Computer Vision and Image Understanding 244 (2024) 104016

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Structure-aware feature stylization for domain generalization

Milad Cheraghalikhani [*,1], Mehrdad Noori[1], David Osowiechi, Gustavo A. Vargas Hakim, Ismail Ben Ayed, Christian Desrosiers

LIVIA, ÉTS, Montreal, Quebec, Canada
International Laboratory on Learning Systems (ILLS), Canada

ARTICLE INFO

Communicated by Sifei Liu

MSC:
41A05
41A10
65D05
65D17

Keywords:
Computer vision
Image classification
Domain generalization

ABSTRACT

Generalizing to out-of-distribution (OOD) data is a challenging task for existing deep learning approaches. This problem largely comes from the common but often incorrect assumption of statistical learning algorithms that the source and target data come from the same i.i.d. distribution. To tackle the limited variability of domains available during training, as well as domain shifts at test time, numerous approaches for domain generalization have focused on generating samples from new domains. Recent studies on this topic suggest that feature statistics from instances of different domains can be mixed to simulate synthesized images from a novel domain. While this simple idea achieves state-of-art results on various domain generalization benchmarks, it ignores structural information which is key to transferring knowledge across different domains. In this paper, we leverage the ability of humans to recognize objects using solely their structural information (prominent region contours) to design a Structural-Aware Feature Stylization method for domain generalization. Our method improves feature stylization based on mixing instance statistics by enforcing structural consistency across the different style-augmented samples. This is achieved via a multi-task learning model which classifies original and augmented images while also reconstructing their edges in a secondary task. The edge reconstruction task helps the network preserve image structure during feature stylization, while also acting as a regularizer for the classification task. Through quantitative comparisons, we verify the effectiveness of our method upon existing state-of-the-art methods on PACS, VLCS, OfficeHome, DomainNet and Digits-DG. The implementation is available at this repository.

1. Introduction

Various recent studies (Hendrycks and Dietterich, 2019) have shown the high sensitivity of deep learning models to domain shift, as well as the significant drop in accuracy of such models when tested on out-of-distribution (OOD) data. This is mainly due to the over-simplistic assumption of statistical learning algorithms, such as deep neural networks, that the training (source) and testing (target) data come from the same domain/dataset and that they follow the same independent and identically distribute (i.i.d.) distribution. In practice, this assumption may not hold, and ignoring the OOD nature of test data can lead to catastrophic failure. The problem of *domain generalization* (DG) was introduced (Blanchard et al., 2011) to learn domain shift without having access to samples of the target domain during training. In other words, the DG setup tries to train a model on related but distinct source domains in such a way that the model can perform well on any other unseen target domain at test time.

Early works on DG (Li et al., 2018c; Muandet et al., 2013; Li et al., 2018b) are based on aligning the distributions of source domains with the goal of learning a domain-invariant representation. The motivation behind this technique is that features which are invariant to the source domains should also be robust to shifts in target domains. Despite their initial success, these methods typically suffer from over-fitting to source domains (Zhou et al., 2021a). Recently, several DG approaches proposed to mitigate the limited number of source domains in training, as well as the shift of target domains, by generating samples from new synthetic domains. Based on this idea, feature-based augmentation or *stylization* methods create samples by transforming the latent representation of training examples so that their semantic information (class label) remains the same but also encode styles (e.g., textures, colors, etc.) that are different from those of source domains. A simple yet powerful feature stylization method, called MixStyle (Zhou et al., 2021b), mixes the feature statistics from instances of different source domains to generate novel ones, and then trains a model on the augmented set of samples. In Jeon et al. (2021), this method is enhanced using a domain-aware supervised

* Corresponding author at: LIVIA, ÉTS, Montreal, Quebec, Canada.
E-mail address: milad.cheraghalikhani.1@ens.etsmtl.ca (M. Cheraghalikhani).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cviu.2024.104016
Received 14 April 2023; Received in revised form 8 October 2023; Accepted 10 April 2024
Available online 22 April 2024

contrastive loss which minimizes the cosine distance between features of same-class examples, regardless of their domains, while pushing away same-domain examples from different classes. Although feature mixing approaches like MixStyle achieve state-of-art performance on various domain generalization benchmarks, it ignores the structural information of images which is key to transferring knowledge across different domains.

