Learning to Diversify for Single Domain Generalization

Zijian Wang Yadan Luo Ruihong Qiu Zi Huang Mahsa Baktashmotlagh
University of Queesland
{zijian.wang, y.luo, r.qiu, m.baktashmotlagh}@uq.edu.au, huang@itee.uq.edu.au

Abstract

Domain generalization (DG) aims to generalize a model trained on multiple source (i.e., training) domains to a distributionally different target (i.e., test) domain. In contrast to the conventional DG that strictly requires the availability of multiple source domains, this paper considers a more realistic yet challenging scenario, namely Single Domain Generalization (Single-DG), where only one source domain is available for training. In this scenario, the limited diversity may jeopardize the model generalization on unseen target domains. To tackle this problem, we propose a style-complement module to enhance the generalization power of the model by synthesizing images from diverse distributions that are complementary to the source ones. More specifically, we adopt a tractable upper bound of mutual information (MI) between the generated and source samples and perform a two-step optimization iteratively: (1) by minimizing the MI upper bound approximation for each sample pair, the generated images are forced to be diversified from the source samples; (2) subsequently, we maximize the MI between the samples from the same semantic category, which assists the network to learn discriminative features from diverse-styled images. Extensive experiments on three benchmark datasets demonstrate the superiority of our approach, which surpasses the state-of-the-art single-DG methods by up to 25.14%. *The code will be publicly available at https://github.com/BUserName/Learning_to_diversify*

1. Introduction

The remarkable success of modern machine learning algorithms is built on the assumption that the source (*i.e.*, training) and target (*i.e.*, test) samples are drawn from similar distributions. In practice, this assumption is commonly violated by various factors, such as the changes of illuminations, object appearance, or background, which are known as the *domain shift* problem [36, 3]. Due to the discrepancy across domains, the performance of a model trained on the source domain can be significantly degraded when applied to the target domain.

To tackle this problem, extensive research has been carried out mainly on domain adaptation and domain generalization. Domain adaptation aims to transfer the knowledge from the labeled source domain(s) to an unlabeled target domain [45, 1, 26, 13], while domain generalization attempts to generalize a model to an unseen target domain by learning from multiple source domains [34, 42, 10, 42]. Compared with domain adaptation, domain generalization (DG) is considered as a more challenging task as the target samples are not exposed in the training phase. Regarding different strategies for transferring knowledge from source domains to the unseen target domain, existing DG techniques can be subsumed under two broad categories, *i.e.*, *alignment-based* [2, 11, 33] and *augmentation-based* [5, 17, 44, 50, 47]. Technically, alignment-based approaches aim to reach the consensus from multiple source domains and learn domain-invariant latent representations for the target domain. Augmentation-based approaches learn to augment the source images with different transformations or generate the pseudo-novel samples for each source domain.

In general, the paradigm of DG relies on the availability of multiple source domains. However, it is more plausible to consider a more realistic scenario namely single domain generalization (Single-DG), where *only one* source domain is at hand during training. Despite DG has been extensively studied, single-DG remains under-explored. It is nontrivial for prior DG methods to handle this new setting, as the samples gathered from multiple sources and the domain identifiers are no longer accessible. Without the domain information to rely on, neither alignment-based nor augmentation-based models could well identify the domain-invariant features or transformations that are robust to unseen target domain shifts. Very recently, a few works [37, 49] are proposed that learn to add adversarial noises on the source images in order to train robust classifiers against unforeseen data shifts. While contributing positively to address the model vulnerability, the manipulated images are indistinguishable from the original source images, which are not sufficiently diverse to cover the real target distributions.

To address the issue of insufficient diversity, in this paper, we propose a novel approach of Learning-to-diversify (L2D), which aims to improve the model generalization capacity by alternating diverse sample generation and discriminative style-invariant representation learning as illustrated in Figure 1. Specifically, we design a style-complement module, which learns to synthesize samples with unseen styles, which are out of original distributions. Different from previous augmentation approaches that quantify diversity with the Euclidean distance in the image space, we diversify the generated samples in the latent feature space. By explicitly posing a greater challenge on the trained classifier, the model resilience against the target shift is enhanced. We gradually enlarge the shift between the distributions of the generated samples and the source samples at the training time and perform the two-step optimization iteratively. By minimizing the tractable upper bound of mutual information for each sample pair, the generated images are forced to diversify from the source samples in subspace. Furthermore, to obtain the style-invariant features, we maximize the mutual information between the images belonging to the same semantic category. Consequently, the style-complement module and the task model compete in a min-max game, which improves the generalization ability of the task model by iteratively generating out-of-domain images and optimizing the style-invariant latent space. Note that, under this objective, the images generated by the proposed style-complement module will not only being diversified from the source ones, but also can be considered as challenging samples to the task model.

The main contributions of our work are summarized as follows.

