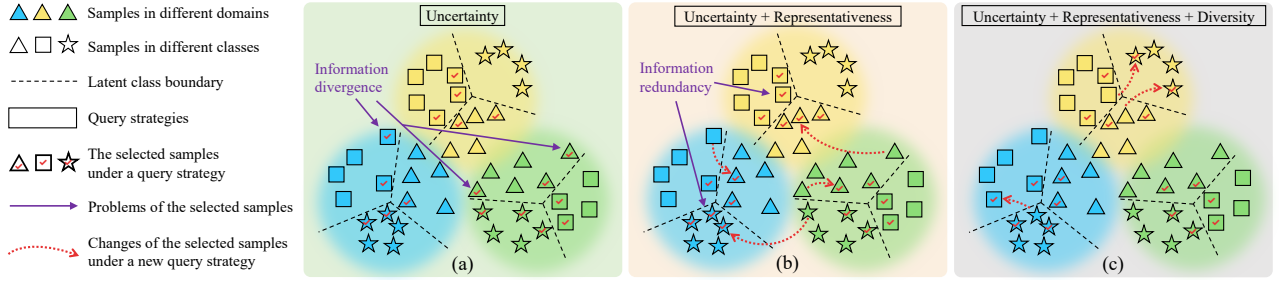
Figure 3: Comparisons of different query strategies for the active exploration. (a): Uncertainty criterion selects class-ambiguous samples but can cause information divergence, i.e., it may fail to select the representative samples for each domain. (b): The introduction of representativeness criterion helps to capture domain characteristics but can lead to information redundancy, i.e., it may select adjacent samples and hence reduce the efficiency of the limited annotation budget. (c): The further introduction of information diversity criterion guides the model to comprehensive knowledge learning of the multi-source distribution.

We tend to query the samples with high uncertainty scores for their class-ambiguous. Labeling these samples provides the key knowledge of class discriminability, which helps the model to figure out class boundary and boosts its performance of class prediction.

Different from the single-domain scenario considered in the AL methods [24, 57], multiple source domains may lead to an information divergence problem here, i.e., the selected high-uncertainty samples are scattered at the domain boundary (see Figure 3 (a)). To sufficiently explore and grasp information of the multi-source distribution, the selected samples are required to represent the characteristics of each source domain. Therefore, given domain label $p^{(u)}$ of each unlabeled sample $\boldsymbol{x}^{(u)}$, we first train a domain discriminator model H with a domain discriminability loss:

$$\mathcal{L}_{dd} = \mathbb{E}_{\boldsymbol{x}^{(u)} \in \mathcal{D}^{(u)}} \ell(H(\boldsymbol{x}^{(u)}), p^{(u)}), \tag{2}$$

where ℓ is the cross-entropy loss. Let H_h be the h-th dimension of the domain prediction of the model H, we then define domain representativeness score S_r for each unlabeled sample $\boldsymbol{x}^{(u)}$ as:

$$S_r(\boldsymbol{x}^{(u)}) = \max_h H_h(\boldsymbol{x}^{(u)}). \tag{3}$$

Note that different from the class discriminability learning with class ambiguous data, here, we select the samples with high representativeness score, i.e., high domain confidence. It prevents the model from learning class discriminability in the remote areas of each source domain and hence loosing domain characteristics.

Then an information redundancy problem has arisen, i.e., the selected samples with high class uncertainty and domain representativeness may gather together (see Figure 3 (b)), which wastes the limited annotation budget. To disperse the information, we choose the samples that are far away from the known domain-class knowledge of the labeled data. We make a knowledge dataset $\mathcal{K}_h^k$ be composed of the labeled data $(\boldsymbol{x}^{(l)}, y^{(l)}) \in \mathcal{D}^{(l)}$ that belongs to domain k and class h (if there is no such a sample in $\mathcal{D}^{(l)}$, then $\mathcal{K}_h^k = \emptyset$). Let $|\mathcal{K}_h^k|$ be the number of samples in the knowledge dataset $\mathcal{K}_h^k$ and F be the feature extractor. We generate *knowledge centroids* $\boldsymbol{\mu}_h^k$ for the known regions in the semantic feature space:

$$\boldsymbol{\mu}_h^k = \frac{1}{|\mathcal{K}_h^k|} \sum_{(\boldsymbol{x}^{(l)}, y^{(l)}) \in \mathcal{K}_h^k} F(\boldsymbol{x}^{(l)}). \tag{4}$$

We let a set $\mathcal{B}$ be composed of all the knowledge centroids $\boldsymbol{\mu}_h^k$ if $\mathcal{K}_h^k \neq \emptyset$. Then, we define information diversity score S_d as

$$S_d(\boldsymbol{x}^{(u)}) = \min_{\boldsymbol{\mu}_h^k \in \mathcal{B}} \text{dist}(\boldsymbol{x}^{(u)}, \boldsymbol{\mu}_h^k), \tag{5}$$

where $\text{dist}(\cdot, \cdot)$ is a distance metric used as cosine distance in experiments. We tend to choose samples with high diversity score, i.e., far away from the closest centroids, facilitating the unknown exploration for comprehensive learning of multi-source distribution.

To avoid numerical issues, we integrate uncertainty score S_u, representativeness score S_r, and diversity score S_d by adopting their rankings, which we denote as S_u', S_r', and S_d', respectively. Finally, we have an overall query ranking R for each unlabeled sample $\boldsymbol{x}^{(u)}$:

$$R(\boldsymbol{x}^{(u)}) = S_u'(\boldsymbol{x}^{(u)}) + \gamma_1 S_r'(\boldsymbol{x}^{(u)}) + \gamma_2 S_d'(\boldsymbol{x}^{(u)}), \tag{6}$$

where γ_1 and γ_2 are trade-off hyper-parameters. Note that we rank each score, i.e., S_u, S_r, and S_d, from high to low, for selecting the most informative samples of the multi-source data distribution.

3.3 Semi-Supervised Generalization

With active query-based exploration, we have a small quota of labeled data and massive unlabeled data, i.e., small range of known regions and large range of unknown regions of the data distribution. In semi-supervised generalization, we aim to expand domain knowledge and learn domain invariance via MixUp-based intra- and inter-domain knowledge augmentation as shown in Figure 4.

