Dynamic Style Transferring and Content Preserving for Domain Generalization

Chaoyi Wang[1], Liang Li[2](✉), Yuhan Gao[3], Jiehua Zhang[1], Yefei Zhang[1], Yaoqi Sun[1], Weijun Qin[4], Jun Yin[5], and Zhongyuan Wang[4]

[1] Hangzhou Dianzi University, Hangzhou, Zhejiang, China
{chaoyiwang,jh.zhang,zhangyf,syq}@hdu.edu.cn

[2] Institute of Computing Technology, CAS, Beijing, China
liang.li@ict.ac.cn

[3] Lishui Institute of Hangzhou Dianzi University, Hangzhou, China
yuhangao@hdu.edu.cn

[4] Kuaishou Technology, Beijing, China
{qinweijun,wangzhongyuan}@kuaishou.com

[5] Zhejiang Dahua Technology CO., LTD., Hangzhou, Zhejiang, China
yin_jun@dahuatech.com

Abstract. Although convolutional neural networks (CNNs) have shown remarkable ability in different computer vision tasks, they do not cope well with domain shifts. Recent studies show that the domain shift mainly results from the style or texture variation of images rather than the content. Inspired by this, we propose dynamic style transferring to overcome the style bias of CNNs. Specifically, we design a knowledge-injected attention mechanism for learning adaptive fusion weights and embedding the style knowledge of dynamic chosen images in latent space. So the extent of transferred style is controlled, and we can retain content-related information. Furthermore, we introduce the content preserving module, which builds an adversarial structure with the encoder to make the extracted style information more precise. For balancing the adversarial relationship between encoder and auxiliary predictor, we also introduce a consistency loss to empower the style-biased predictor and indirectly boost the encoder's ability by extending the back-propagation process. We conduct extensive experiments on PACS and Office-Home datasets to evaluate the effectiveness of our method. Experiment results show remarkable performance over the state-of-the-art methods in the domain generalization.

Keywords: Transfer learning · Domain generalization · Style transfer · Content preserving

1 Introduction

In the past few years, Convolutional Neural Networks (CNNs) achieved satisfactory performance in many computer vision tasks with the help of large scale

Y. Chenggang et al. (Eds.): MobiMedia 2022, LNICST 451, pp. 298–315, 2022.
https://doi.org/10.1007/978-3-031-23902-1_23

well-labelled data. However, most CNN models are trained based on the i.i.d. assumption that training and testing data share the same data distribution. In the real world, the hypothesis is not always satisfied, and models often suffer from poor generalization ability in unseen domains. For boosting the performance of CNN models in unseen environments, annotating extensive amounts of data for each scenario to train networks is expensive and unpractical. To solve this problem, Unsupervised Domain Adaptation (UDA) [1–5] is an alternative method without the requirement of extra labelled data.

UDA methods intend to transfer the knowledge learned from the labelled source domain/task to the unlabeled target domain/task. Most UDA methods strive to align the domain distribution by learning domain invariant features of the source and target domains. In general, UDA methods can be divided into two categories, the discrepancy-based method and the adversarial-based method. The discrepancy-based methods alleviate the domain discrepancy by minimizing some predefined statistic metrics between the source and target domains in a high dimensional space. For aligning domain distributions, adversarial-based methods optimize the feature encoder in an adversarial training paradigm. Despite the success of UDA methods, unlabeled target domain data is still required for aligning domain distributions. Therefore the target domain is fixed during the training process, and models may suffer from poor generalization ability in environments out of training distribution.

As a transfer learning method with relaxed data constraints, Domain Generalization (DG) [6–8] intends to boost the generalization ability of models in arbitrary domains out of source distribution with only labelled source data. The existing DG approaches have tried to learn invariant features across multiple domains by minimizing feature divergences between the source domains [1,9–12], normalizing domain-specific gradients based on meta-learning [7,13–16], robust optimization [17–20], or augmenting source domain examples [6,21–25]. For example, Zhou *et al.* [25]randomly select two instances from different domains and adopt a probabilistic convex combination between instance-level feature statistics of bottom CNN layers. Despite the success of the above methods in mitigating the domain shift, they only reduce the domain gap in an ambiguous manner, which lacks specific optimizing orientation upon the cause of domain shifts. Recent studies [22,26] show that the domain shift mainly results from the style or texture variation of images rather than the content.

Motivated by this observation, Generative Adversarial Networks (GANs) are adopted by many researchers to reduce domain gaps by transferring image appearances. Nevertheless, the GAN-based approaches usually contain large scale parameters, which is time-consuming and hard training. Recently, normalization methods (e.g. BN [27], IN [28], CIN [29]) attracted increasing attention for style transferring as it's efficiency. One of the popular approaches is AdaIn [26], which is proposed to transfer image styles by normalizing feature statistics. Inspired by this work, many normalization-based domain generalization methods are proposed, including SagNet [22] and CrossNorm [30]. SagNet [22] provides a new idea for mitigating the difference between domain distributions, which dis-