- We propose a style-complement module that addresses the single domain generalization by learning to generate diverse images.
- A min-max mutual information optimization strategy is designed to gradually enlarge the distribution shift between the generated and the source images, while simultaneously bringing the samples from the same semantic category close in the latent space.
- To validate the effectiveness of the proposed method, extensive experiments are conducted on three benchmark datasets, including digits recognition, corrupted CIFAR-10, and PACS. We further show the effectiveness of our approach in the standard DG setting under the leave-one-domain-out protocol. The results clearly demonstrate that our method surpasses the state-of-the-art DG and Single-DG methods on all datasets.

2. Related Work

Domain Shift. Most of the existing machine learning methods suffer from performance degradation when the source (*i.e.* training) domain and the target (*i.e.* test) domain follow different distributions. The difference between the distributions is termed as the domain shift [36, 8]. In computer vision applications, such a shift may be brought by but not limited to environment and style changes. To this problem, domain adaptation (DA) approaches have been proposed to minimize the domain shift across the domains by matching the marginal [26, 13, 1] or conditional distributions [27, 29] of the source and target domains. Domain adaptation has been widely studied in various settings, such as semi-supervised [18, 38, 46] and unsupervised scenarios [26, 28], which leverage the partially labelled or unlabelled target domain in the training process. More recently, few-shot DA [32] is proposed, where only a few labelled target samples together with source samples are available in the training phase.

Domain Generalization. The most significant difference between domain adaptation (DA) and domain generalization (DG) is that DG does not require access to the target domain in the training phase. Existing DG methods can be rough classified into two categories of learning the domain invariant representation and data augmentation. The key idea behind the former category is to reduce the discrepancy between representations of multiple source domains. Muandet *et al.* [34] first proposed a kernel-based method to obtain domain invariant features. [14] learns the latent invariant representation by jointly considering the domain reconstruction task. [33] further introduces a contrastive semantic alignment loss, which encourages the intra-class similarity and inter-class difference. Li *et al.* proposed to reduce the gap across the domains by minimizing Maximum Mean Discrepancy (MMD) under adversarial autoencoder framework. [44] aims to learn a domain agnostic representation by enforcing an orthogonal constrain on textural information of images and the corresponding latent representations. Recently, meta-learning procedure has been studied to solve DG problem [10, 24, 11]. Li *et al.* put forward a gradient-based model agnostic meta-learning algorithm for DG. Dou *et al.* [10] exploits the episodic training scheme to enforce the global and local alignment. [11] incorporates variational information bottleneck with meta-learning to narrow the domain gap between the source domains.

The other category is related to data augmentation. This line of work generally aims to generate out-of-domain samples, which are then used to train the network along with the source samples to improve the generalization ability. For instance, [42] exploits the adversarial training scheme to generate 'hard' samples for the classifier. Shankar *et al.* [39] proposed to augment the source samples along the direction of the domain change. [50] exploits a conditional gener-

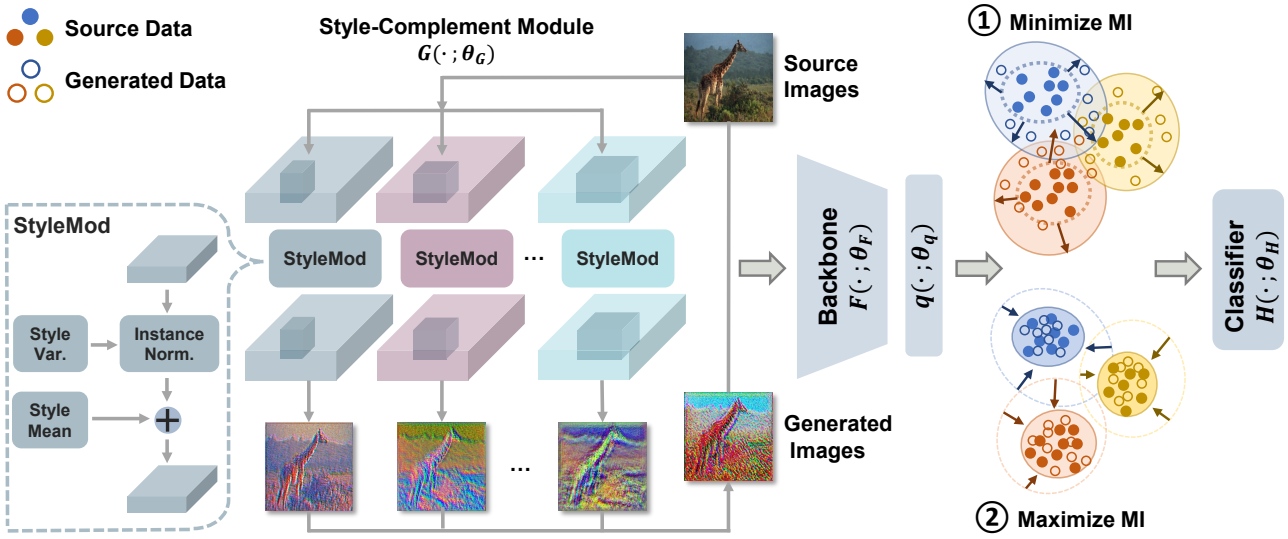


Figure 1: The overall framework of the proposed Learning-to-diversify (L2D). L2D alternatively trains the style-complement module $G(\cdot;\theta_G)$ and the task model $F(\cdot;\theta_F)$, $q(\cdot;\theta_q)$, and $H(\cdot;\theta_H)$. Specifically, (1) the upper bound of mutual information (MI) is minimized between the source and generated images, and (2)MI among samples belonging to the same category is maximized. It enhances the generalization power of the task model in an adversarial min-max manner.

ative adversarial network (GAN) to synthesize data from pseudo-novel domain. [5] is considered as another type of augmentation, which exploits an auxiliary self-supervision training signal from solving a jigsaw puzzle to improve the generalization ability of the classifier.