We start by defining the unlabeled samples that close to the knowledge centroids $\boldsymbol{\mu}_h^k$ as "reliable samples", and construct a reliable dataset $\mathcal{U}$ with expansion threshold T to tune reliable range:

$$\mathcal{U} = \{(\boldsymbol{x}^{(u)}, y^{(u)}) | \min_{\boldsymbol{\mu}_h^k \in \mathcal{B}} \text{dist}(\boldsymbol{x}^{(u)}, \boldsymbol{\mu}_h^k) < T\}, \tag{7}$$

where pseudo label $y^{(u)}$ of each unlabeled sample $\boldsymbol{x}^{(u)}$ is assigned by the nearest knowledge centroids $\boldsymbol{\mu}_h^k$, that is,

$$y^{(u)} = \arg\min_{h:\boldsymbol{\mu}_h^k \in \mathcal{B}} \text{dist}(\boldsymbol{x}^{(u)}, \boldsymbol{\mu}_h^k). \tag{8}$$

A low threshold value T leads to few reliable samples but high dependability of its pseudo labels, and vice versa. To arrange learning tasks in the order of difficulty for helping the model gain sufficient
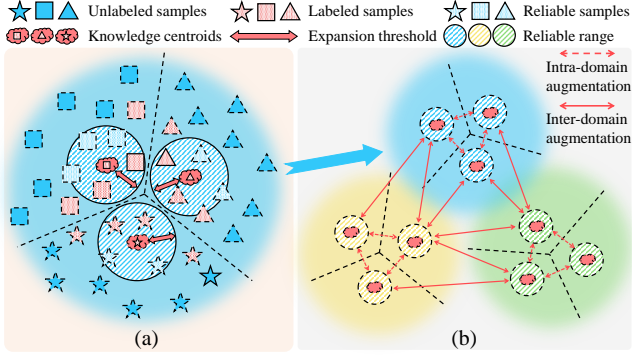
Figure 4: Semi-supervised generalization with knowledge augmentation. (a): Unlabeled samples within reliable range (scaled by expansion threshold), i.e., the reliable samples, are pseudo-labeled by the nearest knowledge centroids, i.e., the centers of the labeled data. (b): The reliable samples are then augmented in the same domain and across domains to expand domain knowledge and generalize domain invariance.

basic and easy knowledge before handling more complex data, we let T increase with epochs and dynamically tune learning difficulty:

$$T = T^{ini} + \frac{T^{fin} - T^{ini}}{E_{tot}} * E_{cur}, \qquad (9)$$

where T^{ini} and T^{fin} are the initial and final threshold values, E_{tot} and E_{cur} are the total and current epochs, respectively. It makes the model expand knowledge stably with the high dependable samples at the beginning, and break through the hard samples gradually.

We expand domain-class knowledge in each domain and across domains with the reliable dataset $\mathcal{U}$, and construct MixUp-based intra- and inter-domain knowledge augmentation datasets, i.e., $\mathcal{M}^{intra}$ and $\mathcal{M}^{inter}$, respectively. That is,

$$\mathcal{M}^{intra} = \{(\lambda x_i^{(u)} + (1-\lambda)x_j^{(u)}, \lambda y_i^{(u)} + (1-\lambda)y_j^{(u)}) | \\ (x_i^{(u)}, y_i^{(u)}) \in \mathcal{D}^m, (x_j^{(u)}, y_j^{(u)}) \in \mathcal{D}^n, m = n\}, \qquad (10)$$

$$\mathcal{M}^{inter} = \{(\lambda x_i^{(u)} + (1-\lambda)x_j^{(u)}, \lambda y_i^{(u)} + (1-\lambda)y_j^{(u)}) | \\ (x_i^{(u)}, y_i^{(u)}) \in \mathcal{D}^m, (x_j^{(u)}, y_j^{(u)}) \in \mathcal{D}^n, m \neq n\}, \qquad (11)$$

where $\lambda \sim Beta(\alpha, \alpha)$ with $\alpha = 0.2$ as in [67]. $\mathcal{M}^{intra}$ and $\mathcal{M}^{inter}$ open up the association among known regions within and across domains, respectively. To broaden the known regions and learn domain-invariant representations for improving out-of-domain generalization ability, we train the model on the union of the augmented datasets by optimizing an expansion and generalization loss $\mathcal{L}_{eg}$:

$$\mathcal{L}_{eg} = \mathbb{E}_{(x^{(u)}, y^{(u)}) \in \mathcal{M}^{intra} \cup \mathcal{M}^{inter}} \ell(G(x^{(u)}), y^{(u)}). \qquad (12)$$

We further utilize the unknown and known regions by adopting augmentation consistency constraint [50] for the unlabeled data and prediction supervision for the labeled data, respectively. Let $\mathcal{A}^w$ and $\mathcal{A}^s$ be weak (flip-and-shift) and strong [11] augmentation functions, respectively. Pseudo labels $\hat{q}$ can be assigned by $\mathcal{A}^w$ via $\hat{q} = \arg\max_h G_h(\mathcal{A}^w(x^{(u)}))$. The augmentation consistency loss

Algorithm 1 Collaborative Exploration and Generalization

Input: Source datasets $\{\mathcal{D}^k\}_{k=1}^K$; Pretraining epochs N_p and learning epochs N_l; Annotation budget B and initial budget B^{ini}.
Output: A well trained prediction model $\hat{G}$;
1: Initialize labeled dataset $\mathcal{D}^{(l)}$ with B^{ini};
2: **for** $n = 1$ to N_l **do**
3:     Get knowledge centroids $\boldsymbol{\mu}_h^k$ with $\mathcal{D}^{(l)}$ via Equation (4);
4:     **if** $n > N_p$ **then** // begin to query after pretraining
5:         Select a subset $\Delta\mathcal{D}^{(u)}$ from $\mathcal{D}^{(u)}$ via Equation (6);
6:         Query labels of $\Delta\mathcal{D}^{(u)}$ for $\Delta\mathcal{D}^{(u)} \to \Delta\mathcal{D}^{(l)}$;
7:         $\mathcal{D}^{(l)} \leftarrow \mathcal{D}^{(l)} \cup \Delta\mathcal{D}^{(l)}, \mathcal{D}^{(u)} \leftarrow \mathcal{D}^{(u)} / \Delta\mathcal{D}^{(u)}$;
8:     **end if**
9:     Train the discriminator H with $\mathcal{D}^{(u)}$ via Equation (2);
10:     Train the model G with $\mathcal{D}^{(l)}$ and $\mathcal{D}^{(u)}$ via Equation (15).
11: **end for**