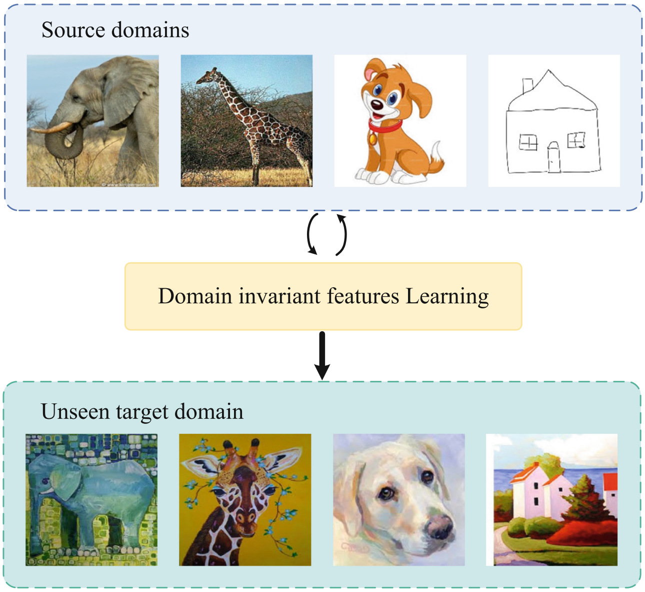

Fig. 1. Classification task about multi-source domain generalization. Given labelled data sampled from several source domains, training the model to learn domain invariant features. Then apply the model to an unseen target domain.

entangles style features from categories information and exchanges statistics of features randomly to prevent style biased predictions and focus more on the contents. Similarly, the CrossNorm also proposes to exchange channel-wise statics between features for enlarging the training distribution. Although both methods improved the model generalization ability under image style shifts, they cannot control the extent of the statistics exchange. Therefore, they tend to ignore some content-related information or pay too much attention to trivial information such as one-sided style features (Fig. 1).

To solve the above problem, we propose the dynamic style transferring and content preserving for domain generalization, which makes the extent of transferred style controllable and reduces the intrinsic style-bias of CNNs in an adversarial learning paradigm. Specifically, we first design a knowledge-injected attention mechanism to learn weight vectors for adaptively fusing style knowledge of mini-batch instances in feature space. This enables an adaptive style integration to capture content-related information hidden in the style knowledge. Second, we introduce the dynamic content preserving module by building an adversarial learning paradigm between the feature encoder and the auxiliary predictor to make the extracted feature irrelevant to the image appearance. Further, in order to enable the encoder to learn the content-biased representation in a long phase, we impose a content consistency loss to boost the optimizing of auxiliary classifier during the minimax game. Therefore, our approach can mitigate the inherent style bias of CNNs by capturing the content-biased representation.

Our main contributions can be summarized as follows:

– We propose a domain generalization method with dynamic style transferring and content preserving, which makes the extent of transferred style controllable and overcomes the intrinsic style bias of CNNs in an adversarial learning paradigm.
– We introduce consistency loss to balance the encoder and auxiliary predictor, which build an adversarial structure. Therefore the encoder can learn the content-biased representation in an extended phase by the back-propagation.
– We conduct extensive experiments on three widely used domain generalization benchmarks, including PACS and Office-Home. The results demonstrate that our method achieves comparable performance to state-of-the-art methods.

2 Related Work

Unsupervised Domain Adaptation (UDA) tackles the domain shift problem where labelled source data and unlabelled target data are available for training. Most UDA methods derive from the point of view of reducing the domain gap between different domains. Maximum Mean Discrepancy (MMD) [31] is an important statistic metric to mitigate the domain distribution shift in previous works. Ganin *et al.* [1] introduced Domain Adversarial Neural Network (DANN) to align the feature distributions. DDC [32] was proposed to alleviate the domain discrepancy by adding adaptation layers for matching high-order moments of feature distributions. Saito *et al.* proposed MCD [33] by devising a domain discriminator to learn domain-invariant features in an adversarial manner for bridging the domain gaps. Inspired by the image translation idea of CycleGAN [34], some methods [35–37] translate the target style into source images to close the domain gap at the image level.

Domain generalization aims to make the model more robust against unseen domains with only access to the source data. Similar to domain adaptation, some multi-source domain generalization works utilize domain alignment methods to minimize the domain discrepancy among source domains for learning domain invariant features. These methods [38,39] argue that feature distributions aligned among source domains should also be robust to unseen target domains. The popular feature aligning methods include minimizing Maximum Mean Discrepancy (MMD) [39,40], minimizing the KL divergence [41] and adversarial learning [9,42]. The works [43,44] design the model with certain parts for learning domain-specific and domain shared representations. For instance, Chattopadhyay *et al.* [43] proposed to learn a balance of domain-invariant and domain-specific features by domain-specific masks. Thus both in-domain and out-of-domain generalization performance are improved. Work [45] proposed to iteratively discard the dominant features activated on the training data, and encourage the network to activate remaining features that correlates with labels. The domain generalization method we proposed is more lightweight compared to previous methods and adaptive learning is performed to enable a controlled degree of style transformation.

Normalization plays a vital role in deep neural network training and image style transferring. Ioffe *et al.* [27] introduced a Batch Normalization (BN) layer to speed up the convergence of models and alleviate the "gradient diffusion" problem in deep networks by normalizing the feature statistics. Batch Normalization is a benchmark technology that has inspired many following normalization methods [26,28,29,46,47]. Ulyanov *et al.* [28] found that significant improvement could be achieved simply by replacing batch normalization with instance normalization. Dumoulin *et al.* [29] proposed a Conditional Instance Normalization (CIN) to learn different affine parameters for different styles. Nevertheless, CIN cannot be adapted to arbitrary new styles without re-training the model. Adaptive Instance Normalization (AdaIN) [26] enables arbitrary style transfer in real-time by aligning the mean and variance of the content features with those of the style features. Compared to BN, IN and CIN, AdaIN adaptively computes affine parameters from style inputs to achieve arbitrary style transfer.