This paper focuses on a more challenging yet realistic setting, namely single domain generalization [37, 49]. In Single-DG, the network is trained on a single source domain, while it is evaluated on multiple unseen domains. Gradient-based image augmentation is an effective strategy for Single-DG. [37] improves the ADA by encouraging semantic consistency between the augmented and source images in the latent space via an auxiliary Wasserstein autoencoder. [49] considers entropy maximization in the adversarial training framework to generate challenging perturbations of the source samples. In all the aforementioned approaches for Single-DG, the visual differences between the source and generated images are mostly depicted in the color and texture of the augmented samples. Different from existing Single-DG methods, our method aims to generate diverse samples with novel style/texture/appearance having a larger shift from the source distribution, and thus can be considered as complementary to the source data distribution.

3. Methodology

Given a source domain $\mathcal{S} = \{x_i, y_i\}_{i=1}^N$ of N samples, the goal of single domain generalization is to learn a model that can generalize to many unseen target domains $\mathcal{T}$. With no prior knowledge on the target domain, we propose a style-complement module $G(\cdot;\theta_G) : x \rightarrow x^+$ to augment the source domain by synthesizing the x^+ that shares the same semantic information with the source image x, albeit with different styles. As shown in Figure 1, we firstly apply the feature extractor $F(\cdot;\theta_F)$ to transform images x and x^+ to latent vectors z and z^+. To diversify the generated features from the source samples, the MI upper bound approximation for each pair z and z^+ is minimized to learn G; subsequently, we freeze G and maximize MI between the z and z^+ from the same semantic category, which assists the task network F and classifier $H(\cdot;\theta_H)$ parameterized by θ_H to learn discriminative features from diverse-styled images.

Style-Complement Module. The style-complement module $G(\cdot;\theta_G)$ consists of K transformations, each of which is comprised of a convolutional layer, a style learning layer, and a transposed convolutional layer. By applying the convolution operations, the source images are projected from the original distribution to a novel distribution with arbitrary style shift. We further enhance the diversity of the generated samples by creating the style shift at the pixel level. Specifically, we add learnable parameters $\theta_{G,k} = \{\mu_k, \sigma_k\}$ for each of the k transformations, where the learnable parameters $\mu_k, \sigma_k \in \mathbb{R}^{h*w*c}$ are the mean shift and variance shift. More concretely, we have:

$$T(f_{i,k};\theta_{G,k}) = \sigma_k * \frac{f_{i,k} - \mu}{\sigma} + \mu_k, \qquad (1)$$

where $f_{i,k} \in \mathbb{R}^{h*w*c}$ is the output of the convolutional operation applied on x_i in the k-th transformation, with h, w

and c representing the height, width, and channel, respectively. μ and σ correspond to the mean and covariance of the $f_{i,k}$. The transformed feature map $f'_{i,k} = T(f_{i,k}; \theta_{G,k})$ are then applied with the transposed convolutional operation to reconstruct to the original image dimension of x. The final outcome x_i^+ of the style-complement module is the linear combination of the augmented images $x_{i,k}^+$ obtained from k-th transformation:

$$x_i^+ = \frac{1}{\sum_{k=1}^{K}(w_k)} \sum_{k=1}^{K}(w_k \sigma(x_{i,k}^+)), \quad w_k \sim Normal(0,1) \tag{2}$$

where w_k is a scalar sampled from a normal distribution and weights the contribution of augmented image $x_{i,k}$ from the transformation k to the output augmented image x_i^+. We apply the activation function $\sigma(\cdot)$, *e.g. tanh*, to scale x_i^+.

Synthesizing Novel Styles. The objective of the style-complement module is to generalize from the source domain distribution to the out-of-domain distribution. To increase the diversity of the created styles, the correlation between the generated images and the source images should be minimized. Mutual information (MI) $I(z; z^+)$ serves as a measure of quantifying the correlation of z and z^+, which is defined as:

$$I(z; z^+) = \mathbb{E}_{p(z,z^+)}\big[\log \frac{p(z^+|z)}{p(z^+)}\big] \tag{3}$$

We minimize the mutual information (MI) between the source and generated images in the latent feature space $\mathbb{Z}$, achieved by passing images through $F(\cdot; \theta_F)$. The upper bound of MI defined in [7] is:

$$I(z; z^+) \leq \mathbb{E}_{p(z,z^+)}[\log p(z^+|z)] - \mathbb{E}_{p(z)p(z^+)}[\log p(z^+|z)], \tag{4}$$

where z and z^+ are the respective latent vectors of the source image x and the generated image x^+.

