

Feature Stylization and Domain-aware Contrastive Learning for Domain Generalization

Seogkyu Jeon
Department of Computer Science
Yonsei University, South Korea
jone9312@yonsei.ac.kr

Kibeom Hong
Department of Computer Science
Yonsei University, South Korea
cha2068@yonsei.ac.kr

Pilhyeon Lee
Department of Computer Science
Yonsei University, South Korea
lph1114@yonsei.ac.kr

Jewook Lee
Department of Computer Science
Yonsei University, South Korea
hooraid@yonsei.ac.kr

Hyeran Byun^{*†}
Department of Computer Science
Yonsei University, South Korea
hrbyun@yonsei.ac.kr

ABSTRACT

Domain generalization aims to enhance the model robustness against domain shift without accessing the target domain. Since the available source domains for training are limited, recent approaches focus on generating samples of novel domains. Nevertheless, they either struggle with the optimization problem when synthesizing abundant domains or cause the distortion of class semantics. To these ends, we propose a novel domain generalization framework where feature statistics are utilized for stylizing original features to ones with novel domain properties. To preserve class information during stylization, we first decompose features into high and low frequency components. Afterward, we stylize the low frequency components with the novel domain styles sampled from the manipulated statistics, while preserving the shape cues in high frequency ones. As the final step, we re-merge both the components to synthesize novel domain features. To enhance domain robustness, we utilize the stylized features to maintain the model consistency in terms of features as well as outputs. We achieve the feature consistency with the proposed domain-aware supervised contrastive loss, which ensures domain invariance while increasing class discriminability. Experimental results demonstrate the effectiveness of the proposed feature stylization and the domain-aware contrastive loss. Through quantitative comparisons, we verify the lead of our method upon existing state-of-the-art methods on two benchmarks, PACS and Office-Home.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision; Image representations.**

^{*}Corresponding Author

[†]Also with Graduate school of Artificial Intelligence, Yonsei University

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3475271>

KEYWORDS

Domain Generalization; Deep learning; Image Classification; Feature Stylization; Contrastive Learning

ACM Reference Format:

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 (MM '21)*, October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3474085.3475271>

1 INTRODUCTION

Since the remarkable advance of deep neural networks, they have become ubiquitous in various fields, especially computer vision systems. However, there still exist potential risks lying on the flip side of their success. One of the main concerns is their vulnerability against visual domain shift [3, 4, 40]. Concretely, deep models react unexpectedly when confronting data unaffiliated to the training distribution. For example, an auto-tagging model trained on clean product images shows poor performance when taking as inputs real product images which are under various viewpoints and light conditions.

To equip the models with the ability in coping with domain shift, previous studies tackle the problem of domain adaptation [8, 26, 42, 48, 50, 60]. In this problem setting, two different domains sharing the same label space are prepared for training and test, which are referred to as the “source” and the “target” domains, respectively. During training, a model has access to both labeled images from the source domain and unlabeled (or partially labeled) images from the target domain. Being aware of the target domain, existing works successfully minimize the discrepancy between two distinct domains, thus leading to large performance boosts [3, 20, 26, 42, 48, 50, 56, 60].

However, the problem setting of domain adaptation is impractical in that domain shift is generally unpredictable in the real-world scenarios, *i.e.*, we do not know the target domain at training time. To this end, a new task has attracted much attention recently, aiming to learn domain robustness without accessing the target domain data, namely domain generalization [5–7, 11, 15, 21, 29–31, 37, 44, 55]. In this setting, multiple datasets from different source domains are typically utilized to learn domain invariant representations.

Previous methods [35, 38, 57, 62] embrace the observation that domain robustness is proportional to the number of domains observable in the training stage [49]. To that end, they utilize generative adversarial networks (GAN) [62] or adaptive instance normalization (AdaIN) [22, 35] for synthesizing novel (unseen) domains. Nonetheless, they have two clear limitations which are critical for domain generalization. First, GAN-based methods become prohibitively difficult to optimize as the number of novel domains increases, which limits the size of observable domain space. Next, AdaIN-based approaches fail to preserve the semantics of original images, as instance normalization (IN) tends to wash away class discriminative information [34, 44].

In this paper, we introduce a novel framework for domain generalization, overcoming the above limitations. Specifically, to synthesize novel domains without losing class discriminative information, we propose a novel *feature stylization block*. First, we calculate batch-wise feature statistics of source domains and sample novel domain styles from the feature distribution. We re-scale the standard deviation of the source feature distribution so that the outlying style statistics are more likely to be sampled. However, the original semantics can be distorted during the stylization process, which will disturb the training. To preserve the original semantics during stylization, inspired by a recent photo-realistic stylization method [59], we decompose original features into high and low frequency components which contain structural and textural information, respectively. Afterwards, we manipulate the low frequency components while remaining shape cues in high frequency ones to prevent semantics distortion. Lastly, we re-merge the stylized low frequency components and the high frequency ones, leading to the stylized features. By incorporating them in the training, our model is allowed to learn robust representation against domain shift.

Rather than naively utilizing stylized features for training, we seek for better strategies that can provide domain robustness guidance. Intuitively, a robust model against domain shift should yield consistent predictions for the stylized features and the original ones. In this point of view, we adopt the *consistency loss* to maximize the agreement between the model predictions for them. Concretely, we measure the KL divergence between two output distributions and minimize it with the consistency loss.

Moreover, we propose the *domain-aware supervised contrastive loss* to minimize distance between the stylized and the original features, in order to achieve feature-level consistency. Although the conventional supervised contrastive loss has proven to be effective, we found that it is unsuitable for domain generalization. The loss expels the samples from different domains and thus disturbs domain invariance, which conflicts with the goal of domain generalization. To this end, we introduce the novel domain-aware supervised contrastive loss which ignores negative samples from different domains, hence preserving domain invariance while empowering class discriminability.