3 Methodology

In this section, we elaborate our dynamic style transferring and content preserving for multiple/single source domain generalization. We first review the background knowledge about instance normalization and style transferring in Sect. 3.1. Secondly, we overview the basic paradigm of domain generalization and the proposed method in Sect. 3.2. Thirdly, we detail the dynamic style transferring and content preserving modules in Sect. 3.3 and Sect. 3.4, respectively.

3.1 Preliminary

As revealed by recent studies [48–50], CNNs are sensitive to the style of images extracted from domains with different data distributions. For inducing the intrinsic style bias of CNNs, both GAN-based and instance normalization-based methods [26,28,29] are proposed. Compared with GAN-based methods, instance normalization-based methods are more efficient and easy inserted into other methods. They [26,51] utilize the channel-wise mean and standard deviation as style representation and transfer image styles by normalizing feature statistics.

Let $x \in \mathbb{R}^{B \times C \times H \times W}$ denotes a batch of feature maps, where B, C, H and W indicate the dimension of batch, channel, height and width, respectively. Instance normalization transforms the normalized feature map, which is formulated as,

$$\mathrm{IN}(x) = \gamma \frac{x - \mu(x)}{\sigma(x)} + \beta, \tag{1}$$

where $\gamma, \beta \in \mathbb{R}^{C}$ are learnable parameters of the affine transformation. And $\mu(x), \sigma(x) \in \mathbb{R}^{B \times C}$ indicate the mean and standard deviation of each feature map at the spatial dimension within the channel according to Eq. 2 and 3, respectively.

$$\mu(x)_{b,c} = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} x_{b,c,h,w} \tag{2}$$

$$\sigma(x)_{b,c} = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{b,c,h,w} - \mu(x)_{b,c}\right)^2} \tag{3}$$

Inspired by this, some researchers [26,28,29] utilize the mean and the variance of the features for style transferring. Huang *et al.* [26] propose Adaptive Instance Normalization (AdaIN) for arbitrary style transfer by recombining the mean and variance of the content features with those of the style features.

$$\mathrm{AdaIN}(x) = \sigma(y)\frac{x - \mu(x)}{\sigma(x)} + \mu(y), \tag{4}$$

where x, y denote the content feature and the target style feature, respectively.

3.2 Overview of Proposed Method

In terms of data availability for training, we assume that we have only access to the source instances $x_s \in X_s$ and the corresponding labels $y_s \in Y_s$ from the source distribution $p_s(x, y)$. Under the scenario of domain generalization, the target image and label are not available for training. Our main goal is to train a neural network on the source domain and generalize well to unseen target domains by inducing the style bias of network.

As illustrated in Fig. 2, our framework consists of three sub-modules, i.e. the shared feature encoder, dynamic style transferring with a style-agnostic classifier and content preserving module with an auxiliary predictor. The shared encoder E extracts features of instances for predicting the categories of inputs. As the target data is unavailable in the training phase, we adopt the dynamic style transferring module to enrich the style information by knowledge-injected attention mechanism for learning content invariant representations. The style-agnostic classifier G_s is supervised by a task loss L_s for accurately predicting image classes. Meanwhile, the dynamic content preserving module builds an adversarial structure between the encoder and auxiliary predictor G_a. The minimax game between them encourages the encoder to generate less style-biased representations. Furthermore, we design a content consistency loss for balancing the adversarial relationship between encoder E and auxiliary predictor G_a. And it also makes the content features extracted by the encoder more style irrelevant. It is worth noting that we only employ the encoder E and auxiliary predictor G_a during the evaluation stage.

3.3 Dynamic Style Transferring

In domain generalization, the alignment of domain distributions plays a significant role. Recent studies [48–50] show that domain distribution shifts mainly result from the style or texture variation of images. Inspired by these studies, SagNet [22] proposed style randomization to reduce this gap so that the encoder can focus on capturing content-related features. However, the extent of

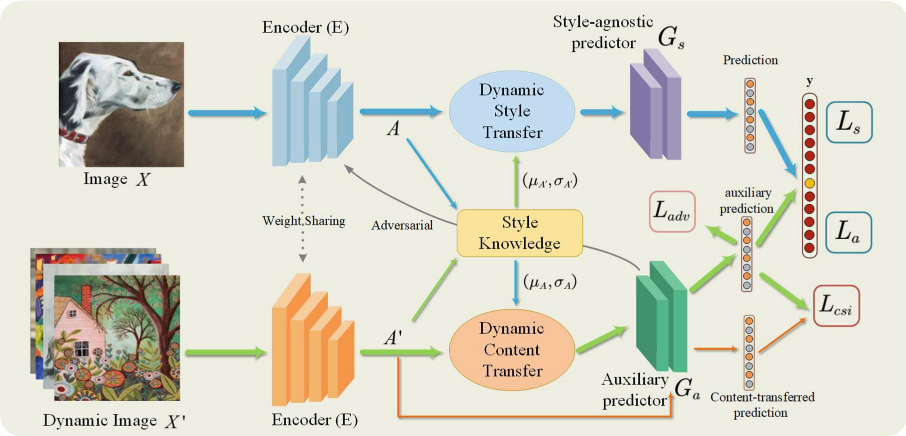

Fig. 2. An illustration of our method, which consists of three sub-modules, including the shared feature encoder, dynamic style transferring with style-agnostic classifier, and dynamic content preserving module with auxiliary predictor. The dynamic style transferring module leads the style-agnostic classifier to focus on content information in the feature map. The dynamic content preserving module guides the auxiliary classifier to focus on the style information, while adversarial learning makes the feature extractor generate less style-related representation.

statistics exchanges is uncontrollable. And they tend to ignore some content-related knowledge hidden in style features. To solve this problem, we propose the dynamic style transferring (DST) module, which can control the extent of transferred knowledge by introducing a knowledge-injected attention mechanism. This mechanism helps achieve the goal of adaptively fusing style knowledge of mini-batch instances in latent space by learning weight vectors. This enables an adaptive style integration to capture content-related information hidden in the style knowledge.