$\mathcal{L}_{ac}$ makes the model predictions of strong augmented data, i.e., $G(\mathcal{A}^s(x^{(u)}))$, and weak augmented label, i.e., $\hat{q}$, to be consistent:

$$\mathcal{L}_{ac} = \mathbb{E}_{x^{(u)} \in \mathcal{D}^{(u)}} I(x^{(u)}) \ell(G(\mathcal{A}^s(x^{(u)})), \hat{q}), \qquad (13)$$

where an indicator $I(x^{(u)}) = \mathbb{1}(\max_h (G_h(\mathcal{A}^w(x^{(u)})) \geq \tau)$ (τ is set to 0.95 as in [50]) selects high dependable data. This constraint helps the model to capture structural knowledge in the unknown regions via unsupervised learning. For prediction supervision, we adopt a cross-entropy classification loss $\mathcal{L}_{ce}$ for the labeled data:

$$\mathcal{L}_{ce} = \mathbb{E}_{(x^{(l)}, y^{(l)}) \in \mathcal{D}^{(l)}} \ell(G(x^{(l)}), y^{(l)}). \qquad (14)$$

A semi-supervised training loss $\mathcal{L}_{ss}$ is then derived as:

$$\mathcal{L}_{ss} = \mathcal{L}_{ce} + \mathcal{L}_{ac} + \delta\mathcal{L}_{eg}, \qquad (15)$$

where δ is a hyper-parameter of knowledge expansion and generalization. We set the weights of $\mathcal{L}_{ce}$ and $\mathcal{L}_{ac}$ to 1 as in [50].

Our framework CEG explores informative unlabeled samples for learning key knowledge of multi-source distribution using limited annotation budget, promoting the expansion and generalization of the key knowledge in semi-supervised training. Then a well trained model is continuously utilized to select more effective samples in the next round of query. The active exploration and semi-supervised generalization are unified in a collaborative way by being repeated alternately. They complement and promote each other to enable label-efficient domain generalization. The learning process of CEG is stated in Algorithm 1. Note that we use an initial budget B^{ini} from the annotation budget B to initialize the labeled dataset $D^{(l)}$ via uniform sample selection, and pretrain the model before active query to solve a cold start problem, to our empirical experience.

4 EXPERIMENTS

In this section, we first evaluate our framework **CEG** in label-limited scenarios, and then give sensitivity analysis of hyper-parameters, ablation studies of the components, and in-depth empirical analysis.

Datasets. We adopt two popular pubilc datasets that are **PACS** [30] and **Office-Home** [54]. PACS contains 7 categories within 4 domains, i.e., *Art*, *Cartoon*, *Sketch*, and *Photo*. Office-Home has 65 classes in 4 domains, i.e., *Art*, *Clipart*, *Product*, and *Real-World*.

Table 1: Performance (%) comparisons between CEG and DG methods on PACS dataset with 5% annotation budget. The results with fully labeled source data (100% annotation) are given in parentheses. The best results are emphasized in bold.

Methods	Art	Cartoon	Photo	Sketch	Average
DeepAll	$55.79_{\pm2.92}$ ($73.59_{\pm2.89}$)	$61.84_{\pm1.98}$ ($70.63_{\pm2.33}$)	$80.21_{\pm1.14}$ ($89.36_{\pm1.21}$)	$63.00_{\pm1.60}$ ($80.06_{\pm0.95}$)	$65.21_{\pm0.46}$ ($78.41_{\pm0.44}$)
JiGen [6]	$53.03_{\pm5.19}$ ($77.05_{\pm1.83}$)	$55.09_{\pm1.85}$ ($76.16_{\pm2.02}$)	$78.62_{\pm5.83}$ ($93.81_{\pm1.51}$)	$22.28_{\pm4.05}$ ($70.93_{\pm0.72}$)	$52.26_{\pm0.25}$ ($79.49_{\pm0.67}$)
FACT [61]	$74.21_{\pm0.30}$ ($84.76_{\pm0.77}$)	$65.82_{\pm2.03}$ ($77.52_{\pm0.70}$)	$90.48_{\pm1.17}$ ($95.29_{\pm0.31}$)	$55.24_{\pm3.62}$ ($78.97_{\pm0.32}$)	$71.44_{\pm1.02}$ ($84.13_{\pm0.24}$)
DDAIG [71]	$62.87_{\pm3.13}$ ($77.80_{\pm1.09}$)	$57.64_{\pm6.59}$ ($75.35_{\pm3.21}$)	$81.48_{\pm3.82}$ ($89.66_{\pm1.74}$)	$36.61_{\pm2.23}$ ($73.70_{\pm2.99}$)	$59.65_{\pm2.22}$ ($79.13_{\pm0.91}$)
RSC [21]	$59.57_{\pm3.37}$ ($77.88_{\pm0.66}$)	$59.61_{\pm3.22}$ ($73.90_{\pm2.12}$)	$84.43_{\pm3.59}$ ($93.85_{\pm0.80}$)	$57.38_{\pm6.61}$ ($80.66_{\pm0.81}$)	$65.25_{\pm2.95}$ ($81.57_{\pm0.71}$)
CrossGrad [46]	$56.06_{\pm7.04}$ ($75.69_{\pm2.25}$)	$52.73_{\pm4.17}$ ($76.51_{\pm3.24}$)	$80.51_{\pm1.97}$ ($91.33_{\pm0.50}$)	$41.25_{\pm5.21}$ ($70.50_{\pm0.97}$)	$57.64_{\pm2.13}$ ($78.51_{\pm1.40}$)
DAEL [72]	$66.24_{\pm1.86}$ ($83.51_{\pm0.83}$)	$61.72_{\pm1.89}$ ($72.31_{\pm2.67}$)	$89.98_{\pm0.37}$ ($95.74_{\pm0.08}$)	$32.50_{\pm2.20}$ ($78.87_{\pm0.59}$)	$62.61_{\pm0.46}$ ($82.61_{\pm0.98}$)
CEG (ours)	$\mathbf{80.12}_{\pm0.37}$	$\mathbf{71.11}_{\pm0.96}$	$\mathbf{92.32}_{\pm1.68}$	$\mathbf{73.13}_{\pm2.87}$	$\mathbf{79.17}_{\pm0.83}$

Table 2: Performance (%) comparisons between CEG and DG methods on Office-Home dataset with 5% annotation budget. The results with fully labeled source data (100% annotation) are given in parentheses. The best results are emphasized in bold.