The contributions of this paper can be summarized into three folds. Firstly, we propose a novel domain generalization framework, where diverse domain styles are generated and leveraged through the proposed feature stylization block. The stylized features are in turn used to enhance domain robustness by encouraging the model to produce consistent outputs. Secondly, we introduce the novel domain-aware supervised contrastive loss. The proposed loss

strengthens the domain invariance by contrasting features with respect to domain and class labels. Lastly, we demonstrate the effectiveness of each component of our model through analyses and ablation studies. Furthermore, experimental results show that our method surpasses previous methods with obvious margins, achieving a new state-of-the-art on the widely used benchmarks: PACS and Office-Home. Even on the single-source domain generalization task, our method shows delightful performance improvements over the baseline.

2 RELATED WORKS

2.1 Domain Adaptation

Domain adaptation aims to transfer learned knowledge from source domains to a target domain. In this setting, the source domain is usually a large scale dataset with annotations, and the target domain data is either partially labeled or completely unlabeled. They are referred to semi-supervised domain adaptation (SSDA) [1, 9, 41, 58] and unsupervised domain adaptation (UDA) [8, 20, 26, 42, 48, 50, 60], respectively.

Semi-supervised domain adaptation methods impose constraints on both labeled and unlabeled instances of the target domain in various ways. Donahue *et al.* [9] build a similarity graph to constrain unlabeled data and transfer knowledge with a projective model transfer method. Ao *et al.* [1] distill knowledge from the source domain by generating pseudo labels for the unlabeled target data. Saito *et al.* [41] estimate class-specific prototypes with sparsely labeled examples of the target domain, then update them by solving a minimax game on the unlabeled data.

In unsupervised domain adaptation, most methods [8, 20, 26, 48, 50] conduct feature alignment between source and target domains. To this end, CORAL [48] minimizes the distance between the covariance matrices, while ADDA [50] employs a domain discriminator for adversarial learning. Meanwhile, CyCADA [20] adopts an image-to-image translation framework to transfer the source domain data to the target domain data on image-level. Recently, domain randomization [26, 49, 60, 61] is another generative stream which diversifies the textures of source domain images, allowing the model to learn texture invariant representations. Yue *et al.* [60] manipulate images into an external class from ImageNet [27], and LTIR [26] exploits an artistic style transfer method to alter the textures of the source and target domains.

Our method relates to the domain randomization approaches in that it aims to generate features with diverse domain characteristics. However, they are not suitable for domain generalization as they require external datasets [26, 27, 60]. On the contrary, our proposed feature stylization block is able to generate various stylized features based on the statistics of source domains, without access to additional data.

2.2 Domain Generalization

The goal of domain generalization is to learn domain invariant representations based on only source domains. Different from unsupervised domain adaptation, target domain data is inaccessible during training, making the task more challenging. In addition, multiple domains are typically utilized to achieve domain-agnostic

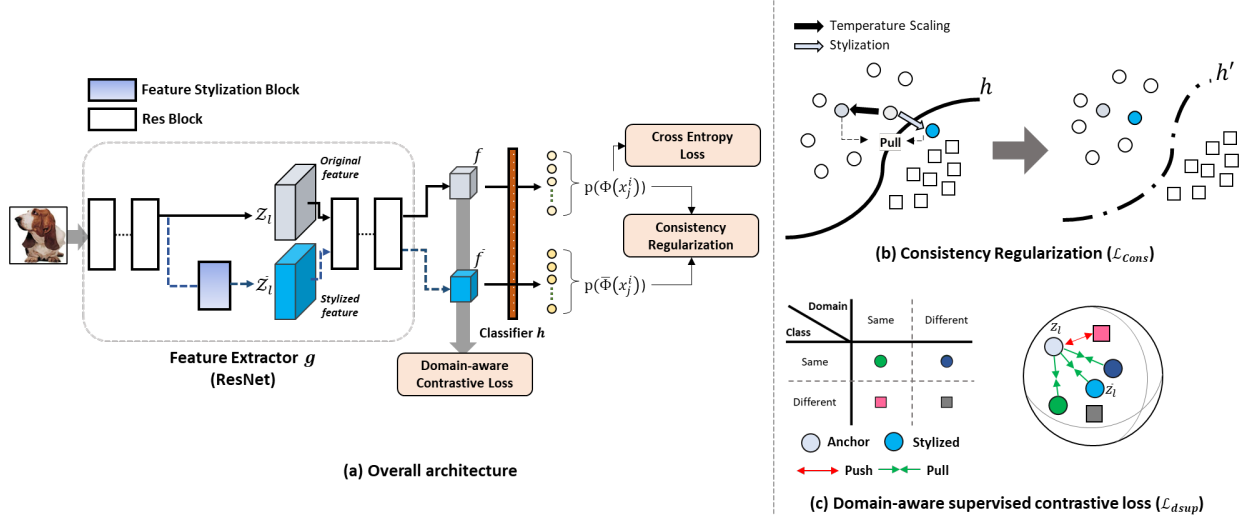


Figure 1: The overview of our proposed methods. We illustrate the overall architecture on (a). The backbone consists of multiple convolutional layers, and we stylize the feature on the intermediate layer of backbone. Both original and stylized features follow the same forward path, producing class predictions $p(\Phi(x_j^i))$ and $p(\Phi(X_j^i))$. The predictions and features are exploited with the consistency regularization (\mathcal{L}_{cons}) and domain-aware supervised contrastive loss (\mathcal{L}_{dsup}) which are described in (b) and (c), respectively.

representation without having target domain data. Previous domain generalization methods can be roughly categorized into four groups: meta-learning, architectural modification, regularization, and generative approaches.

The first group exploits meta-learning techniques [13, 39] to align different domains [2, 11, 29, 30]. These approaches borrow the powerful adaptability of meta-learning algorithms whose effectiveness is proven in the field of few-shot learning. Representatively, Li *et al.* [30] separates the training set into multiple episodes, each of which handles only a single domain. During training, they update the backbone with aggregated regularization losses from domain specific networks. Meanwhile, MASF [11] simulates domain shift using different episodes. They perform global alignment of class relationships while clustering local samples.