Our knowledge-injected attention mechanism is inspired by the channel-attention mechanism [52], and we implement the attention function $a(\cdot,\cdot)$ by a single linear layer to process both the mean and standard deviation of feature maps. Given a training image x and a dynamic selected image x', we extract their intermediate feature maps by $E(x) = \mathbf{A}, E(x') = \mathbf{A}' \in \mathbb{R}^{D\times H\times W}$ from the encoder E, where H and W indicate spatial dimensions, and D is the number of channels. Then we calculate the statistic $\mu_A, \sigma_A \in \mathbb{R}^D$ as the style representation by Eq. 5 and 6:

$$\mu(A)_{b,c} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} A_{h,w}; \tag{5}$$

$$\sigma(\mathbf{A}) = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(\mathbf{A}_{hw} - \mu(\mathbf{A}))^2}. \tag{6}$$

The dynamic style transferring constructs the transferred style knowledge $\mu, \sigma \in \mathbb{R}^D$ through the knowledge-injected attention mechanism based on A and A':

$$\mu = a\left(\mu_A, \mu_{A'}\right); \tag{7}$$

$$\sigma = a\left(\sigma_A, \sigma_{A'}\right), \tag{8}$$

where $a(\cdot, \cdot)$ denote our attention function, and $\mu_{A'}, \sigma_{A'}$ indicate the channel-wise statics of A'.

Then we implement style transfer by conducting affine transformation on normalized A:

$$\mathrm{DST}(A, A') = \sigma \cdot \frac{A - \mu_A}{\sigma_A} + \mu. \tag{9}$$

The transferred feature map is fed to G_s for a content-biased prediction, and we obtain the cross-entropy loss L_s to jointly optimize $\mathbf{E}$ and $\mathbf{G}_\mathrm{s}$:

$$\underset{\mathbf{E}, \mathbf{G}_\mathrm{s}}{arg\ \min} L_\mathrm{s} = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}) \in S} \sum_{k=1}^{K} \mathbf{y}_k \log \mathbf{G}_\mathrm{s}\left(\mathrm{DST}\left(A, A'\right)\right)_k, \tag{10}$$

where K is the number of classification categories, $\mathbf{y} \in \{0, 1\}^K$ is the one-hot label for input x and $S = \{X_s, Y_s\}$ is the training set.

The introduction of the knowledge-injected attention mechanism re-calibrates the statistics to control fusion extents and increases the diversity of style combinations by adjusting the injecting weights. Therefore the style-agnostic classifier G_s can be robust against the style change and predicts categories based on the content information.

3.4 Dynamic Content Preserving

In addition to learn a style-agnostic predictor, we design the dynamic content preserving module to further mitigate the style bias of network by learning style-related feature representations. Concretely, we first construct an adversarial structure between feature encoder E and auxiliary predictor G_a to constrain the encoder to learn content-agnostic features. In other words, the style knowledge would be captured by the feature encoder to contain as little content information as possible. To achieve this goal, G_a is encouraged to make auxiliary decisions according to the content preserved features $DCP(A, A')$ by L_a. Besides, the predictor tries to predict x accurately, which adversarially makes the encoder capture the discriminative representations. Lastly, we impose a content consistency loss L_{csi} to balance the adversarial structure in the minimax game and preserve the ability of the encoder to encode content-related feature.

In contrast to the dynamic style transferring module, this module retains the style knowledge of feature map A and replaces its content with a dynamically selected representation A' as,

$$\mathrm{DCP}(A, A') = \sigma_A \cdot \frac{A' - \mu_{A'}}{\sigma_{A'}} + \mu_A. \tag{11}$$

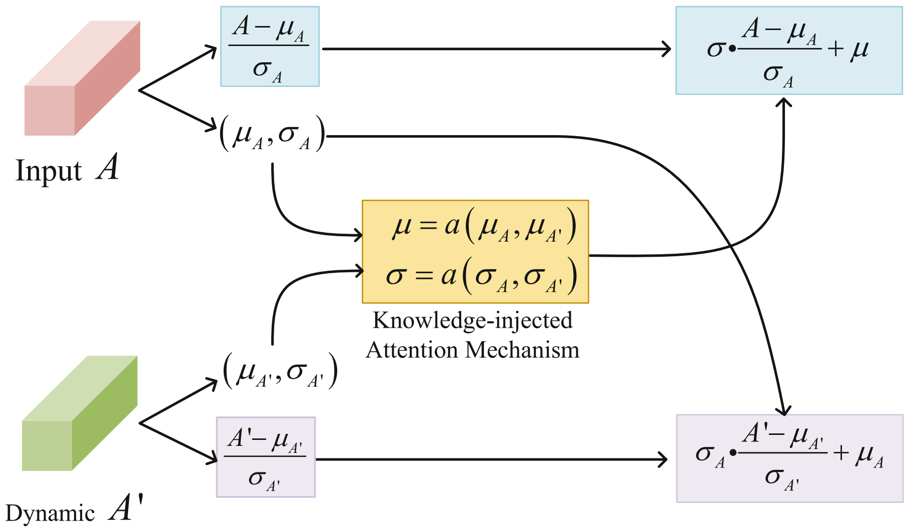

Fig. 3. Illustration of the knowledge-injected attention mechanism. It learns adaptive fusion weights and embedding the style knowledge of dynamic chosen images in latent space.