Methods	Art	Clipart	Product	Real-World	Average
DeepAll	$34.73_{\pm1.13}$ ($47.06_{\pm1.35}$)	$34.46_{\pm2.11}$ ($47.50_{\pm0.91}$)	$46.20_{\pm1.51}$ ($64.89_{\pm0.65}$)	$48.89_{\pm0.87}$ ($65.16_{\pm0.62}$)	$41.07_{\pm0.93}$ ($56.15_{\pm0.59}$)
JiGen [6]	$29.62_{\pm2.25}$ ($52.67_{\pm0.95}$)	$25.52_{\pm2.12}$ ($50.40_{\pm0.97}$)	$37.91_{\pm1.33}$ ($71.21_{\pm0.12}$)	$39.84_{\pm0.61}$ ($72.24_{\pm0.15}$)	$33.22_{\pm0.93}$ ($61.63_{\pm0.25}$)
FACT [61]	$40.71_{\pm0.08}$ ($58.98_{\pm0.29}$)	$32.12_{\pm0.17}$ ($53.53_{\pm0.35}$)	$48.05_{\pm0.14}$ ($74.47_{\pm0.56}$)	$49.16_{\pm0.17}$ ($75.63_{\pm0.67}$)	$42.51_{\pm0.09}$ ($65.65_{\pm0.41}$)
DDAIG [71]	$35.20_{\pm1.06}$ ($55.05_{\pm0.69}$)	$29.75_{\pm0.50}$ ($52.37_{\pm0.58}$)	$42.42_{\pm0.58}$ ($72.00_{\pm0.58}$)	$43.07_{\pm0.12}$ ($73.54_{\pm0.19}$)	$37.61_{\pm0.16}$ ($63.24_{\pm0.35}$)
RSC [21]	$31.95_{\pm1.24}$ ($56.06_{\pm0.71}$)	$28.62_{\pm1.53}$ ($52.95_{\pm0.31}$)	$40.88_{\pm1.87}$ ($72.61_{\pm0.39}$)	$42.43_{\pm0.69}$ ($73.42_{\pm0.38}$)	$35.97_{\pm0.61}$ ($63.76_{\pm0.25}$)
CrossGrad [46]	$35.05_{\pm0.37}$ ($54.42_{\pm0.55}$)	$30.86_{\pm1.74}$ ($52.63_{\pm0.77}$)	$45.10_{\pm1.76}$ ($73.00_{\pm0.47}$)	$44.41_{\pm2.08}$ ($73.42_{\pm0.74}$)	$38.86_{\pm0.27}$ ($63.37_{\pm0.24}$)
DAEL [72]	$35.93_{\pm0.57}$ ($59.20_{\pm0.56}$)	$30.71_{\pm0.86}$ ($50.97_{\pm2.63}$)	$42.79_{\pm0.99}$ ($73.53_{\pm0.52}$)	$43.95_{\pm0.86}$ ($76.56_{\pm0.45}$)	$38.35_{\pm0.12}$ ($65.06_{\pm0.55}$)
CEG (ours)	$\mathbf{47.60}_{\pm1.32}$	$\mathbf{42.01}_{\pm1.19}$	$\mathbf{56.20}_{\pm1.79}$	$\mathbf{57.69}_{\pm1.18}$	$\mathbf{50.87}_{\pm0.99}$

Table 3: Comparisons between CEG and AL, SSL, SSDG methods on PACS and Office-Home datasets with 5% annotation budget.

Methods	PACS dataset (%)					Office-Home dataset (%)				
	Art	Cartoon	Photo	Sketch	Average	Art	Clipart	Product	Real-World	Average
Uniform	$55.56_{\pm2.92}$	$61.61_{\pm1.98}$	$79.98_{\pm1.14}$	$62.77_{\pm1.60}$	$64.98_{\pm0.46}$	$34.27_{\pm1.13}$	$34.00_{\pm2.11}$	$45.74_{\pm1.51}$	$48.43_{\pm0.87}$	$40.61_{\pm0.93}$
Entropy [57]	$58.79_{\pm2.36}$	$63.49_{\pm1.78}$	$82.47_{\pm0.99}$	$61.67_{\pm0.82}$	$66.61_{\pm0.58}$	$34.06_{\pm1.59}$	$32.02_{\pm1.84}$	$46.72_{\pm0.99}$	$47.03_{\pm0.96}$	$39.96_{\pm0.90}$
BvSB [24]	$62.85_{\pm1.83}$	$63.17_{\pm1.02}$	$79.57_{\pm3.46}$	$63.61_{\pm5.97}$	$67.30_{\pm0.97}$	$35.58_{\pm1.44}$	$35.32_{\pm2.29}$	$47.66_{\pm1.42}$	$50.39_{\pm0.54}$	$42.24_{\pm0.82}$
Confidence [57]	$58.02_{\pm2.14}$	$59.48_{\pm1.61}$	$81.75_{\pm3.40}$	$61.04_{\pm2.86}$	$65.07_{\pm1.60}$	$36.35_{\pm1.23}$	$36.21_{\pm1.24}$	$47.88_{\pm0.75}$	$50.46_{\pm2.20}$	$42.73_{\pm0.83}$
CoreSet [45]	$61.48_{\pm4.57}$	$58.74_{\pm2.66}$	$79.03_{\pm3.71}$	$60.61_{\pm2.25}$	$64.96_{\pm1.06}$	$37.54_{\pm0.76}$	$35.75_{\pm2.55}$	$49.44_{\pm1.21}$	$51.06_{\pm1.75}$	$43.45_{\pm0.66}$
BADGE [1]	$54.49_{\pm1.67}$	$63.10_{\pm1.60}$	$80.84_{\pm1.19}$	$65.57_{\pm5.99}$	$66.50_{\pm2.11}$	$37.81_{\pm1.07}$	$36.86_{\pm2.62}$	$49.90_{\pm2.00}$	$51.26_{\pm2.77}$	$43.96_{\pm0.82}$
MeanTeacher [52]	$53.84_{\pm6.41}$	$54.86_{\pm4.14}$	$78.86_{\pm4.63}$	$35.52_{\pm4.57}$	$55.77_{\pm1.43}$	$32.70_{\pm1.56}$	$27.25_{\pm2.92}$	$43.01_{\pm2.04}$	$42.41_{\pm4.03}$	$36.35_{\pm1.25}$
MixMatch [4]	$63.92_{\pm1.77}$	$61.37_{\pm2.31}$	$81.14_{\pm4.12}$	$55.46_{\pm0.61}$	$65.47_{\pm1.70}$	$25.65_{\pm0.66}$	$22.90_{\pm2.24}$	$33.80_{\pm0.93}$	$28.35_{\pm2.24}$	$27.68_{\pm0.80}$
FixMatch [50]	$78.60_{\pm1.47}$	$71.14_{\pm2.49}$	$92.17_{\pm1.02}$	$69.16_{\pm0.94}$	$77.77_{\pm0.91}$	$36.76_{\pm1.84}$	$31.09_{\pm2.53}$	$44.79_{\pm4.20}$	$45.07_{\pm5.18}$	$39.43_{\pm2.21}$
StyleMatch [70]	$72.67_{\pm1.08}$	$\mathbf{73.07}_{\pm0.81}$	$89.61_{\pm0.74}$	$\mathbf{76.46}_{\pm0.93}$	$77.95_{\pm0.56}$	$42.01_{\pm0.68}$	$40.95_{\pm0.97}$	$47.65_{\pm1.70}$	$51.93_{\pm0.26}$	$45.63_{\pm0.29}$
CEG (ours)	$\mathbf{80.12}_{\pm0.37}$	$71.11_{\pm0.96}$	$\mathbf{92.32}_{\pm1.68}$	$73.13_{\pm2.87}$	$\mathbf{79.17}_{\pm0.83}$	$\mathbf{47.60}_{\pm1.32}$	$\mathbf{42.01}_{\pm1.19}$	$\mathbf{56.20}_{\pm1.79}$	$\mathbf{57.69}_{\pm1.18}$	$\mathbf{50.87}_{\pm0.99}$