Since the conditional distribution $p(z^+|z)$ is intractable, the upper bound of $I(z; z^+)$ cannot be directly minimized. Therefore, we adopt a variational distribution $q(z^+|z)$, that employs a neural network parameterized by θ_q to approximate the upper bound $\hat{I}(z; z^+)$ of mutual information:

$$\hat{I}(z; z^+) = \frac{1}{N}\sum_{i=1}^{N}[\log q_\theta(z_i^+|z_i) - \frac{1}{N}\sum_{j=1}^{N}\log q_\theta(z_j^+|z_i)] \tag{5}$$

By minimizing Eq. (5), the learnable mean/variance shift parameters in our model are trained to complement the style of the source domain. Although $\hat{I}(z; z^+)$ is no longer an upper bound for MI as the conditional distribution $p(z^+|z)$ is substituted with a variational approximation $q_\theta(z^+|z)$, $\hat{I}(z; z^+)$ can be a reliable upper bound estimator if the difference between the two distributions is small. Specifically, we estimate the the difference Δ between $\hat{I}(z; z^+)$ and the upper bound of $I(z; z^+)$ by using the Kullback–Leibler divergence (KLD):

$$\begin{aligned}\Delta &= KLD(p(z^+, z) \parallel q_\theta(z^+, z)) \\ &= \mathbb{E}_{p(z,z^+)}[\log(p(z^+|z)p(z)) - \log(q_\theta(z^+|z)p(z))] \\ &= \mathbb{E}_{p(z,z^+)}[\log p(z^+|z)] - \mathbb{E}_{p(z,z^+)}[\log q_\theta(z^+|z)].\end{aligned} \tag{6}$$

The above equation shows that the difference Δ is affected by two terms. Since the first term of Eq. 6 is not related to θ_q, we minimize the negative log-likelihod between z_i and z_i^+ instead of directly minimizing Δ:

$$L_{likeli} = -\frac{1}{N}\sum_{i=1}^{N}\log q_\theta(z_i^+|z_i). \tag{7}$$

Semantic Consistency. While the style-complement module can generate images with diverse styles, it may introduce noise or generate images with distorted semantic information from original source images (*e.g.*, when the variance shift σ_k equals to 0, the generated images will become meaningless). Therefore, it is important to limit the conditional distribution shift from the source distribution to the out-of-domain distribution, thereby avoiding generating semantically unrelated images. To achieve this, we minimize the class-conditional maximum mean discrepancy (MMD) in the latent space as follows,

$$L_{const} = \frac{1}{C}\sum_{m=1}^{C}(\|\frac{1}{n_s^m}\sum_{i=1}^{n_s^m}\phi(z_i^m) - \frac{1}{n_t^m}\sum_{i=1}^{n_t^m}\phi(z_j^{m+})\|^2), \tag{8}$$

where z_i^m and z_i^{m+} are the i-th latent vector of the source and augmented sample of the class m, respectively. n_s^m and n_t^m are the total number of original and augmented sample of class m. $\phi(\cdot)$ represents the kernel function. Conditional MMD mitigates the potential semantic information distortion by constraining the distribution shift of samples that belong to the same class.

MI Maximization. We aim to obtain a generalizable and robust model by playing a min-max game between the style-complement module $G(\cdot; \theta_G)$ and the task model $F(\cdot; \theta_F)$. While the style-complement module aims to generate diverse images that have minimum information on the source images, the task model can cluster images with same semantic label in the embedding space. The lower bound on MI between two varaibles proposed in [40] is:

$$I(z; z^+) \geq \mathbb{E}\big[\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{f(z,z^+)}}{\sum_{j=1}^{N}e^{f(z_i,z_j^+)}}\big], \tag{9}$$

where $f(\cdot,\cdot)$ is a critic function.

However, directly maximizing the lower bound MI of the generated and source images without leveraging the semantic label may falsely reduce the shared information of same class samples. To alleviate this problem, we employ the supervised contrastive loss [19] to increase the mutual information among samples from the same class, defined as:

$$L_{supcon} = -\sum_{i=0}^{N} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{(z_i \cdot z_p/\tau)}}{\sum_{a \in A(i)} e^{(z_i \cdot z_a/\tau)}}$$
$$P(i) = \{p \in A(i) : y_p = y_i\}, \tag{10}$$

where $A(i)$ is the set of the source and generated latent representations z, z^+ of the same class. τ is the temperature coefficient.

To further enhance the semantic consistency, we minimize the cross-entropy loss on both the source images X and the generated images X^+:

$$L_{task} = -\frac{1}{2N}[\sum_{i=0}^{N} y_i \log(\hat{y_i}) + \sum_{j=0}^{N} y_j^+ \log(\hat{y}_j^+)], \tag{11}$$

where $\hat{y}$ and $\hat{y}^+$ are the prediction of the source and generated images, respectively.