The method proposed in this paper is inspired by the natural ability of humans to recognize objects using only structural information represented by the contours of prominent regions in the image. For example, a child can recognize a dog from a simple, imperfect drawing as well as from real images of the animal. Following this idea, we design a Structure-Aware Feature Stylization method for DG which improves feature stylization based on instance statistics by enforcing structural consistency across different style-augmented samples. Toward this goal, our method leverages a multi-task learning model that classifies original and style-augmented images while also reconstructing prominent edges in a secondary task. This reconstruction task helps the encoder preserve important structural information during feature stylization and acts as a regularization prior for the classification task. The contributions of our work are the following:

1. We propose a novel feature stylization method for DG which enforces both semantic information (class labels) and structural information (prominent edges) consistency in style-augmented features via a multi-task learning model.
2. Using the proposed method, we shows a robust and consistent improvement in five popular DG benchmarks for classification, PACS (Li et al., 2017), VLCS (Fang et al., 2013), OfficeHome (Venkateswara et al., 2017), DomainNet (Peng et al., 2019), and Digits-DG (Zhou et al., 2020b), outperforming several recent DG approaches.

2. Related works

A wide range of deep learning methods have been proposed to tackle the problem of DG. Recent approaches for this task can be grouped in three broad categories (Zhou et al., 2021a): *Data Augmentation* based methods, *Self-Supervised Learning* based methods and *Disentangled Representation Learning* based methods.

Data Augmentation (DA) In supervised learning, DA is commonly used to regularize the training of over-parameterized neural networks to avoid over-fitting. This well-known technique uses a given set of transformations to augments original training pairs so that their label is preserved. In DG, since the target domain data is not accessible in training, the transformation is applied to simulate domain shifts. This can be achieved in three different ways: (1) learnable augmentation, (2) off-the-shelf style transfer, and (3) feature-based augmentation. The first approach uses an augmentation network to synthesize images from source samples so that the joint distribution of synthesized pairs is different from the one of existing source domains. The classifier is then trained with both source images and synthesized images. Based on this idea, the *Deep Domain-Adversarial Image Generation* (DDAIG) (Zhou et al., 2020a) method trains a domain transformation network such that the class label of transformed images can be recognized but not their domain label. Leveraging a similar approach, ADAGE (Carlucci et al., 2019b) generates images from an agnostic synthetic domain with a Hallucinator network so that the domain cannot be recovered from the augmented image (pixels) nor its extracted features. *Learning to Augment by Optimal Transport* (L2A-OT) (Zhou et al., 2020b) is another learnable augmentation method that generates pseudo-domain images by maximizing the distance between source domains and the new pseudo-domains, as measured by optimal transport (OT). Cycle-consistency and classification losses are used to preserve the semantics and global structure of generated images. Off-the-shelf style transfer approaches for DG exploit the recent advances in style transfer (Huang and Belongie, 2017) and try to map input images from one domain to another domain (Somavarapu et al., 2020) or even to external styles (Yue et al., 2019). As an example, the method in Somavarapu et al. (2020) uses a transformation network based on AdaIN (Huang and Belongie, 2017) and, for each source domain, maps an input image to the target style of randomly selected domain. In contrast to the above-mentioned approaches, which mainly operate on pixels, feature-level augmentation (Mancini et al., 2020; Zhou et al., 2021b) are motivated by the fact that style-related information is captured in statistics of CNN features. MixStyle (Zhou et al., 2021b) introduced a plug-and-play module, inserted between CNN layers, that mixes the feature statistics of two instances with a random convex weight to simulate new styles. The *Feature Stylization and Domain-Aware Contrastive Learning* (Jeon et al., 2021) approach instead supposes that instance-wise statistics come from a normal distribution characterizing the batch. They then compute the batch-wise statistics and sample a new distribution from these. Original features are decomposed into high-frequency and low-frequency components, and feature stylization is only applied on the low frequency one. To encourage semantic consistency, a loss maximizing the agreement between the model prediction for the original and augmented feature maps is also proposed. Unlike this approach, which explicitly adds high-frequency features over stylized low-frequency ones, our method enforces structural consistency in a more flexible way using a secondary reconstruction task.