Secondly, some works [6, 15, 31, 33, 44, 55] try architectural changes to model a shared embedding space [15, 31, 33] or to build domain-specific networks [6, 44]. Exploiting auxiliary pretext tasks are also favored as a sub-stream [5, 54]. As a pioneer, JiGen [5] proposes to solve jigsaw puzzles as an auxiliary task to induce the model to learn the concepts of spatial correlation. Inheriting from JiGen, EIS-Net [54] employs a momentum metric learning task to provide extrinsic relationship supervision. Other approaches [7, 23, 46, 52, 53] apply diverse regularization during training. HEX [53] employs the neural gray-Level co-occurrence matrix to find superficial representations related to the task. PAR [52] penalizes the predictive power of earlier layers so that the model relies more on global representations from deeper layers. RSC [23] masks out both spatial regions and channels which have high contributions to the task. RobustNet [7] encourage model to utilize domain-invariant features by selectively whitening domain-variant feature channels in the gram matrix during training.

Lastly, based on the intuition that the generalization ability can be boosted with samples from more diverse domains [49], generative approaches arises [35, 38, 57, 62]. They augment the training set with samples similar in semantics but different in domain characteristics. L2A-OT [62] adopts generative adversarial networks to synthesize images which are distant from original ones in terms of the Wasserstein distance. Qiao *et al.* [38] apply adversarial perturbations on the images to augment the source domains. From the perspective of frequency, FACT [55] analyze the frequency components of the image with the fourier transformation, and conduct data augmentation by mixing the amplitude information.

Our method can be viewed as a harmonious combination of the generative method and the regularization-based approach. We generate features of novel domains via the novel feature stylization block during training, then apply regularization in terms of output consistency and feature similarity. The efficacy of our method is demonstrated through extensive experiments in Sec. 4.

3 METHODS

In this section, we first describe the baseline setup of multi-source domain generalization for image classification, then introduce our novel feature stylization method and consistency learning process thereafter. The overall framework of our method is illustrated in Fig. 1 (a).

3.1 Baseline

In the multi-source domain generalization task, multiple datasets of K source domains $\mathcal{D} = \{D_1, D_2, \dots, D_K\}$ are accessible during training. Each dataset D_i contains a set of images $X^i = \{x_1^i, x_2^i, \dots, x_{n_i}^i\}$ with the corresponding class label set $Y^i = \{y_1^i, y_2^i, \dots, y_{n_i}^i\}$, where n_i is the number of images in the i -th dataset. Naturally, the domain

label of x_j^i can be obtained as $d_j^i = i$. We also note that all datasets share the same label space, i.e., $y_j^i \in \mathcal{Y}$. We train a neural network Φ which consists of a feature extractor g and a following classifier h . The feature extractor is composed of multiple convolutional layers and we denote the output features of the l -th convolutional layer by $z_l \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$, where B is the cardinality of a mini-batch, and C_l is the number of channels, while H_l and W_l are the height and width of the feature, respectively. The classifier is a single fully-connected layer. We train the network Φ by minimizing the cross-entropy loss as follows.

$$\mathcal{L}_{ce} = -\frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} y_j^i \log(p(\Phi(x_j^i))), \quad (1)$$

where $p(\cdot)$ indicates the softmax function. Consequently, with the above baseline setup, the network Φ is trained to classify an image x_j^i into its corresponding label y_j^i .

3.2 Feature Stylization

An intuitive way to improve the generalization ability of a model would be allowing it to see diverse samples from different domains [49]. In this point of view, we augment the source domains by synthesizing novel domains by manipulating feature statistics. Before stylization, we note that it should be ensured that the generated feature should maintain the original semantics. To this end, we borrow the feature decomposition of a photo-realistic style transfer model [59], where structural features and textural features are separated into high frequency and low frequency components, respectively. The feature decomposition process is formulated as:

$$\begin{aligned} z_l^L &= \text{UP}(\text{AvgPool}(z_l)), \\ z_l^H &= z_l - z_l^L, \end{aligned} \quad (2)$$

where ‘‘AvgPool’’ denotes spatial average pooling operation with the kernel size of 2, and ‘‘UP’’ indicates nearest neighbor upsampling operation. After decomposition, we perform stylization on the low frequency feature z_l^L only to preserve structural information.

Since neither extra datasets nor pre-trained networks are available in our setting, we stylize the feature by utilizing its batch-wise statistics. Firstly, the mean and variance are obtained as follows:

$$\begin{aligned} \mu_l^L &= \frac{1}{BH_lW_l} \sum_{m=1}^{BH_lW_l} \text{flat}(z_{m,l}^L), \\ (\sigma_l^L)^2 &= \frac{1}{BH_lW_l} \sum_{m=1}^{BH_lW_l} (\text{flat}(z_{m,l}^L) - \mu_l^L)^2, \end{aligned} \quad (3)$$

where $\text{flat}(\cdot) : \mathbb{R}^{B \times C_l \times H_l \times W_l} \rightarrow \mathbb{R}^{BH_lW_l \times C_l}$ indicates flattening operation, while $\mu_l^L, \sigma_l^L \in \mathbb{R}^{C_l}$ denote the mean and the variance of feature style, respectively.

As thoroughly investigated in previous studies [32, 36], these batch-wise statistics are highly related to the domain characteristics. In order to generate the domain statistics, we model the prior distributions of both style vectors (μ_l^L, σ_l^L) as gaussians. For this purpose, we calculate the channel-wise mean and variance of style

vectors as follows.

$$\begin{aligned} \hat{\mu}_l^L &= \frac{1}{C_l} \sum_{c=1}^{C_l} \mu_{c,l}^L, \quad (\hat{\sigma}_l^L)^2 = \frac{1}{C_l} \sum_{c=1}^{C_l} (\mu_{c,l}^L - \hat{\mu}_l^L)^2, \\ \hat{\sigma}_l^L &= \frac{1}{C_l} \sum_{c=1}^{C_l} \sigma_{c,l}^L, \quad (\tilde{\sigma}_l^L)^2 = \frac{1}{C_l} \sum_{c=1}^{C_l} (\sigma_{c,l}^L - \hat{\sigma}_l^L)^2, \end{aligned} \quad (4)$$

where $\hat{\mu}_l^L$ and $(\hat{\sigma}_l^L)^2$ denote channel-wise statistics of μ_l^L , while $\tilde{\mu}_l^L$ and $(\tilde{\sigma}_l^L)^2$ are statistics of σ_l^L . C_l is the number of channels of z_l .