Objective Function. We adopt a two-step training, in which we optimize the style-complement module $G(\cdot;\theta_G)$ and a task model, including $F(\cdot;\theta_F)$, $q(\cdot;\theta_q)$ and $H(\cdot;\theta_H)$ in an iterative manner. Specifically, we train the task module F with both the source images X and the generated images X^+ with the weighted combination of Eq. (5), (7), and (11) as follows:

$$\min_{\theta_F,\theta_q,\theta_H} L = L_{task} + \alpha_1 L_{supcon} + \alpha_2 L_{likli}. \tag{12}$$

Notably, α_1 and α_2 are the hyper-parameters for balancing the losses. To optimize G, we consider solving Eq. (5) and (8) jointly:

$$\min_{\theta_G} L = \hat{I}(z; z^+) + \beta L_{const}, \tag{13}$$

with β being the balancing weight between the mutual information upper bound estimation $\hat{I}(z; z^+)$ and semantic consistent loss L_{const}.

Implementation Notes: To capture multi-scale information from images, we apply different transformations (*i.e.,* kernel sizes) in the convolution and transposed convolution layers. Moreover, to avoid potential distortion of the semantic information, we fix the number of output channels to the number of input color channels (*i.e.,* 3 output channels for RGB images). We re-initialize the weights of the convolutional layer and transposed convolutional layer by sampling from uniform distribution of $(-\frac{1}{\sqrt{size(kernel)}}, \frac{1}{\sqrt{size(kernel)}})$ in each iteration.

Table 1: Single domain generalization accuracy (%) comparison on digits dataset. Models are trained on MNIST and evaluated on the rest of the digits datasets. Best performances are highlighted in bold.

	SVHN	MNIST-M	SYN	USPS	Avg.
ERM	27.83	52.72	39.65	76.94	49.29
CCSA	25.89	49.29	37.31	83.72	49.05
d-SNE	26.22	50.98	37.83	**93.16**	52.05
JiGen	33.80	57.80	43.79	77.15	53.14
ADA	35.51	60.41	45.32	77.26	54.62
M-ADA	42.55	67.94	48.95	78.53	59.49
ME-ADA	42.56	63.27	50.39	81.04	59.32
Ours	**62.86**	**87.30**	**63.72**	83.97	**74.46**

4. Experiments

Datasets. To evaluate the effectiveness of the proposed method, we conduct experiments over three single-dg benchmark datasets. **Digits** consists of 5 different datasets, which are MNIST[22], SVHN[35], MNIST-M[12], SYN[12] and USPS[9]. Each dataset is considered as a unique domain which may be different from the rest of the domains in font style, background, and stroke color. **PACS** [23] is a recent proposed DG benchmark dataset that has four domains, including photo, art painting, cartoon, and sketch. Each domain contains 224×224 images belonging to seven categories, and there are 9,991 images in total. Compared with the digits dataset, PACS is considered as a more challenging dataset due to the large style shift among domains. For fair comparison, we follow the official split of the train, validation, and test. **Corrupted CIFAR-10** [16, 20] contains 32×32 tiny RGB images from CIFAR-10 that are corrupted by different types of noises. There are 15 corruptions from 4 main categories, including weather, blur, noise, and digital. Each of the corruptions has 5 levels serverities and '5' indicates the severest corruption.

4.1. Comparisons on Digits

Experiment Setup. Following [42, 37, 49], we select 10,000 images from MNIST as the source domain and test the generalization performance of models on the other four digits datasets. We resize all the images to 32×32 and duplicate their channels to convert all the grayscale images to RGB. We employ the LeNet [22] as the backbone and set the batch size as 32. We use SGD to optimize both the style-complement module and the task model.

Results. Table 1 shows that our model achieves the highest average accuracy compared to the other baselines. Specifically, we observe clear improvements of 20.3%, 29.36%, 13.33% and 14.9% on SVHN, MNIST-M, SYN, and overall accuracy, respectively. Previous Single-DG methods di-

Table 2: Single domain generalization accuracy (%). Models are trained on CIFAR-10 dataset and evaluated on CIFAR-10-C dataset with corruption severity level 5. We report the average accuracy over 4 main categories of corruption: Weather, Blur, Noise, and Digital. Best performances are highlighted in bold. * indicates our implementation.

	Weather	Blur	Noise	Digits	Avg.
ERM	67.28	56.73	30.02	62.30	54.08
CCSA	67.66	57.81	28.73	61.96	54.04
d-SNE	67.90	56.59	33.97	61.83	55.07
ADA*	72.67	67.04	39.97	66.62	61.58
M-ADA	75.54	63.76	54.21	65.10	64.65
ME-ADA*	74.44	**71.37**	66.47	70.83	70.77
Ours	**75.98**	69.16	**73.29**	**72.02**	**72.61**

rectly generate the auxiliary training samples by applying the adversarial perturbation. Compared to the adversarial gradient-based methods, L2D creates larger domain shifts between the generated images and the source images by incorporating the style-complement module. The substantial increase of the accuracy over the baselines demonstrates the importance of generating diverse-styled images on improving the generalization power of a model. Moreover, we observe that single-DG methods generally achieve better performance in this task, reflecting the dependency of previous DG methods on multiple source domains to learn a generalizable model. Our method achieves the second best performance on USPS. We infer this might be related to the fact the USPS and the source domain MNIST share very similar stroke styles. In such case, the diverse generated images might not benefit the generalizibility of the model as much as the task with a larger domain shift.