Figure 3 illustrates the process of the dynamic content preserving. Once the content transformation is finished, the transferred feature maps are fed into the auxiliary classifier G_a to compute the auxiliary predictions. We employ a cross-entropy loss L_a to optimize $\mathbf{G}_\mathrm{a}$:

$$\underset{\mathbf{G}_\mathrm{a}}{arg\min} L_\mathrm{a} = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\in S} \sum_{k=1}^{K} \mathbf{y}_k \log \mathbf{G}_\mathrm{a} \left(\mathrm{DCP}\left(A, A'\right)\right)_k, \tag{12}$$

where $\mathbf{y}_k$ is the label of the instance x, and the optimization goal of the loss function L_a is to promote $\mathbf{G}_\mathrm{a}$ make correctly predictions based on the content transferred feature of x $(DST(A, A'))$. On the other hand, we train encoder E to fool G_a by minimizing an adversarial loss L_adv as follows.

$$\underset{\mathbf{E}}{arg\min} L_\mathrm{adv} = -\lambda_\mathrm{adv} \mathbb{E}_{(\mathrm{x},\cdot)\in S} \sum_{k=1}^{K} \frac{1}{K} \log \mathbf{G}_\mathrm{a} \left(\mathrm{DCP}\left(A, A'\right)\right)\right)_k \tag{13}$$

where λ_adv is a hype-parameter for adjusting the adversarial extent.

Although the two networks and objective functions construct an adversarial structure, the capability between them may not be balanced as the discriminator is weak, which results in the learning of the encoder being terminated in early phase. Essentially, the improvements of generators or encoders come from the gradient back-propagation of the discriminator's loss. When the predictor is fooled easily by generators, the marginal cost in the later training phase is insufficient to drive the generator to jump out the local optimum point in

Algorithm 1. Training algorithm of our method.

Input: training data $S = (x_i, y_i)_{i=1}^{M}$; batch size$N$;
Initialize: feature extractor $\mathbf{E}$; style-agnostic classifier $\mathbf{G}_s$; auxiliary classifier $\mathbf{G}_a$
While not converged $\mathbf{X}, \mathbf{Y} = Minibatch(S, N)$
$\mathbf{A} = \mathbf{E}(\mathbf{X})$
$\mathbf{A'} = SHUFFLE(\mathbf{A})$
$\mathbf{A}^s = DST(\mathbf{A}, \mathbf{A}')$

$\underset{\mathbf{E},\mathbf{G}_\mathrm{s}}{arg} \min L_s = -\frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{K} \mathbf{Y}_{j,k} \log \mathbf{G}_\mathrm{s}(A_j^s)_k$

$\mathbf{A}^c = DCP(\mathbf{A}, \mathbf{A}')$

$\underset{\mathbf{G}_\mathrm{a}}{arg} \min L_a = -\frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{K} \mathbf{Y}_{j,k} \log \mathbf{G}_\mathrm{a}(A_j^c)_k$

$\underset{\mathbf{E}}{arg} \min L_\mathrm{adv} = -\lambda_\mathrm{adv} \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{K} \frac{1}{K} \log \mathbf{G}_\mathrm{a} \left(\mathbf{A}_j^\mathrm{c}\right)_k$

$\underset{\mathbf{E},\mathbf{G}_\mathrm{a}}{arg} \min L_\mathrm{csi} = -\lambda_\mathrm{csi} \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{K} \|\mathbf{G}_\mathrm{a}(A_j^c) - A_j^c\|_k^2$
end
Output: $\mathbf{E} \circ \mathbf{G}_\mathrm{s}$

optimization space. Motivated by this, we impose the consistency loss L_{csi} to empower the auxiliary classifier, which minimizes the mean square error between auxiliary predictions and content preserving predictions. Therefore, it not only promotes the adversarial relationship between encoder and auxiliary predictor, but also makes the encoder learn content-biased representations in an extended phase and protects the encoder's ability to capture content-related features. The above consistency loss is computed as,

$$\underset{\mathbf{E},\mathbf{G}_\mathrm{a}}{arg} \min L_\mathrm{csi} = \lambda_\mathrm{csi} \mathbb{E}_{\mathbf{x} \in S} \sum_{k=1}^{K} \left\{\mathbf{G}_\mathrm{a} \left(\mathrm{DCP}\left(A, A'\right)\right)_k - \mathbf{G_a}\left(A'\right)_k\right\}^2 \tag{14}$$

where λ_csi is the weight coefficient which controls balance of the adversarial game. We analyze the influence of different λ_csi value in ablation study (Sect. 4.3).

4 Experiments

In this section, we conduct extensive experiments to validate the effectiveness of our methods on three widely-used benchmarks, including PACS [44] and Office-Home [53]. We first introduce the datasets and implementation details. Then we analyse the experimental results on datasets and discuss the ablation study. Finally, we analyse the visualization result domain alignment of the proposed method.

4.1 Datasets and Implementation Details

In this section, we first introduce three widely used benchmarks, including PACS [44] and Office-Home [53]. Then we describe the implementation details of our method.

PACS is a domain generalization dataset that consists of 9991 images across four domains, namely Photo, Art Painting, Cartoon and Sketch. Each domain contains seven categories. Following the official split [44], we split the data of PACS into 70% training and 30% validation.