To generate the novel domain styles, we manipulate the variance of distributions with the scaling parameters s_μ and s_σ , then sample new style vectors from its distribution as:

$$\begin{aligned} \mu_l^{\text{new}} &\sim \mathcal{N}(\hat{\mu}_l^L, s_\mu (\hat{\sigma}_l^L)^2), \\ \sigma_l^{\text{new}} &\sim \mathcal{N}(\tilde{\mu}_l^L, s_\sigma (\tilde{\sigma}_l^L)^2). \end{aligned} \quad (5)$$

As the variance increases, outlying style vectors, i.e., outliers, are more likely to be sampled from the distributions, whereas in-liners are sampled with higher probability in the opposite case. We show the effects of scale parameters through experiments in section 4.4.

After the sampling stage, the style vectors μ_l^{new} and σ_l^{new} are applied to the original low frequency component z_l^L via the affine transformation as follows.

$$\bar{z}_l^L = \sigma_l^{\text{new}} \left(\frac{z_l^L - \mu_l^L}{\sigma_l^L} \right) + \mu_l^{\text{new}}. \quad (6)$$

We can interpret the sampling and transformation process as generating arbitrary domain statistics and applying on the original one. Notably, our affine transformation process is analogous to the batch normalization (BN) [24] where the affine parameters are μ_l^{new} and σ_l^{new} . Compared to adaptive instance normalization (AdaIN) [22] which washes away discriminative features [34, 44], our feature stylization process conserves them while generating novel domain styles.

Lastly, the stylized low frequency feature \bar{z}_l^L are then combined with the original high frequency feature z_l^H via the following equation.

$$\bar{z}_l = z_l^H + \bar{z}_l^L. \quad (7)$$

We note that our feature stylization block can be inserted into any layer of the feature extractor g . Analyses on the best location l of the feature stylization block will be discussed through ablation studies.

3.3 Consistency Regularization

The augmented feature \bar{z}_l is passed through the remaining layers, the same as the original feature z_l . Given the stylized feature, we further encourage the model to output a consistent prediction with the original one. For this, we minimize the discrepancy between output predictions from the original and the stylized features, which is formulated as:

$$\mathcal{L}_{cons} = -\frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} p(\Phi(x_j^i), \tau) \log(p(\Phi(x_j^i))), \quad 0 \leq \tau \leq 1, \quad (8)$$

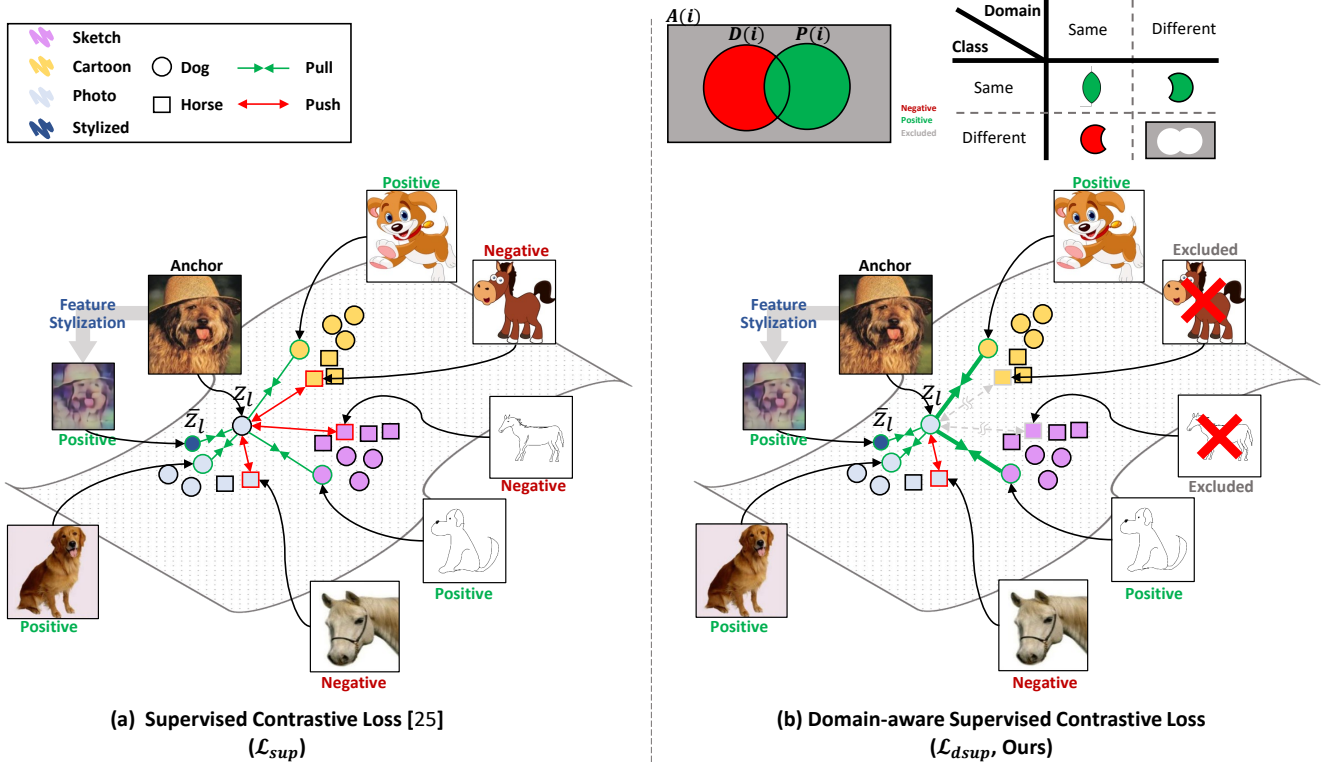


Figure 2: We illustrate the difference between the conventional supervised contrastive loss [25] (\mathcal{L}_{sup}) and the proposed domain-aware supervised contrastive loss ($\mathcal{L}_{dsupsup}$). \mathcal{L}_{sup} only considers the class label to compose the positive and negative sets. From the domain perspective, the sketch and cartoon domain images in $P(i)$ are attracted to the anchor, but those domains included in the negatives are expelled. This is contradictory since the positives contribute to the domain invariance while the negatives cause the domain discrepancy. For this, we propose $\mathcal{L}_{dsupsup}$ where we exclude the samples with different domains from the negative set. Consequently, the class discriminative feature is attainable and the domain-invariance is also accomplished by attracting positive samples from different domains.

where $p(\cdot)$ indicates the softmax function, $\Phi(x_j^i)$ denotes the neural network output in which feature stylization is performed on an intermediate layer, and τ is the temperature hyper-parameter.