Baseline methods. We implement four types of baselines. (1) Domain generalization (**DG**): DeepAll (training with mixed multi-source data), JiGen [6], CrossGrad [46], DDAIG [71], DAEL [72], RSC [21], and FACT [61]. (2) Active learning (**AL**): Uniform (uniform selection), Entropy [57], BvSB [24], Confidence [57], CoreSet [45], and BADGE [1]. (3) Semi-supervised learning (**SSL**): MeanTeacher [52], MixMatch [4], and FixMatch [50]. (4) Semi-supervised domain generalization (**SSDG**): StyleMatch [70]. See Section 2 for details.

Implementation details. Following [6, 21, 61], we use a pre-trained ResNet-18 [17] as the backbone and conduct leave-one-domain-out experiments by choosing one domain to hold out as the target domain. For fair comparisons, we implement all the methods with the same settings, i.e., SGD optimizer with learning rate 0.003 for feature extractor and 0.01 for classifier, pre-training/learning epochs on PACS and Office-Home datasets are 30/30 and 15/15, respectively, and batch-size is 16, et al. In experiments, we directly use "T" to represent "T^{fin}" in Equation (9) for simplicity. We adopt
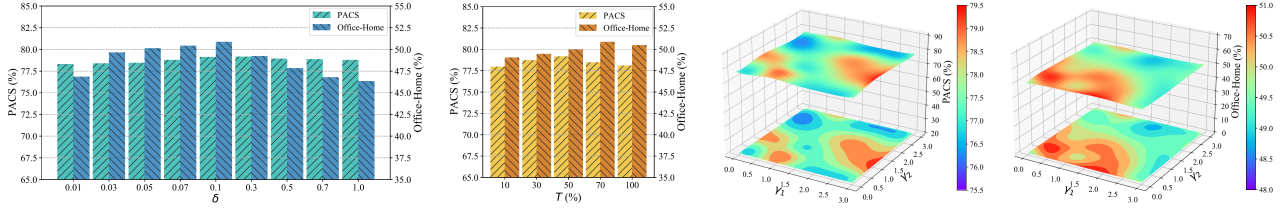
Figure 5: Sensitivity analysis of hyper-parameters. From left to right: δ, T, γ_1 and γ_2 on PACS, γ_1 and γ_2 on Office-Home.

Table 4: Ablation studies on PACS and Office-Home datasets (5% annotation). The best results are emphasized in bold.