Office-Home is a benchmark dataset for domain adaptation that contains four domains, including Art, Clipart, Product and Real-World. Each domain consists of 65 categories with an average of about 70 instances per class. We split the 15588 instances of Office-Home into 90% training and 10% validation following [8].

Implementation Details. The proposed method is implemented on PyTorch and trained on a single NVIDIA RTX 2080TI GPU. During the training, we adopt the Stochastic Gradient Descent (SGD) optimizer with a weight decay of 0.0001, momentum of 0.9 and an initial learning rate of 0.0004 for all datasets. We adopt the cosine scheduling for the learning rate adjusting, and the adjusting iterations is 2k. The adversarial weight are fixed to 0.1. Limited by the GPU memory, we set the training batchsize to 96 on PACS and 32 on others.

4.2 Results on PACS

In this section, we conduct the experiments of single-source domain generalization and multi-source domain generalization on PACS. We employ the ResNet-18 [54] as our feature extractor for all experiments on the PACS dataset.

For demonstrating the effectiveness of our method under single-source domain generalization setting, we conduct experiments on the PACS dataset where only a single source domain data is accessible. We train our network on each domain of PACS and validate the model on the remaining domains. As can be seen in Table 1, our model outperforms the state-of-the-art method by a large margin in most domains. Specifically, we outperform JiGen [6] by 11.7% and higher than Sagnet [22] by 4.4%, in average. The reason may be that the proposed dynamic style transferring significantly benefits the robustness of CNNs against the appearance variation. And the content preserving module further improves the performance by balancing the adversarial learning between the encoder and auxiliary classifier.

Furthermore, for demonstrating the effectiveness of our method under multi-source domain generalization task, we compare it with recent works, including Epi-FCR [15], D-SAM [8], JiGen [6], MASF [7], MMLD [10], RSC [45], StableNet [56] and SagNet [22]. The results of RSC and StableNet are from the original paper, and the results of other methods in Tabel 2 are copied from [22]. The experimental comparison results are shown in Table 2. We observe that the proposed method achieves competitive performance against all the methods with 0.45–7.64% improvements in average accuracy, proving our method is robust to

Table 1. Performance comparison at the single-source domain generalization on PACS (A: Art Painting, C: Cartoon, S: Sketch, P: Photo).

Method	A→C	A→S	A→P	C→A	C→S	C→P	S→A	S→C	S→P	P→A	P→C	P→S	Avg.
ResNet-18 [54]	62.3	49.0	95.2	65.7	60.7	83.6	28.0	54.5	35.6	64.1	23.6	29.1	54.3
JiGen [6]	57.0	50.0	96.1	65.3	65.9	85.5	26.6	41.1	42.8	62.4	27.2	35.5	54.6
ADA [55]	64.3	58.5	94.5	66.7	65.6	83.6	37.0	58.6	41.6	65.3	32.7	35.9	58.7
SagNet [22]	67.1	56.8	95.7	**72.1**	69.2	85.7	41.1	62.9	46.2	**69.8**	35.1	40.7	61.9
Ours	**67.2**	**62.4**	**96.3**	67.8	**69.7**	**87.7**	**59.0**	**65.1**	**54.3**	67.2	**41.3**	**56.3**	**66.3**

Table 2. Performance comparison at multi-source domain generalization on PACS.

Method	Venue	Art painting	Cartoon	Sketch	Photo	Avg.
Epi-FCR [15]	AAAI 2018	82.10	77.00	73.00	93.9	81.50
D-SAM [8]	ICPR 2018	77.33	72.43	77.83	95.30	80.72
JiGen [6]	CVPR 2019	79.42	75.25	71.35	96.03	80.51
MASF [7]	NIPS 2019	80.29	77.17	71.69	94.99	81.04
RSC [45]	ECCV 2020	75.72	68.50	66.10	93.93	76.06
MMLD [10]	AAAI 2020	81.28	77.16	72.29	96.09	81.83
SagNet [22]	CVPR 2021	**83.58**	**77.66**	76.30	95.47	83.25
StableNet [56]	CVPR 2021	80.16	74.15	70.10	94.24	79.66
Ours	–	82.32	76.62	**80.00**	**95.87**	**83.70**

the style variations. Particularly, our approach brings significant improvement on the Sketch domain, with a maximum improvement of 13.9% and a minimum improvement of 2.17%. The improvement demonstrates that the dynamic style transferring and content preserving are effective in reducing the intrinsic style bias of the feature extractor. Therefore the generalization ability of our model against style variations is improved. Besides, we also find SagNet [22] exceed our model in Art Painting and Cartoon domains. We attribute this inferior performance to two aspects. On the one hand, our method may not be as stable as other methods due to the hype parameters are not fine-tuned. On the other hand, we may suffer from a relatively high source risk in mussy backgrounds such as Art painting, as our method performs well in Sketch and Photo domains which have salient objects and brief background.

We also notice that the performance of our model decreases sharply under the single-source DG setting, compared to the results of multi-source DG in Table 2. On the one hand, this demonstrates the number of training instances plays a vital role in DG, and the single-source DG is a challenging task. On the other hand, this validates that the rich variations of domain style benefit our method to capture and induce the style bias as the multi-source experiments are conducted on mixed data.

4.3 Ablation Study

In this section, we verify the effectiveness of the proposed dynamic style transferring (DST) and content preserving (DCP) under a single source domain gener-

Table 3. Ablation study of our method on PACS for single-source domain generalization (A: Art Painting, C: Cartoon, S: Sketch, P: Photo). λ_{csi} is the coefficient of content consistency loss.