The effect of the consistency loss is illustrated in Fig. 1 (b). Through the consistency loss, the log-likelihood between the predictions is maximized. In addition, we apply temperature scaling with τ on the original prediction, denoted as $p(\Phi(x_j^i), \tau)$, to encourage the prediction of stylized feature to have low entropy [16, 19].

3.4 Domain-aware Supervised Contrastive Loss

Furthermore, we bring another intuition that a robust feature extractor should embed stylized features adjacent to original ones. Hence, we minimize the distance between the original feature (anchor) and the stylized one in terms of the dot-product similarity. This is accomplished by contrasting stylized features (positives), with other samples (negatives) [17].

In addition, to encourage class discriminability, we adopt a supervised contrastive learning framework [25] where output features from augmented samples and those with the same

class label are treated as positive. Meanwhile, the others in the mini-batch are considered negative. The basic formulation of supervised contrastive learning is defined as:

$$\mathcal{L}_{sup} = - \sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(f'_i \cdot f'_p / \tau)}{\sum_{a \in A(i)} \exp(f'_i \cdot f'_a / \tau)}, \quad (9)$$

where $I \equiv \{1 \dots 2B\}$ indicates a set of indices of features and its stylized augmentation, $A(i) \equiv \{I \setminus i\}$ contains the indices of all samples but the anchor, $P(i)$ denotes the set of indices of all positives to the anchor, and f' denotes an output feature from the feature extractor g after L2 normalization. The softmax function with temperature scaling is applied on the similarity matrix of the anchor.

As shown in Fig. 2 (a), the loss induces the model to attract positive features while repulsing negative ones from the anchor. However, the performance degradation occurs when the loss is directly adopted for the domain generalization task. Concretely, the feature space becomes domain-discriminative since the samples from different domains are pushed aside from the anchor. This is widely known to be detrimental for achieving the domain-invariance [8, 31, 48, 50]. To this end, we propose to modify Eq. (9) to

be more suitable for domain generalization, namely a *domain-aware supervised contrastive loss*:

$$\mathcal{L}_{dsup} = - \sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(f'_i \cdot f'_p / \tau)}{\sum_{a \in P(i) \cup D(i)} \exp(f'_i \cdot f'_a / \tau)}, \quad (10)$$

where $D(i)$ is a set containing the indices of samples sharing the same domain label with the anchor.

As shown in Fig. 2 (b), $D(i)^c \cap P(i)^c$, i.e., samples from the different domain which were included in the earlier negative set, are excluded. For example, when an anchor i is a “photo dog”, its positive set $P(i)$ is {“stylized photo dog”, “cartoon dog”, “sketch dog”, ...}, whereas remaining negatives belongs to different classes in “photo” domain.

Consequently, with the proposed domain-aware supervised contrastive loss, our feature extractor produces features not only discriminative to class labels but also invariant by attracting samples from different domain, i.e., $D(i)^c \cap P(i)$.

3.5 Overall Training and Inference

We train the neural network Φ with the weighted sum of losses as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{cons} \mathcal{L}_{cons} + \lambda_{dsup} \mathcal{L}_{dsup}, \quad (11)$$

where λ_* is the weighting factor. The overall training is conducted in end-to-end manner. The network Φ is updated with respect to \mathcal{L}_{ce} and \mathcal{L}_{cons} , while \mathcal{L}_{dsup} affects only the feature extractor g . During the inference, we detach the feature stylization module from the forward path so that the model predicts based on the original feature.

4 EXPERIMENTS

4.1 Experiment Details

Datasets. As our evaluation benchmarks, we use PACS [28] and Office-Home [13] following conventional settings. PACS is made up of four domains, i.e., Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images) and Sketch (3,929 images). The total number of images is 9,991 and the image resolution is 227×227 . The dataset contains 7 common categories: ‘dog’, ‘elephant’, ‘giraffe’, ‘guitar’, ‘horse’, ‘house’, ‘person’. Another benchmark is Office-Home which consists of images from four different domains, namely Artistic, Clip Art, Product, and Real world images. Each domain contains images of 65 object categories which are found in office and home. The total number of images is 15,500.

Evaluation. A common evaluation protocol in domain generalization is leave-one-domain-out evaluation [28]. Specifically, we first select one domain as the target domain. Then, the other domains are set as source domains. We train our model on the source domains and evaluate it on the target domain. We note that any sample from the target domain is not allowed during the training step. This procedure is repeated to ensure that every domain is chosen to be the source domain exactly once, and we report the averaged accuracy.

Implementation details. For fair comparison with previous studies, we adopt ResNet-18 and ResNet-50 [18] pre-trained on ImageNet [27]. We optimize our network with Stochastic Gradient

Table 1: Quantitative leave-one-domain-out results on PACS. Entries are sorted in the chronological order and separated based on the backbones.