Self-Supervised Learning (SSL) Methods based on SSL seek to find a good representation by solving a pretext task that does not require any label (e.g., predicting the transformation applied to the image Gidaris et al., 2018 or whether two transformed images come from the same original one Grill et al., 2020). The driving hypothesis of such technique is that the learned representation captures generic but useful features which help learn a downstream task, typically in a fine-tuning step. In DG, SSL methods help avoid over-fitting to domain-specific biases. As an example, Carlucci et al. (2019a) trained an encoder to solve a Jigsaw puzzle problem in addition to a regular classification task, so that the network can learn features that are more generalizable across domains. In Bucci et al. (2021), authors combined jigsaw puzzle solving and rotation prediction tasks to increase the robustness of encoded features to domain shift. Similarly, the DG approach in Albuquerque et al. (2020) combines rotation prediction with the task of predicting responses to Gabor filter banks to improve generalization. While our method also reconstructs prominent edges of the image using a separate task, we do so in a consistency loss, jointly optimized with the classification loss, which preserves the structure of images for different feature-based augmentations.

Distangled Representation Learning (DRL) Instead of forcing the model to learn a domain-invariant representation, DRL methods split it in a domain-specific part and a domain-agnostic part, the latter one used to extract domain-invariant features. In Ilse et al. (2019), the authors train three independent encoders, the first for domain-specific features, the second for class-specific features, and the third for capturing residual variations. The representations of these encoders are used to reconstruct the original input via a Variational Auto Encoder (VAE). Two adversarial classifiers, trying to predicting the domain and class of samples from their representation, are added to disentangle the corresponding features.

3. Method

Our framework follows a multi-task learning approach that improves feature stylization based on instance statistics by enforcing structural consistency across different style-augmented samples. In this section, we first describe the baseline setup of multi-source domain generalization for image classification, then introduce our novel structure-aware feature stylization method.

3.1. Problem definition

For a classification task, denoting the input space as $\mathcal{X}$ and the target space as $\mathcal{Y}$, a *domain* is defined as the joint distribution of P_{XY} on

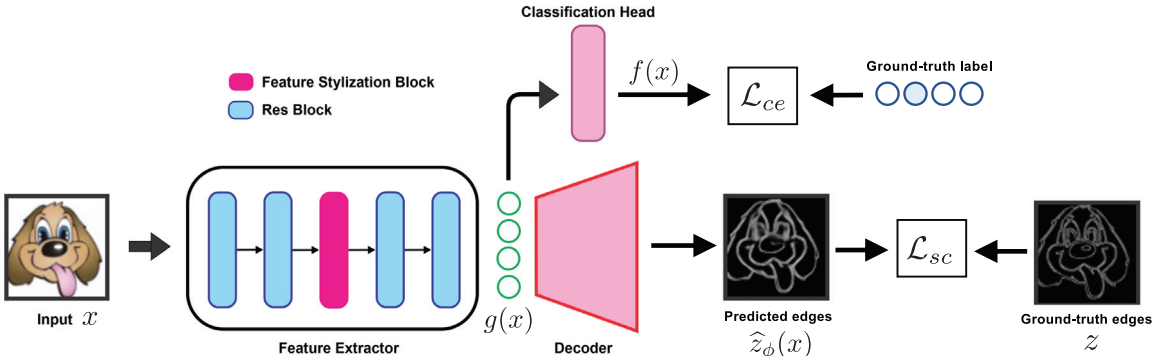

Fig. 1. The overall architecture of the proposed method.