Method	A→C	A→S	A→P	C→A	C→S	C→P	S→A	S→C	S→P	P→A	P→C	P→S	Avg.
Baseline	67.1	56.8	95.7	**72.1**	69.2	85.7	41.1	62.9	46.2	**69.8**	35.1	40.7	61.9
DST w/o DCP	63.7	**65.1**	**96.8**	70.1	**72.2**	**89.9**	43.7	62.1	53.1	68.8	32.9	50.0	64.0
DST w/DCP ($\lambda_{\mathrm{csi}} = 1$)	66.2	61.2	96.3	66.0	70.3	86.8	41.2	64.3	47.7	67.2	**41.6**	48.8	63.1
DST w/DCP ($\lambda_{\mathrm{csi}} = 0.01$)	66.8	60.1	95.9	70.2	69.4	89.0	50.7	62.1	**54.5**	66.5	40.8	52.2	64.8
DST w/DCP ($\lambda_{\mathrm{csi}} = 0.1$)	**67.2**	62.4	96.3	67.8	69.7	87.7	**58.9**	**65.1**	54.3	67.2	41.3	**56.3**	**66.3**

alization setting on PACS, and we also show the influence of content consistency loss in Eq. 14.

We adopt SagNet [22] as our baseline. First, we validate the effect of the proposed DST based on the baseline (denoted as 'DST w/o DCP' in Table 3), where the content consistency loss is not involved into the optimizing process. Then we balance the adversarial learning between feature encoder and auxiliary predictor by introducing the content consistency criterion. We explore the consistency loss by varying the coefficient λ_{csi} in Eq. 14 while the DST module is fixed.

As illustrated in Table 3, all variants of our method significantly surpass the baseline [22] in average accuracy. The comparison between baseline and 'DST w/o DCP' demonstrates that the adaptive style fusion weights can guide the model to concentrate on useful information by dynamically adjusting the extent of style knowledge transferring. As can be seen in the second and third rows of Table 3, the performance of 'DST w/ DCP($\lambda_{\mathrm{csi}} = 1$)' is worse than original 'DST w/o DCP'. Such this accuracy degeneration may results from the imbalance between feature extractor and auxiliary predictor. The large coefficient of content consistency loss leads to the encoder failing in the minimax game, and the style bias of CNNs impacts the extraction of domain invariant features.

The last three rows of Table 3 show the influence of hype parameters λ_{csi}. We observe that the best performance occurs in 'DST w/ DCP ($\lambda_{\mathrm{csi}} = 0.1$)'. And the performance of 'DST w/ DCP ($\lambda_{\mathrm{csi}} = 0.01$)' surpasses 'DST w/ DCP ($\lambda_{\mathrm{csi}} = 1$)' with 1.7% in average accuracy. Particularly, 'DST w/ DCP ($\lambda_{\mathrm{csi}} = 0.01$)' exceed 'DST w/ DCP ($\lambda_{\mathrm{csi}} = 1$)' in task $S \rightarrow A$ by 9.5%. We attribute such phenomenon to the balance of adversarial training process between encoder and auxiliary predictor. The large coefficient of content consistency leads to the encoder failing to reduce the style bias. On the contrary, a small coefficient makes the encoder surpass the predictor in early-stage and unable to learn further via the adversarial process. Therefore a proper coefficient is necessary for the content preserving loss which intends to balance the minimax two-player game. And the encoder can preserves the content feature benefiting from the DCP. Finally, we set DCP($\lambda_{\mathrm{csi}} = 0.1$) in the following experiments.

Table 4. Performance comparison at single-source domain generalization on Office-Home (A: Art , C: Clipart, P: Product, R: Real_World).

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
ResNet-50 [54]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
ERM [61]	40.0	47.7	57.9	39.0	51.8	52.3	37.4	30.3	56.4	53.3	41.5	69.7	48.1
ARM [57]	41.0	48.7	63.1	36.4	45.6	47.7	44.6	32.8	56.9	51.8	42.6	69.2	48.4
Fish [20]	40.5	54.4	63.1	41.5	52.8	58.5	41.0	36.4	66.7	53.8	41.5	73.8	52.0
SD [58]	45.6	57.4	66.2	43.6	53.3	52.8	40.0	34.4	65.1	56.9	47.7	72.8	53.0
DAN [2]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [1]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	**63.2**	51.8	76.8	57.6
SagNet [22]	48.7	61.0	70.3	48.7	55.4	62.1	50.8	45.6	69.2	62.6	**54.9**	**76.9**	58.8
Ours	**49.7**	**62.6**	**73.8**	**50.8**	**59.0**	**64.1**	**52.3**	**46.7**	**73.3**	63.1	48.2	**76.9**	**60.0**

4.4 Results on Office-Home

In this section, we further compare our method with recent state-of-the-art works [1,2,20,22,54,57,58] under single-source domain generalization setting on the Office-Home dataset. The results of ResNet-50 [54], DAN [2] and DANN [1] are copied from [59], and the results of other methods in Tabel 4 are reproduced from the code in DomainBed library [60].

The comparison results are shown in Table 4. We can observe that the proposed method exceeds all of the comparison approaches in most tasks and averages accuracy by a large margin. In particular, our approach achieves 4.1% gain on $P \rightarrow R$ task and 1.2% gain on average, compared to the state-of-the-art SagNet [22] on accuracy. We note that SagNet exchanges the style statistics in a random interpolation manner. Therefore, they tend to ignore some content-related information or pay too much attention to trivial information such as one-sided style features which are unrelated to the prediction. Besides, the adversarial training in SagNet is also unbalanced, which results in the back-propagation disabled in the later phase.