	Method	Accuracy(%)				Avg.
		Photo	Art	Cartoon	Sketch	
ResNet-18	Baseline [28]	95.19	77.87	75.89	69.27	76.56
	D-SAM [12]	95.30	77.33	72.43	77.83	80.72
	MetaReg [2]	95.50	83.70	77.20	70.30	81.68
	JiGen [5]	96.03	79.42	75.25	71.35	80.51
	MASF [11]	94.99	80.29	77.17	71.69	81.04
	Epi-FCR [30]	93.90	82.10	77.00	73.00	81.50
	InfoDrop [46]	96.11	80.27	76.54	76.38	82.33
	DMG [6]	93.35	76.90	80.38	75.21	81.46
	EISNet [54]	95.93	81.89	76.44	74.33	82.15
	L2A-OT [62]	96.20	83.30	78.20	73.60	82.83
	DSON [44]	95.87	84.67	77.65	82.23	85.11
	RSC [23]	95.99	83.43	80.31	80.85	85.15
	MixStyle [64]	96.10	84.10	78.80	75.90	83.73
	pAdaIN [35]	96.29	81.74	76.91	75.13	82.52
	<i>Ours</i>	95.63	85.30	81.31	81.19	85.86
ResNet-50	Baseline [28]	97.66	86.20	78.70	70.63	83.30
	MetaReg [2]	97.60	87.20	79.20	70.30	83.58
	MASF [11]	95.01	82.89	80.49	72.29	82.67
	DMG [6]	94.49	82.57	78.11	75.21	82.60
	EISNet [54]	97.11	86.64	81.53	78.07	85.84
	DSON [44]	95.99	87.04	80.62	82.90	86.64
	RSC [23]	97.92	87.89	82.16	83.35	87.83
	pAdaIN [35]	97.17	85.82	81.06	77.37	85.36
	<i>Ours</i>	96.59	88.48	83.83	82.92	87.96

Descent (SGD) optimizer. We set an initial learning rate as 0.004 and train for 40 epochs. The decay rate is set to 0.0005 which is applied after 20 epochs. A single mini-batch contains a total of 126 images, 42 images for each source domain.

Inspired by FixMatch [47] and SupCon [25], the temperature parameters τ of \mathcal{L}_{cons} and \mathcal{L}_{dsup} are set to 0.5 and 0.15, respectively. Considering the scale of loss functions, weighting factors (λ_{cons} , λ_{dsup}) are set to (0.3, 12) and (0.9, 6) for ResNet-18 and ResNet-50, respectively. Besides, the scale parameters (s_μ , s_σ) are set to 10 and 20 for ResNet-18 and ResNet-50, respectively. Our model is built upon the popular implementation of Zhou et al. [63].

4.2 Comparison with State-of-the art Methods

Results on PACS. In Table 1, we compare our method with previous domain generalization methods on PACS dataset [28]. Recognizably, our method beats previous approaches and achieves a new state-of-the-art performance with the average accuracy of 85.86% with ResNet-18. Consistently, our method shows large improvements when adopting ResNet-50 as our backbone, accomplishing a new record with the average accuracy of 87.86% across the leave-one-domain-out scenarios.

Through the experiments, vivid performance gains are observed when art and sketch domains are set as target domains. This is reasonable since our method has the strength in generating novel

Table 2: Quantitative leave-one-domain-out results on Office-home. Entries are sorted in the chronological order.

	Method	Accuracy(%)				
		Art	Clipart	Product	Real	Avg.
ResNet-18	Baseline [51]	52.15	45.86	70.86	73.15	60.51
	CCSA [33]	59.90	49.9	74.10	75.7	64.90
	D-SAM [12]	58.03	44.37	69.22	71.45	60.77
	CrossGrad [45]	58.40	49.40	73.90	75.80	64.38
	MMD-AAE [31]	56.50	47.30	72.10	74.80	62.68
	JiGen [5]	53.04	47.51	71.47	72.79	61.20
	L2A-OT [62]	60.60	50.10	74.80	77.00	65.63
	DSO [44]	59.37	45.70	71.84	74.68	62.90
	RSC [23]	58.42	47.90	71.63	74.54	63.12
	MixStyle [64]	58.70	53.40	74.20	75.90	65.55
	<i>Ours</i>	60.24	53.54	74.36	76.66	66.20

Table 3: Results of single source domain generalization on PACS. Each row and column indicates the source and target domain, respectively. We report the accuracy with the absolute gain from baseline in brackets. Positive and negative gains are colored green and red, respectively.

	Accuracy (%)			
	(Absolute gain from baseline)			
	Photo	Art painting	Cartoon	Sketch
Photo	99.88 (+0.00)	63.18 (+5.61)	21.84 (+2.43)	54.71 (+28.59)
Art painting	96.53 (-0.06)	99.46 (+0.00)	67.79 (+11.56)	53.59 (+9.60)
Cartoon	84.97 (+0.54)	70.56 (+9.09)	99.57 (+0.12)	70.90 (+8.50)
Sketch	41.56 (+9.16)	43.07 (+13.53)	60.20 (+15.96)	99.36 (-0.08)

styles while preserving the shape cues which are essential for accurate classification in those domains. Despite the slight performance degradation in the photo domain, our method outperforms the competitors in terms of the average accuracy, validating the better domain generalization ability. We note that our feature stylization module does not require additional network parameters, which makes our model more competitive in terms of memory.

Results on office-Home. We also provide the results on Office-Home benchmark [13] in Table 2. Again, our method breaks the record with ResNet-18, achieving the average accuracy of 66.2%. Conspicuously, ours makes the performance improvements regardless of the target domain. Overall comparisons verify the effectiveness of our feature stylization strategy and the proposed contrastive loss.

4.3 Single-source Domain Generalization

In Table 3, we present the single source domain generalization results with ResNet18 on PACS benchmark. In this setting, only a single domain is selected as a source dataset and the trained model is tested on other target domains. Rows and columns indicate

Table 4: Ablation study on each component. † denotes that stylized features are aggregated for cross-entropy loss (\mathcal{L}_{ce}).

Feature Transform	\mathcal{L}_{cons}	\mathcal{L}_{dsup}	Accuracy(%)				
			P	A	C	S	Avg.
			96.29	76.90	77.30	68.81	79.83
✓†			96.11	82.03	80.12	77.52	83.95
✓	✓		95.45	84.18	80.08	79.30	84.75
✓		✓	95.15	80.42	80.59	73.75	82.48
✓	✓	✓	95.63	85.30	81.31	81.19	85.86

source and target domains, respectively. We use the same hyper-parameter settings described in the previous section, except for the scale parameters s_μ , s_σ both of which are scaled down to 5. In addition, since the source domain is single in this setting, $D(i)$ in \mathcal{L}_{dsup} is naturally ignored.