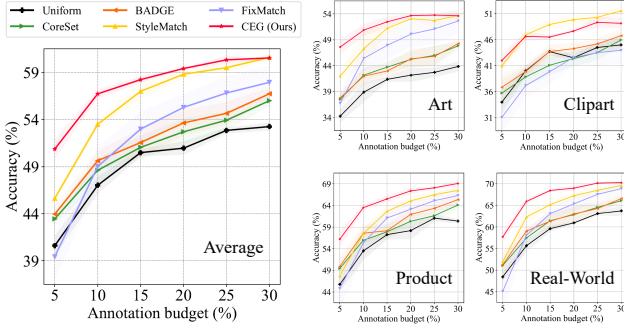| Strategies	Cases	PACS dataset (%)					Office-Home dataset (%)				
		Art	Cartoon	Photo	Sketch	Average	Art	Clipart	Product	Real-World	Average
Active Exploration	w/ Uniform	$76.42_{\pm1.27}$	$69.92_{\pm3.01}$	$87.09_{\pm2.03}$	$71.92_{\pm4.97}$	$76.34_{\pm2.05}$	$45.25_{\pm1.29}$	$40.48_{\pm2.18}$	$53.48_{\pm2.00}$	$55.95_{\pm1.08}$	$48.79_{\pm0.96}$
	w/o S_u	$80.00_{\pm2.08}$	$67.56_{\pm2.34}$	$89.43_{\pm0.71}$	$71.11_{\pm5.64}$	$77.03_{\pm1.28}$	$46.12_{\pm2.04}$	$40.84_{\pm0.98}$	$52.56_{\pm0.75}$	$55.83_{\pm1.47}$	$48.84_{\pm0.84}$
	w/o S_r	$76.77_{\pm2.74}$	$67.91_{\pm4.38}$	$90.90_{\pm1.19}$	$72.46_{\pm1.44}$	$77.01_{\pm0.74}$	$46.14_{\pm1.05}$	$39.59_{\pm1.25}$	$55.99_{\pm2.06}$	$57.02_{\pm2.06}$	$49.68_{\pm0.71}$
	w/o S_d	$78.21_{\pm0.97}$	$68.95_{\pm2.54}$	$91.03_{\pm1.50}$	$72.58_{\pm3.24}$	$77.69_{\pm0.61}$	$45.78_{\pm2.34}$	$40.79_{\pm1.78}$	$54.39_{\pm1.46}$	$\mathbf{58.24_{\pm1.83}}$	$49.84_{\pm0.82}$
Semi-Supervised Generalization	w/o $\mathcal{L}_{ac}$ w/o $\mathcal{L}_{eg}$	$62.61_{\pm2.81}$	$68.51_{\pm2.34}$	$82.84_{\pm3.85}$	$48.65_{\pm4.32}$	$65.65_{\pm1.35}$	$38.51_{\pm1.06}$	$33.62_{\pm1.50}$	$48.60_{\pm2.59}$	$49.89_{\pm5.07}$	$42.66_{\pm1.44}$
	w/o $\mathcal{L}_{ac}$	$76.16_{\pm2.11}$	$67.05_{\pm3.37}$	$88.07_{\pm1.80}$	$60.37_{\pm5.36}$	$72.91_{\pm1.31}$	$44.65_{\pm1.69}$	$37.99_{\pm1.25}$	$55.41_{\pm2.16}$	$56.42_{\pm2.09}$	$48.62_{\pm0.59}$
	w/o $\mathcal{L}_{eg}$	$72.13_{\pm1.43}$	$68.09_{\pm3.37}$	$86.03_{\pm4.02}$	$75.12_{\pm2.31}$	$75.34_{\pm0.84}$	$40.38_{\pm2.15}$	$34.68_{\pm1.29}$	$46.10_{\pm2.20}$	$47.94_{\pm1.66}$	$42.27_{\pm1.18}$
	w/o $\mathcal{L}_{eg}$ ($\mathcal{M}^{intra}$)	$76.52_{\pm2.01}$	$67.69_{\pm3.50}$	$89.89_{\pm1.72}$	$74.72_{\pm2.88}$	$77.20_{\pm1.53}$	$45.85_{\pm2.72}$	$38.50_{\pm1.47}$	$54.12_{\pm2.65}$	$55.80_{\pm1.66}$	$48.57_{\pm1.36}$
	w/o $\mathcal{L}_{eg}$ ($\mathcal{M}^{inter}$)	$75.36_{\pm2.59}$	$68.20_{\pm2.66}$	$91.18_{\pm1.05}$	$72.49_{\pm2.28}$	$76.81_{\pm1.32}$	$47.47_{\pm3.43}$	$38.97_{\pm2.26}$	$52.99_{\pm2.77}$	$55.10_{\pm1.48}$	$48.63_{\pm1.32}$
	w/ static T	$76.86_{\pm2.37}$	$69.61_{\pm1.63}$	$90.79_{\pm1.13}$	$73.96_{\pm2.06}$	$77.81_{\pm1.09}$	$45.90_{\pm1.76}$	$40.46_{\pm2.33}$	$55.90_{\pm1.50}$	$56.21_{\pm2.30}$	$49.62_{\pm1.02}$
CEG		$\mathbf{80.12_{\pm0.37}}$	$\mathbf{71.11_{\pm0.96}}$	$\mathbf{92.32_{\pm1.68}}$	$73.13_{\pm2.87}$	$\mathbf{79.17_{\pm0.83}}$	$\mathbf{47.60_{\pm1.32}}$	$\mathbf{42.01_{\pm1.19}}$	$\mathbf{56.20_{\pm1.79}}$	$57.69_{\pm1.18}$	$\mathbf{50.87_{\pm0.99}}$

Figure 6: Comparisons with increasing annotation budget.

the percentage of the unlabeled samples for T instead of a distance value. We set $T^{ini} = \frac{T}{2}$. The hyper-parameters $\{\delta, T, \gamma_1, \gamma_2\}$ are set to $\{0.3, 50\%, 3, 1\}$ and $\{0.1, 70\%, 0.5, 0.5\}$ for PACS and Office-Home, respectively. Half annotation budget is used as the initial budget to initialize the labeled dataset. We report the results over five runs.

4.1 Main Results of CEG

CEG vs DG methods. Table 1 and 2 report the results with 5% annotation budget on PACS and Office-Home datasets, respectively. We observe that the accuracy of DG methods drops rapidly when only 5% labeled data is given. In comparison, our method CEG can select the informative data to label and utilize both the labeled and unlabeled data to boost generalization performance in this challenging label-limited scenario. Most notably, CEG can even achieve competitive results with only 5% annotation budget compared to the DG methods with full annotation on the PACS dataset.

It reveals that CEG generally realizes label-efficient domain generalization by exploiting only a small quota of labeled data and massive unlabeled data. We attribute this success to the effective collaboration mechanism between active exploration and semi-supervised generalization, which unleashes the latent power of the limited annotation budget. Since the DG methods may not be good at tackling the label-limited task as they can only use the labeled data, we further compare our CEG method with AL, SSL, and SSDG methods.

CEG vs AL, SSL, SSDG methods. Table 3 reports the results with 5% annotation budget on PACS and Office-Home datasets. CEG outperforms other methods on half of the tasks and yields the best average accuracy on the PACS dataset. It is probably because the AL and SSL methods rely on the i.i.d. assumption, and the SSDG method does not label and exploit the important source data. In contrast, CEG selects the most informative samples for query via active exploration, and hence captures multi-source distribution and boosts generalization ability more accurately. Besides, the performance of CEG is significantly better than other methods on the Office-Home dataset. We attribute it to the construction of domain-class knowledge centroids, which greatly helps CEG to precisely explore unknown regions during active exploration, and effectively expand knowledge and generalize domain invariance during semi-supervised generalization on the Office-Home dataset (because Office-Home has 65 classes but PACS only has 7 classes).