$\mathcal{X}\times\mathcal{Y}$. For a specific domain, we denote as P_X the marginal distribution on $\mathcal{X}$, $P_{Y|X}$ the posterior distribution of $\mathcal{Y}$ given X, and $P_{X|Y}$ the class-conditional distribution of $\mathcal{X}$ given Y. In the multi-source domain generalization setup, we have access to M similar but distinct source domains, $\mathcal{S} = \{S_i\}_{i=1}^{M}$. In general, we assume that the joint distribution of each domain $P_{XY}^{(i)}$ is different from that of others, $P_{XY}^{(i)} \neq P_{XY}^{(i')}$ when $i \neq i'$. Each source domain consists of N_i samples, $S_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{N_i}$. The target domain, whose joint distribution is also different from source domain ones, is denoted by $\mathcal{T} = \{x_j^{\mathcal{T}}\}_{j=1}^{N_T}$. The labels for the target domain are unknown and need to be predicted. The goal is to find the learning function $f : \mathcal{X} \to \mathcal{Y}$ estimating $P_{Y|X}$, by minimizing a given loss function $\mathcal{L} : \mathcal{Y}\times\mathcal{Y} \to [0,\infty]$.

3.2. Structure-aware feature stylization

The proposed framework for Domain Generalization is illustrated in Fig. 1. Our Structure-Aware Feature Stylization model is added on top of a baseline CNN classifier, composed of an encoder followed by a classification head. It boosts the classifier's ability to generalize to new domains by mixing the feature statistics of source images and forcing the edge reconstruction of style-augmented samples to be similar to the true edges of original images. As baseline classifier, we train a neural network $f : \mathcal{X} \to [0,1]^K$ which consists of a feature extractor $g(\cdot)$ made of multiple convolutional layers, followed by a classifier $h(\cdot)$ with a single fully-connected layer and softmax output. Here, $K = |\mathcal{Y}|$ is the number of classes, which is same for all domains. We train f by minimizing the cross-entropy loss,

$$\mathcal{L}_{ce} = -\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\sum_{k=1}^{K} y_{j,k}^{(i)} \log f_k\left(x_j^{(i)}\right), \tag{1}$$

where $y_{j,k}^{(i)} = 1$ if the class label of $x_j^{(i)}$ is k, else 0. The next sections detail the feature stylization and structural consistency loss components of our model.

3.2.1. Feature stylization

While any other technique can be employed, our feature-based augmentation method is based on MixStyle (Zhou et al., 2021b). This approach is inspired by the adaptive instance normalization (AdaIN) method for style transfer (Huang and Belongie, 2017), which replaces feature statistics of a content image with statistics of a style image. For feature stylization, we choose two random instances $(x, \tilde{x})$ in a batch and compute feature statistics as

$$\begin{aligned}\gamma_{mix} &= \alpha\sigma(x) + (1-\alpha)\sigma(\tilde{x})\\ \beta_{mix} &= \alpha\mu(x) + (1-\alpha)\mu(\tilde{x})\end{aligned} \tag{2}$$

where α are instance-wise weights sampled from the Beta distribution, $\alpha \sim Beta(0.1, 0.1)$, and $\mu(x), \sigma(x)$ are the mean and standard deviation computed across the spatial dimension within each channel of each instance, as follows:

$$\begin{aligned}\mu_{b,c}(x) &= \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{b,c,h,w}\\ \sigma_{b,c}(x) &= \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(x_{b,c,h,w} - \mu_{b,c}(x)\right)^2}\end{aligned} \tag{3}$$

Finally, the mixed feature statistics are obtained as

$$\phi(x) = \gamma_{mix}\frac{x-\mu(x)}{\sigma(x)} + \beta_{mix}. \tag{4}$$

As it requires no explicit image synthesis mechanism and can be applied to any mini-batch training algorithm, this feature stylization method is simple to design and implement. Yet, as shown in our experimental results, it yields state-of-art performance when combined with the proposed structural consistency loss.