Except for the diagonal elements where train and test domains are the same, we achieve improvements on the most of domain generalization scenarios. We can observe remarkable performance gain on “Photo-to-Sketch”, “Art-to-Cartoon”, “Sketch-to-Art”, and “Sketch-to-Cartoon” settings. Especially, a huge performance improvement is observed when the sketch is used for the source domain, *i.e.*, only coarse shape information is available for training. This demonstrates the style diversification ability of our feature stylization block. The “Art-to-Photo” setting is the only generalization scenario where a performance degradation is observed but still in an acceptable margin.

4.4 Analysis

In this section, we analyze our method and conduct ablation studies on the PACS benchmark with ResNet-18 backbone. To be specific, we first analyze the effect of each loss function, then we investigate the correlation between the performance and the scale parameter s . Thereafter, we examine the most suitable location of the feature stylization block followed by the analysis on the feature decomposition with frequency components. We note that all remaining hyper-parameters are fixed through ablation studies.

Ablation study on components. We conduct an ablation study to inspect the contribution of the feature stylization along with loss functions. As shown in Table 4, every single component enlarges the generalization capacity of the model compared to the baseline. In detail, the effectiveness of feature stylization is observable with the overall performance improvement of $\sim 4.08\%$, in terms of the average accuracy. Specifically, the performance gain on the sketch domain is delightful, reaching $\sim 8.71\%$. In addition, it is verified that the consistency loss boosts domain robustness by regularizing the output discrepancy between original and stylized features. Moreover, the proposed domain-aware contrastive loss enhances the performance by pursuing feature similarities between different domains but with the same category. Consequently, with the harmonious combination of aforementioned components, we can find that all components are complementary and have a positive effect in the domain generalization task.

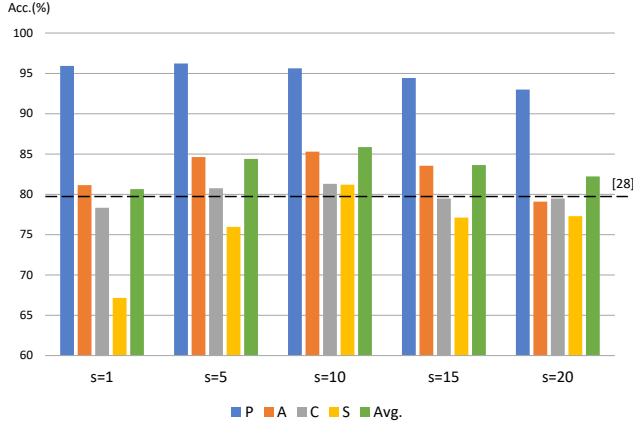


Figure 3: Ablation of scale parameter. The value of scale parameters and the accuracy is on the x and y axis, respectively. We also draw the average accuracy of baseline as a dotted line for better comparison.

The scale parameter. In Fig. 3, we compare different scale parameters in the feature stylization block. We adjust scale parameters (s_μ , s_σ) in $\{1, 5, 10, 15, 20\}$. With the scale parameters of 1, augmented style vectors follow the original style distribution, leading to a marginal improvement. As the scale parameter increases, the outlying style vectors are more likely to be sampled, thus increasing the generalization ability. However, excessive scale values produce too distant features and outputs from the original one, resulting in high favoritism on shape cues. This is undesirable since shape cues are not sufficient for visual recognition and texture cues still contains class-discriminative information as in the human visual system [43]. The best performance is found at the scale parameter of 10, which may be the “sweet spot” of exploiting both shape and textural cues.

Feature decomposition strategy. We verify the effect of decomposing the feature into high frequency and low frequency components. In Table 5, we compare between exploiting whole feature without decomposition (-), high frequency components z^H , and low frequency components z^L for feature stylization. As shown in Table 5, the best performance is achieved when the feature stylization is applied only on low frequency components. Applying stylization on high frequency feature falls behind the other strategies, since only the shape information is partially distorted. Meanwhile, although transforming the whole feature seems to be a good strategy overall, it shows inferior performances on the domains where shape cues are crucial, such as art and sketch domains.

Location of the feature stylization block. We discuss on where the proposed feature stylization block should be located. We denote these stack of residual blocks by re-grouping the ResNet architecture into 5 groups of layers, Conv and ResBlock 1–4, where Conv denotes the first convolutional layer before residual blocks.

As shown in Table 6, the best spot of the proposed feature stylization is right after the second residual blocks. This is quite reasonable considering the nature of deep neural networks [10, 14, 46], where

Table 5: Ablation study on feature decomposition strategies. The column “Frequency component” denotes the component where feature stylization is applied. “-” denotes use of the original feature z , while z^H and z^L indicate the high frequency and low frequency features respectively.

Frequency component	$\mathcal{L}_{cons}, \mathcal{L}_{dsup}$	Accuracy(%)				
		P	A	C	S	Avg.
-		95.45	81.79	79.44	77.47	83.54
z^H		96.05	77.34	78.58	72.73	81.18
z^L (Ours)		96.11	82.03	80.12	77.52	83.95
-	✓	95.87	83.98	81.06	80.30	85.30
z^H	✓	96.41	78.61	79.78	72.25	81.76
z^L (Ours)	✓	95.63	85.30	81.31	81.19	85.86

Table 6: Ablation study on the location of feature stylization block.

Layer	Accuracy(%)				
	P	A	C	S	Avg.
Conv	94.91	81.25	79.10	77.85	83.28
ResBlock ₁	96.11	83.74	80.33	76.02	84.05
ResBlock ₂	95.63	85.30	81.31	81.19	85.86
ResBlock ₃	95.99	82.19	78.67	68.13	81.24
ResBlock ₄	95.39	82.18	79.14	70.77	81.87

features at this level adequately represent low-level structural information as well as high-level semantic information.