4.2 Sensitivity Analysis

As shown in Figure 5, CEG is generally robust to the hyper-parameters and outperforms other methods even with the default settings, i.e., $\delta = 1.0$ (78.79% on PACS and 46.34% on Office-Home), $T = 100\%$ (78.09% on PACS and 50.49% on Office-Home), $\gamma_1 = \gamma_2 = 1.0$ (78.12%
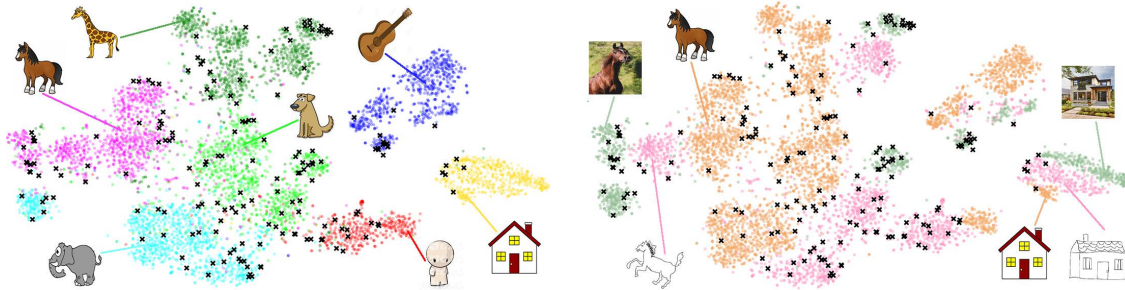
Figure 7: T-SNE visualization [40] of the selected samples. Different colors represent different classes (left) and domains (right). Each dot is the embedding of a sample and the black crosses are the selected samples. Dataset: PACS; Target domain: Art.
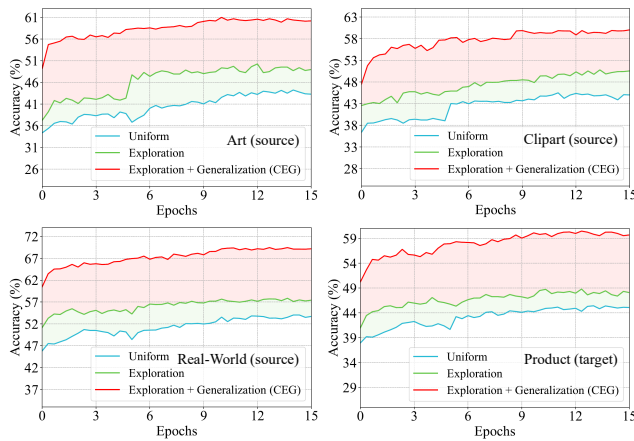

Figure 8: Accuracy curve of w/ uniform, w/ active exploration, w/ active exploration + semi-supervised generalization (CEG).

on PACS and 50.37% on Office-Home), indicating that exhaustive hyper-parameter fine-tuning is not necessary for CEG to achieve excellent performance in label-efficient generalization learning.

4.3 Results with Increasing Annotation Budget

Figure 6 shows the results with increasing annotation on Office-Home dataset. CEG consistently outperforms other methods by sharp margins on the average accuracy and three of the four tasks. The significant performance achieved by CEG when given a low budget is probably due to the query based-active exploration, but this advantage could be weakened when given a higher budget.

4.4 Why does CEG Work?

Ablation studies are reported in Table 4. The three criteria of active exploration, i.e., uncertainty S_u, representativeness S_r, and diversity S_d, are all important for learning multi-source distribution, and the integration of them further make full use of the limited annotation, compared with uniform selection. For semi-supervised generalization, both knowledge expansion and generalization $\mathcal{L}_{eg}$ and augmentation consistency $\mathcal{L}_{ac}$ are necessary to yield remarkable results. The intra- and inter-domain knowledge augmentation datasets, i.e., $\mathcal{M}^{intra}$ and $\mathcal{M}^{inter}$, both play vital roles in improving generalization performance. It is noteworthy that the proposed $\mathcal{L}_{eg}$

significantly improves average accuracy from 42.27% to 50.87% on Office-Home. Besides, the devised dynamic threshold T shows its effectiveness of learning with increasing difficulty compared to the static one. The above results illustrate that each component is indispensable, and the exploration and generalization complement and promote each other for achieving the excellent performance.

T-SNE visualization is shown in Figure 7. The left figure shows class-ambiguous samples, i.e., the samples distribute on class boundary, are selected for learning class discriminability. The right figure shows that, in general, the selected samples distribute uniformly and representatively in each domain, illustrating the effectiveness of the domain representativeness and information diversity criteria. These three criteria help CEG to select the most informative samples for learning multi-source distribution, which facilitates the generalizable model training in semi-supervised generalization.

Accuracy curve on Office-Home dataset is shown in Figure 8. Active exploration selects the most important samples and grasps the key knowledge of multi-source distribution to effectively improve performance on each domain, compared with uniform sample selection. Semi-supervised generalization further markedly boosts performance by expanding the obtained knowledge and generalizing domain invariance. They promote each other to achieve remarkable generalization performance on the target domain.

5 CONCLUSION

We introduce a practical task named label-efficient domain generalization, and propose a novel method called CEG for this task via active exploration and semi-supervised generalization. The two modules promote each other to improve model generalization with the limited annotation. In future work, we may extend our method to a more challenging setting that domain labels are unknown.

REFERENCES

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems* 31 (2018), 998–1008.

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

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

[5] Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems* 24 (2011), 2178–2186.

[6] Fabio Maria Carlucci, Antonio D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. 2019. Domain Generalization by Solving Jigsaw Puzzles. *IEEE Conference on Computer Vision and Pattern Recognition* (2019), 2224–2233.

[7] Yang Chen, Yingwei Pan, Yu Wang, Ting Yao, Xinmei Tian, and Tao Mei. 2021. Transferrable Contrastive Learning for Visual Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3399–3408.

[8] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. 2021. Pareto Self-Supervised Training for Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13663–13672.

[9] Zhengyu Chen and Donglin Wang. 2021. Multi-Initialization Meta-Learning with Domain Adaptation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1390–1394.

[10] Zhengyu Chen, Teng Xiao, and Kun Kuang. 2022. BA-GNN: On Learning Bias-Aware Graph Neural Network. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE.

[11] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation strategies from data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 113–123.

[12] Wanxia Deng, Yawen Cui, Zhen Liu, Gangyao Kuang, Dewen Hu, Matti Pietikäinen, and Li Liu. 2021. Informative Class-Conditioned Feature Alignment for Unsupervised Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1303–1312.

[13] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems* 32 (2019), 6450–6461.

[14] Zhekai Du, Jingjing Li, Ke Lu, Lei Zhu, and Zi Huang. 2021. Learning Transferrable and Interpretable Representations for Domain Generalization. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3340–3349.

[15] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. 2021. Adaptive Methods for Real-World Domain Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14340–14349.

[16] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9710–9719.

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

[18] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. 2021. Fsdr: Frequency space domain randomization for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6891–6902.

[19] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. 2021. Semi-supervised active learning with temporal output discrepancy. In *IEEE/CVF International Conference on Computer Vision*. 3447–3456.

[20] Shengqi Huang, Wanqi Yang, Lei Wang, Luping Zhou, and Ming Yang. 2021. Few-shot Unsupervised Domain Adaptation with Image-to-class Sparse Similarity Encoding. In *Proceedings of the 29th ACM International Conference on Multimedia*. 677–685.

[21] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. 2020. Self-challenging Improves Cross-Domain Generalization. In *European Conference on Computer Vision*. 124–140.