4.2. Comparisons on Corrupted CIFAR-10

Experiment Setup. We train all models on the training split of CIFAR-10 dataset (50,000 images) and evaluate them on the corrupted test split of CIFAR-10 (10,000 images). Following [37], we select WideResNet (16-4) [48] as the backbone network with a batch size of 256. We optimize the model by using SGD with Nesterov momentum and weight decay rate of 0.0005. The learning rate is initialized to 0.2 which is gradually decreased by using the cosine annealing scheduler.

Results. We report the average accuracy of four categories under level-5 severity corruption in Table 2. Our method achieves the highest average performance, surpassing the best baseline by approximately 2.6%. Notably, for noise corruptions, accuracy is significantly improved by approximately 13.2%. We also report the performance of methods against all five levels of Noise corruption and Digits corruption in Figure 3. From the figure, we can see that the performance margin between our method and others are relatively small under severity level one, and it is gradually enlarged while the severity level goes up. This further validates that our model not only can achieve the highest average performance, but also is resilient towards the severe corruptions.

Table 3: Single domain generalization accuracy (%) on PACS. Models are trained on *photo* and evaluated on the rest of the target domains (*i.e., art painting, cartoon, and sketch*). The comparison is based on our implementation. Best performances are highlighted in bold.

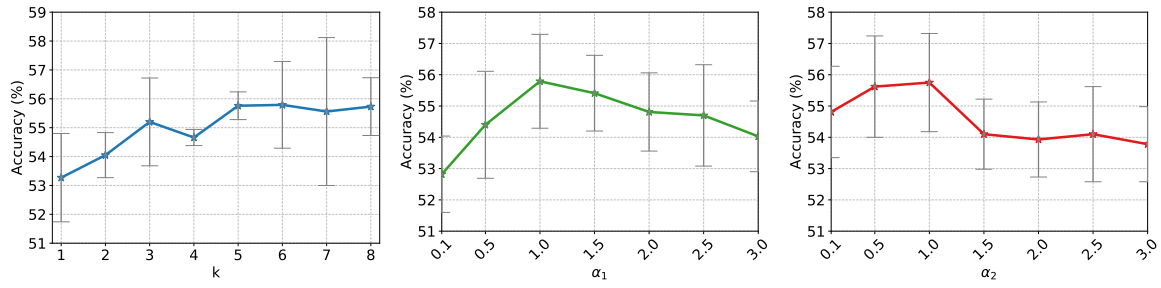
	A	C	S	Avg.
ERM	54.43	42.74	42.02	46.39
JiGen	54.98	42.62	40.62	46.07
RSC	56.26	39.59	47.13	47.66
ADA	58.72	45.58	48.26	50.85
ME-ADA	**58.96**	44.09	49.96	51.00
Ours w/o Style-comp.	53.27	41.00	41.92	45.39
Ours w/o Mod.	58.48	48.96	53.20	53.54
Ours w/o Min. MI	56.49	48.08	56.32	53.63
Ours w/o Max. MI	56.64	47.08	49.68	51.13
Ours (Full Model)	56.26	**51.04**	**58.42**	**55.24**

4.3. Comparisons on PACS

Experiment Setup. For the single domain generalization task, we consider a practical case, where we utilize a set of easy-to-collect realistic images (*i.e.*, *photo*) as the source domain, and evaluate models on the rest of diverse-styled domains (*i.e.*, *art painting*, *cartoon*, and *sketch*). AlexNet [21] is employed as the backbone which is pretrained on Imagenet and finetuned on the source domain. We also evaluate the effectiveness of our approach on PACS under the standard leave-one-domain-out protocol, where one domain is selected as the test domain and the rest are treated as the source domain. We employ pretrained Alexnet and ResNet-18 [15] as the backbone networks for the leave-one-domain-out setting. Please refer to the supplementary material for more implementation details.

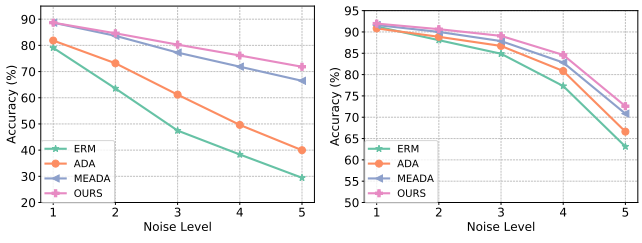
Results. From Table 3, we can see that our method can achieve the best average classification accuracy compared to the baselines. Importantly, our method can achieve a relatively large performance margin over other methods on the *sketch* domain, which has the largest domain shift from *photo* due to its highly abstracted shapes. This result verifies that our method takes the advantage of diverse style images that are generated by the style-complement module.