5 Conclusion

This paper proposes the dynamic style transferring and content preserving to alleviate the style bias of CNNs. Concretely, we design a knowledge-injected attention mechanism to control the extent of embedding the style knowledge of dynamic chosen images in latent space. So the content-related information hidden in style knowledge can be retained. Furthermore, we introduce the content preserving module, which builds an adversarial structure with the encoder to make the captured style information more precise. Experiment results show our method achieve remarkable performance over the SOTA methods in the single/multiple source domain generalization.

Limitation. Because of the adversarial relationship between the encoder and the auxiliary classifier, our model suffers from the performance degradation in some sub-tasks. As our approach does not leverage domain labels, it may be significant to further improve the performance under multi-source domain setting by adding a domain discriminator to capture the domain information.

References

1. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2130 (2016)
2. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning, pp. 97–105. PMLR (2015)
3. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. Adv. Neural Inf. Process. Syst. **31** (2018)
4. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1406–1415 (2019)
5. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176 (2017)
6. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2229–2238 (2019)
7. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. Adv. Neural Inf. Process. Syst. **32** (2019)
8. D'Innocente, A., Caputo, B.: Domain generalization with domain-specific aggregation modules. In: Brox, T., Bruhn, A., Fritz, M. (eds.) GCPR 2018. LNCS, vol. 11269, pp. 187–198. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12939-2_14
9. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2018)
10. Matsuura, T., Harada, T.: Domain generalization using a mixture of multiple latent domains. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11749–11756 (2020)
11. Sun, B., Saenko, K.: Deep CORAL: Correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 443–450. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_35
12. Zhao, S., Gong, M., Liu, T., Fu, H., Tao, D.: Domain generalization via entropy regularization. Adv. Neural. Inf. Process. Syst. **33**, 16096–16107 (2020)
13. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: MetaReg: towards domain generalization using meta-regularization. Adv. Neural. Inf. Process. Syst. **31** (2018)
14. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: meta-learning for domain generalization. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
15. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., Hospedales, T.M.: Episodic training for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1446–1455 (2019)
16. Zhang, M.M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., Finn, C.: Adaptive risk minimization: a meta-learning approach for tackling group shift (2020)

17. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
18. Krueger, D., et al.: Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning, pp. 5815–5826. PMLR (2021)
19. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2019)
20. Shi, Y., et al.: Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 (2021)
21. Bai, H., et al.: DecAug: out-of-distribution generalization via decomposed feature representation and semantic augmentation. arXiv preprint arXiv:2012.09382 (2020)
22. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8690–8699 (2021)
23. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745 (2018)
24. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12361, pp. 561–578. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_33
25. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008 (2021)
26. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
27. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
28. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6924–6932 (2017)
29. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
30. Tang, Z., Gao, Y., Zhu, Y., Zhang, Z., Li, M., Metaxas, D.N.: Crossnorm and selfnorm for generalization under distribution shifts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 52–61 (2021)
31. Long, M., Cao, Y., Cao, Z., Wang, J., Jordan, M.I.: Transferable representation learning with deep adaptation networks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 3071–3085 (2018)
32. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
33. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3723–3732 (2018)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
35. Hoffman, J., et al.: CyCADA: cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning, pp. 1989–1998. PMLR (2018)

36. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6936–6945 (2019)
37. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
38. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5715–5725 (2017)
39. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International Conference on Machine Learning, pp. 10–18. PMLR (2013)
40. Ghifary, M., Balduzzi, D., Kleijn, W.B., Zhang, M.: Scatter component analysis: a unified framework for domain adaptation and domain generalization. IEEE Trans. Pattern Anal. Mach. Intell. **39**(7), 1414–1430 (2016)
41. Li, H., Wang, Y., Wan, R., Wang, S., Li, T.Q., Kot, A.: Domain generalization for medical imaging classification with linear-dependency regularization. Adv. Neural. Inf. Process. Syst. **33**, 3118–3129 (2020)
42. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 624–639 (2018)
43. Chattopadhyay, P., Balaji, Y., Hoffman, J.: Learning to balance specificity and invariance for in and out of domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 301–318. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_18
44. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5542–5550 (2017)
45. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 124–140. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_8
46. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
47. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19 (2018)
48. Baker, N., Lu, H., Erlikhman, G., Kellman, P.J.: Deep convolutional networks do not classify based on global object shape. PLoS Comput. Biol. **14**(12), e1006613 (2018)
49. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
50. Hermann, K., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. Adv. Neural. Inf. Process. Syst. **33**, 19000–19015 (2020)
51. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
52. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

53. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5018–5027 (2017)
54. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
55. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. Adv. Neural Inf. Process. Syst. **31** (2018)
56. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z.: Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5372–5382 (2021)
57. Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., Finn, C.: Adaptive risk minimization: learning to adapt to domain shift. Adv. Neural Inf. Process. Syst. **34** (2021)
58. Pezeshki, M., Kaba, O., Bengio, Y., Courville, A.C., Precup, D., Lajoie, G.: Gradient starvation: a learning proclivity in neural networks. Adv. Neural Inf. Process. Syst. **34** (2021)
59. Bucci, S., D'Innocente, A., Liao, Y., Carlucci, F.M., Caputo, B., Tommasi, T.: Self-supervised learning across domains. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 5516–5528 (2021)
60. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. arXiv preprint arXiv:2007.01434 (2020)
61. Volk, G., Müller, S., Von Bernuth, A., Hospach, D., Bringmann, O.: Towards robust cnn-based object detection through augmentation with synthetic rain variations. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 285–292. IEEE (2019)