5 CONCLUSION

In this paper, we proposed a novel framework for domain generalization, where the features are stylized into diverse domains. In detail, we sampled domain style vectors from the manipulated distribution of batch-wise feature statistics, then utilized the style vectors for affine transformation. To achieve the domain robustness, we exploited stylized features for regularization in terms of output consistency and feature similarity via consistency loss and novel domain-aware supervised contrastive loss, respectively. Through comparisons and extensive analyses on two popular benchmarks, we demonstrated the effectiveness of the proposed feature stylization and two losses.

ACKNOWLEDGMENTS

This research was partly supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program (No. 2021-0-01696) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2003760), and the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-01361: Artificial Intelligence Graduate School Program (YONSEI UNIVERSITY)). This project was also supported by Microsoft Research Asia.

REFERENCES

- [1] Shuang Ao, Xiang Li, and Charles Ling. 2017. Fast generalized distillation for semi-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.
- [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] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1 (2010), 151–175.
- [4] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. *Advances in neural information processing systems* 19 (2007), 137.
- [5] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2229–2238.
- [6] Prithvijit Chattopadhyay, Y. Balaji, and Judy Hoffman. 2020. Learning to Balance Specificity and Invariance for In and Out of Domain Generalization. *ArXiv abs/2008.12839* (2020).
- [7] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryoung Kim, and Jaegul Choo. 2021. RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11580–11590.
- [8] Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).
- [9] Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. 2013. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 668–675.
- [10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*. PMLR, 647–655.
- [11] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. *arXiv preprint arXiv:1910.13580* (2019).
- [12] Antonio D’Innocente and Barbara Caputo. 2018. Domain generalization with domain-specific aggregation modules. In *German Conference on Pattern Recognition*. Springer, 187–198.
- [13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.
- [14] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.
- [15] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*. 2551–2559.
- [16] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*. 529–536.
- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).
- [20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*. PMLR, 1989–1998.
- [21] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. 2021. Fsd: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6891–6902.
- [22] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
- [23] Zeyi Huang, Haohan Wang, E. Xing, and Dong Huang. 2020. Self-Challenging Improves Cross-Domain Generalization. In *ECCV*.
- [24] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.
- [25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems* 33 (2020).
- [26] Myeongjin Kim and Hyeran Byun. 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12975–12984.
- [27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.
- [28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2017. Deeper, Broader and Artier Domain Generalization. 2017 *IEEE International Conference on Computer Vision (ICCV)* (2017), 5543–5551.
- [29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2018. Learning to Generalize: Meta-Learning for Domain Generalization. In *AAAI*.
- [30] Da Li, J. Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. 2019. Episodic Training for Domain Generalization. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1446–1455.
- [31] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5400–5409.
- [32] Yanguo Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. 2016. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779* (2016).
- [33] Saied Motiian, Marco Piccirilli, D. Adjeroh, and Gianfranco Doretto. 2017. Unified Deep Supervised Domain Adaptation and Generalization. 2017 *IEEE International Conference on Computer Vision (ICCV)* (2017), 5716–5726.
- [34] Hyeonseob Nam and Hyo-Eun Kim. 2018. Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks. *Advances in Neural Information Processing Systems* 31 (2018), 2558–2567.
- [35] Oren Nuriel, Sagie Benaïm, and Lior Wolf. 2020. Permuted AdaIN: Enhancing the Representation of Local Cues in Image Classifiers. *International Conference on Learning Representations*.
- [36] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 464–479.
- [37] Fengchun Qiao and Xi Peng. 2021. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6790–6800.
- [38] Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12556–12565.
- [39] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. (2016).
- [40] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *European conference on computer vision*. Springer, 213–226.
- [41] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8050–8058.
- [42] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3723–3732.
- [43] Coralie Sann and Arlette Streri. 2007. Perception of object shape and texture in human newborns: evidence from cross-modal transfer tasks. *Developmental science* 10, 3 (2007), 399–410.
- [44] Seonguk Seo, Yumin Suh, D. Kim, Jongwoo Han, and B. Han. 2020. Learning to Optimize Domain Specific Normalization for Domain Generalization. In *ECCV*.
- [45] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing Across Domains via Cross-Gradient Training. In *International Conference on Learning Representations*.
- [46] Baifeng Shi, Dinghui Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. 2020. Informative dropout for robust representation learning: A shape-bias perspective. In *International Conference on Machine Learning*. PMLR, 8828–8839.
- [47] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685* (2020).
- [48] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
- [49] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 23–30.
- [50] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on*

- computer vision and pattern recognition. 7167–7176.
- [51] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5018–5027.
 - [52] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. 2019. Learning robust global representations by penalizing local predictive power. arXiv preprint arXiv:1905.13549 (2019).
 - [53] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. 2018. Learning Robust Representations by Projecting Superficial Statistics Out. In International Conference on Learning Representations.
 - [54] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and P. Heng. 2020. Learning from Extrinsic and Intrinsic Supervisions for Domain Generalization. In ECCV.
 - [55] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A Fourier-based Framework for Domain Generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14383–14392.
 - [56] Ruijia Xu, Ziliang Chen, W. Zuo, J. Yan, and L. Lin. 2018. Deep Cocktail Network: Multi-source Unsupervised Domain Adaptation with Category Shift. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), 3964–3973.
 - [57] Zhenlin Xu, Deyi Liu, Junlin Yang, and Marc Niethammer. 2021. Robust and generalizable visual representation learning via random convolutions. (2021).
 - [58] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. 2015. Semi-supervised domain adaptation with subspace learning for visual recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2142–2150.
 - [59] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. 2019. Photorealistic Style Transfer via Wavelet Transforms. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019), 9035–9044.
 - [60] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2100–2110.
 - [61] Sergey Zakharov, Wadim Kehl, and Slobodan Ilic. 2019. Deceptionnet: Network-driven domain randomization. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 532–541.
 - [62] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Learning to generate novel domains for domain generalization. In European Conference on Computer Vision. Springer, 561–578.
 - [63] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2020. Domain Adaptive Ensemble Learning. arXiv preprint arXiv:2003.07325 (2020).
 - [64] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021. Domain generalization with mixstyle. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.