To further validate the performance of our method, we conduct the leave-one-domain-out domain generalization task on PACS. We compare our approach with the two categories of recent state-of-the-art DG methods. The first category,including DSN [4], Fusion [30], MetaReg [2], Epi-



(a) (b) (c)

Figure 2: Sensitivity analysis of L2D with respect to parameters k, α_1, and α_2 on the PACS dataset. The reported accuracy is the averaged over all three unseen target domains.



(a) Noise (b) Digital

Figure 3: Average classification accuracy (%) of different methods under five levels of severity on *Noise* and *Digital* corruptions.

FCR [25], MASF [10] and DMG [6], requires domain identifications in the training phase. Methods in the second category are inline with a more realistic mixed latent domain setting [31], where domain identifiers are unavailable in the training phase. AGG [25], HEX [44], PAR [43], JiGen [5], ADA [42], MEADA [49], MMLD [31] and our method belong to the latter category. We report the results with different backbone networks in Table 4. Without leveraging the domain identifier, our method can still achieve the state-of-the-art performance on the leave-one-domain-out generalization task on PACS. In the training stage, the image augmentation module gradually enlarges the domain shift between the generated images and the source images.

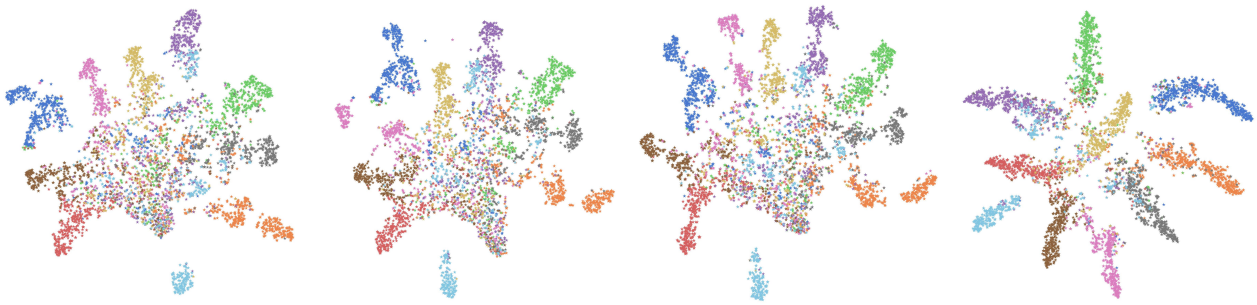
Table 4: Leave-one-domain-out classification accuracy(%) on PACS. Best performances are highlighted in bold.

	D_ID	P	A	C	S	Avg.
AlexNet						
DSN	✓	83.30	61.10	66.50	58.60	67.40
Fusion	✓	90.20	64.10	66.80	60.10	70.30
MetaReg	✓	87.40	63.50	69.50	59.10	69.90
Epi-FCR	✓	86.10	64.70	72.30	65.00	72.00
MASF	✓	90.68	70.35	72.46	67.33	75.21
DMG	✓	87.31	64.65	69.88	**71.42**	73.32
HEX	✗	87.90	66.80	69.70	56.20	70.20
PAR	✗	89.60	66.30	66.30	64.10	72.08
JiGen	✗	89.00	67.63	71.71	65.18	73.38
ADA	✗	85.10	64.30	69.80	60.40	69.90
MEADA	✗	88.60	67.10	69.90	63.00	72.20
MMLD	✗	88.98	69.27	**72.83**	66.44	74.38
Ours	✗	**90.96**	**71.19**	72.18	**67.68**	**75.50**
ResNet-18						
Epi-FCR	✓	93.90	**82.10**	77.00	73.00	81.50
MASF	✓	94.99	80.29	77.17	71.68	81.03
DMG	✓	93.55	76.90	**80.38**	75.21	81.46
Jigen	✗	96.03	79.42	75.25	71.35	80.51
ADA	✗	95.61	78.32	77.65	74.21	81.44
MEADA	✗	95.57	78.61	78.65	75.59	82.10
MMLD	✗	**96.09**	81.28	77.16	72.29	81.83
Ours	✗	95.51	**81.44**	**79.56**	**80.58**	**84.27**

4.4. Empirical Analysis

Ablation Study. We conduct the ablation study on PACS dataset to verify the effectiveness of each component in our framework. Table 3 reports the classification results of the 4 variants of our original framework. We report the baseline result without incorporating the style-complement module as *w/o Style-comp.*. Without the generated images, the model is degraded to the backbone model with a variational approximation of z in the embedding space. The large performance margin between *w/o Style-comp.* and our full model demonstrates the importance of the style-complement module on improving the generalization ability of the model. *w/o Mod.* shows the result after removing the style modification from the full model. The removal of the style modification triggers 1.35% absolute performance decline. We infer this relates to the limited style diversity of the generated images.

To provide an understanding of how mutual information affects the learning framework, in *w/o Min. MI* and *w/o*



(a) ERM (b) ADA (c) ME-ADA (d) OURS

Figure 4: The t-SNE visualizations of extracted unseen target feature distribution for different methods on digits Single-DG task. Features with the same semantic label are plotted in the same color.