[22] Seogkyu Jeon, Kibeom Hong, Pilhyeon Lee, Jewook Lee, and Hyeran Byun. 2021. Feature stylization and domain-aware contrastive learning for domain generalization. In *Proceedings of the 29th ACM International Conference on Multimedia*. 22–31.

[23] Ziqi Jiang, Shengyu Zhang, Siyuan Yao, Wenqiao Zhang, Sihan Zhang, Juncheng Li, Zhou Zhao, and Fei Wu. 2022. Weakly-supervised Disentanglement Network for Video Fingerspelling Detection. In *ACM MM*.

[24] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2372–2379.

[25] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. 2021. Task-aware variational adversarial active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8166–8175.

[26] Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. 2018. Stable prediction across unknown environments. In *proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1617–1626.

[27] Kun Kuang, Haotian Wang, Yue Liu, Ruoxuan Xiong, Runze Wu, Weiming Lu, Yue Ting Zhuang, Fei Wu, Peng Cui, and Bo Li. 2022. Stable Prediction with Leveraging Seed Variable. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[28] Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. 2020. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4485–4492.

[29] Kun Kuang, Hengtao Zhang, Runze Wu, Fei Wu, Yueting Zhuang, and Aijun Zhang. 2021. Balance-Subsampled stable prediction across unknown test data. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 3 (2021), 1–21.

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

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*.

[32] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5400–5409.

[33] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. 2021. Progressive Domain Expansion Network for Single Domain Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 224–233.

[34] Xinhao Li, Jingjing Li, Lei Zhu, Guoqing Wang, and Zi Huang. 2021. Imbalanced Source-free Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3330–3339.

[35] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. 2019. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*. PMLR, 3915–3924.

[36] Yixiao Liao, Ruyi Huang, Jipu Li, Zhuyun Chen, and Weihua Li. 2020. Deep semisupervised domain generalization network for rotary machinery fault diagnosis under variable speed. *IEEE Transactions on Instrumentation and Measurement* 69, 10 (2020), 8064–8075.

[37] Chang Liu, Lichen Wang, Kai Li, and Yun Fu. 2021. Domain Generalization via Feature Variation Decorrelation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1683–1691.

[38] Jianming Lv, Kaijie Liu, and Shengfeng He. 2021. Differentiated Learning for Multi-Modal Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1322–1330.

[39] Xu Ma, Junkun Yuan, Yen-wei Chen, Ruofeng Tong, and Lanfen Lin. 2022. Attention-based cross-layer domain alignment for unsupervised domain adaptation. *Neurocomputing* 499 (2022), 1–10.

[40] L. V. D. Maaten and Geoffrey E. Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[41] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *International Conference on Machine Learning*. PMLR, 7313–7324.

[42] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh AP. 2021. Generalization on Unseen Domains via Inference-Time Label-Preserving Target Projections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12924–12933.

[43] Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12556–12565.

[44] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press.

[45] Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

[46] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing across domains via cross-gradient training. *International Conference on Learning Representation* (2018).

[47] Hossein Sharifi-Noghabi, Hossein Asghari, Nazanin Mehrasa, and Martin Ester. 2020. Domain generalization via semi-supervised meta learning. *arXiv preprint arXiv:2009.12658* (2020).

[48] Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kunag. 2020. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5692–5699.

[49] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. 2021. Open Domain Generalization with Domain-Augmented Meta-Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9624–9633.

[50] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin Dogus Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. 2020. Fix-Match: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems* 33 (2020).

[51] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. 2021. Bottleneck transformers for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16519–16529.

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

[53] Vladimir Vapnik. 1992. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*. 831–838.

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

[55] Riccardo Volpi, Diane Larlus, and Grégory Rogez. 2021. Continual Adaptation of Visual Representations via Domain Randomization and Meta-learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4443–4453.

[56] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Efty-chios Protopapadakis. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018 (2018).

[57] Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. In *International joint conference on neural networks*. IEEE, 112–119.

[58] Mengzhu Wang, Wei Wang, Baopu Li, Xiang Zhang, Long Lan, Huibin Tan, Tianyi Liang, Wei Yu, and Zhigang Luo. 2021. InterBN: Channel Fusion for Adversarial Unsupervised Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3691–3700.

[59] Ruiqi Wang, Lei Qi, Yinghuan Shi, and Yang Gao. 2021. Better Pseudo-label: Joint Domain-aware Label and Dual-classifier for Semi-supervised Domain Generalization. *arXiv preprint arXiv:2110.04820* (2021).

[60] Yufei Wang, Haoliang Li, Lap-pui Chau, and Alex C Kot. 2021. Embracing the Dark Knowledge: Domain Generalization Using Regularized Knowledge Distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2595–2604.

[61] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A Fourier-based Framework for Domain Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14383–14392.

[62] Zizheng Yan, Xianggang Yu, Yipeng Qin, Yushuang Wu, Xiaoguang Han, and Shuguang Cui. 2021. Pixel-level intra-domain adaptation for semantic segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 404–413.

[63] Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. 2021. Source Data-free Unsupervised Domain Adaptation for Semantic Segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2233–2242.

[64] Junkun Yuan, Xu Ma, Defang Chen, Kun Kuang, Fei Wu, and Lanfen Lin. 2021. Collaborative Semantic Aggregation and Calibration for Separated Domain Generalization. *arXiv e-prints* (2021), arXiv–2110.

[65] Junkun Yuan, Xu Ma, Defang Chen, Kun Kuang, Fei Wu, and Lanfen Lin. 2021. Domain-Specific Bias Filtering for Single Labeled Domain Generalization. *arXiv preprint arXiv:2110.00726* (2021).

[66] Junkun Yuan, Xu Ma, Kun Kuang, Ruoxuan Xiong, Mingming Gong, and Lanfen Lin. 2021. Learning domain-invariant relationship with instrumental variable for domain generalization. *arXiv preprint arXiv:2110.01438* (2021).

[67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. *International Conference on Learning Representation* (2018).

[68] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. 2021. Deep Stable Learning for Out-Of-Distribution Generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5372–5382.

[69] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. 2020. Domain generalization via entropy regularization. *Advances in Neural Information Processing Systems* 33 (2020).

[70] Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. 2021. Semi-Supervised Domain Generalization with Stochastic StyleMatch. *arXiv preprint arXiv:2106.00592* (2021).

[71] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Deep domain-adversarial image generation for domain generalisation. In *AAAI Conference on Artificial Intelligence*.

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

[73] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain Generalization with Mixstyle. In *International Conference on Learning Representation*.