Max. MI, we remove the mutual information minimization and maximization process, respectively. We observe the accuracy of *w/o Min. MI* is 1.26% lower than the full model which is similar to the results of *w/o Mod.*. That means, without the mutual information constraint term, the style-complement module tends to generate images by following the source distribution, which limits their diversity. Meanwhile, without the mutual information maximization process, *w/o Max. MI* a clear performance margin of 3.76% is shown compared to the full model. This indicates that mapping diverse styled images from the same class closer in the embedding space is helpful for improving the generalization power of the model.

Parameter Sensitivity. To validate the significance of the total number of transformations k, and weighting parameters α_1, and α_2 in the loss, we conduct sensitivity analysis on PACS dataset. In the experiments, we initially set $k = 6$, $\alpha_1 = 1$, and $\alpha_2 = 1$. When we analyze the sensitivity to a specific parameter, we fix the values of the other two parameters. Figure 2 shows the results of sensitivity analysis under the Single-DG setting. From Figure 2(a), we can see that the performance is gradually increasing as the style-complement module combines more transformations. Also, the average performance become relatively stable after $k = 5$. The results also indicates that jointly considering multiple transformations increases the diversity of the generated images. Meanwhile, excessive transformations might produce extra noise, which hinders the further performance gain. As can be seen from Figure 2 (b) and (c), our method surpasses the state-of-the-art performance in a wide range of α settings, and achieves the best classification accuracy with $\alpha_1 = \alpha_2 = 1$.

t-SNE Visualizations. To further demonstrate the effectiveness of the proposed method, we use t-SNE [41] to visualize the distribution of the unseen target features in the digits dataset (*i.e.*, SVHN, SYN, USPS and MNIST-M). Specifically, we train different models on MNIST and randomly select 1000 samples from each unseen target domains to visualize. As shown in Fig. 4, our method clearly achieve better class-wise separation than the baselines. Moreover, from the distribution of features extracted by ERM, ADA and ME-ADA, we observe that the features within the same class can have multiple sub-clusters. This indicates that it is hard for the methods to learn domain invariant representations and the large intra-class variations may hinder them obtaining a clear decision boundary on the targets. In contrast, it is clearly seen that our approach creates tighter clusters with a good class-wise mix compared to the baselines. This strongly supports the idea of diverse image generation and the maximization of mutual information among the samples from the same class.

5. Conclusion

This paper presented Learning-to-Diversify (L2D), a novel approach to address single domain generalization. Unlike previous domain generalization methods which exploit multiple source domains to learn domain invariant representations, the proposed approach designs a style-complement module to generate diverse out-of-domain images from single source domain. An iterative min-max mutual information (MI) optimization strategy is used to boost the generalization power of the model. The tractable MI upper bound is minimized to further enhance the diversity of the generated image, while MI among same category samples is maximized to obtain the style-invariant representations in the maximization step. Extensive experiments on three benchmark datasets demonstrate that the proposed method outperforms the state-of-the-art methods on both single domain generalization and the standard leave-one-domain-out domain generalization.

Acknowledgement This work was supported by Australian Research Council (ARC DP190102353, and CE200100025)

References

[1] Mahsa Baktashmotlagh, Mehrtash Tafazzoli Harandi, Brian C. Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013.

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018.

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 2010.

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NeurIPS*, 2016.

[5] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

[6] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, 2020.

[7] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *ICML*, 2020.

[8] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*. 2017.

[9] John S. Denker, W. R. Gardner, Hans Peter Graf, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, Lawrence D. Jackel, Henry S. Baird, and Isabelle Guyon. Neural network recognizer for hand-written zip code digits. In *NeurIPS*, 1988.

[10] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019.

[11] Ying-Jun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees G. M. Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *ECCV*, 2020.

[12] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

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

[14] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

[17] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020.

[18] Hal Daumé III, Abhishek Kumar, and Avishek Saha. Co-regularization based semi-supervised domain adaptation. In *NeurIPS*, 2010.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.

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

[25] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. In *ICCV*, 2019.

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

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

[28] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[29] Yadan Luo, Zijian Wang, Zi Huang, and Mahsa Baktashmotlagh. Progressive graph learning for open-set domain adaptation. In *ICML*, 2020.

[30] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: Domain generalization through source-specific nets. In *ICIP*, 2018.

[31] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020.

[32] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *NeurIPS*, 2017.

[33] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017.

[34] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013.

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[36] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 2010.

[37] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, 2020.

[38] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, 2019.

[39] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018.

[40] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 2008.

[42] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018.

[43] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[44] Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *ICLR*, 2019.

[45] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018.

[46] Zijian Wang, Yadan Luo, Zi Huang, and Mahsa Baktashmotlagh. Prototype-matching graph network for heterogeneous domain adaptation. In *MM*, 2020.

[47] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *ICLR*, 2021.

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

[49] Long Zhao, Ting Liu, Xi Peng, and Dimitris N. Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *NeurIPS*, 2020.

[50] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, 2020.