3.2.2. Structural consistency loss

The human vision system strongly relies on structural cues to locate and identify objects in a scene. From a young age, we can easily recognize a broad range of objects from very sparse structural information, for instance, a sketch with a few lines. Usually, these objects can still be recognized when colors or textures are modified in complex ways (e.g., changing the color of a giraffe from yellow to green).

Based on this idea, we define a loss to enforce structural consistency between source images and their style-augmented version. Let x be a training image and $\hat{x} = \phi(x)$ its stylized version, where $\phi(\cdot)$ is a feature-based augmentation function from a set $\mathcal{A}$. We extract the structural information in x using a Canny edge detector (Canny, 1986) which comprises five steps: (1) removing the noise with a Gaussian filter, (2) finding intensity gradients in the image, (3) using minimum cut-off suppression of gradient magnitudes to thin out edges, (4) applying a double threshold to remove spurious edge responses, (5) tracking edges by hysteresis to suppress weak edges that are not connected to strong ones. Compared to simple filter-based detectors, the Canny detector produces a sparser edge response that better corresponds to the true contours of objects in the image (Canny, 1986).

Denote as $z \in [0,1]^{W\times H}$ the edge map produced in an unsupervised manner by the detector for an image $x \in \mathbb{R}^{W\times H}$. To ensure that structural information is preserved for different feature-based augmentations ϕ, we add a decoder $d(\cdot)$ that reconstructs z from the output of the feature extractor, $g(x)$. Let $\hat{z}_\phi(x) = d\left(\phi(g(x))\right)$ be the predicted edge map for features stylized using transformation ϕ. Our structural consistency loss is defined as

$$\mathcal{L}_{sc} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sum_{j=1}^{N_i}\mathbb{E}_{\phi\sim\mathcal{A}}\left[\ell\left(z, \hat{z}_\phi(x_j^{(i)})\right)\right], \tag{5}$$

where $\ell(\cdot)$ is a combination of Dice loss (Sudre et al., 2017) and binary cross-entropy. We note that the edge reconstruction should also

Table 1
Ablation study for our method on the PACS dataset, reporting the mean and standard deviation across three runs.

Method	Accuracy (%)				
	Art	Cartoon	Photo	Sketch	Avg
Baseline	80.49 ±0.71	74.84 ±0.52	95.89 ±0.37	68.54 ±0.91	79.94
Only feature stylization	84.42 ±0.79	78.65 ±0.64	96.27 ±0.08	75.66 ±0.03	83.75
Only edge reconstruction	79.23 ±1.27	78.85 ±0.29	93.35 ±0.73	79.26 ±0.99	82.67
Both (Ours)	**85.53 ±0.14**	**79.89 ±0.97**	**96.75 ±0.45**	**82.28 ±1.04**	**86.11**

Table 2
Average accuracy of our method for the Sketch class of the PACS dataset, using different values of hyper-parameters λ and p. Reported values are the mean and standard deviation across 3 runs.

λ	p			
	0.1	0.3	0.5	0.8
0.01	81.52 ±1.16	80.57 ±0.69	81.49 ±0.47	81.61 ±0.75
0.05	80.90 ±1.37	80.21 ±1.26	81.16 ±0.26	**82.28 ±1.04**
0.1	79.94 ±1.27	82.16 ±1.06	81.72 ±0.70	80.54 ±0.59
0.4	80.92 ±0.89	80.55 ±1.09	80.90 ±1.22	80.88 ±0.52
1.0	79.66 ±1.40	80.31 ±0.87	79.83 ±0.48	80.07 ±0.80

be accurate for the original images. To account for this, we define a hyper-parameter $p \in [0, 1]$. Then, with probability p, the feature transformation function is drawn randomly from $\mathcal{A}$ and, with probability $1 - p$, the identity function is used for ϕ (no stylization). Using a value of $p = 0$ thus encourages the features of source domains to encode structural information without explicitly considering their generalizability to new domains, similar to the SSL approach in Albuquerque et al. (2020).

The structural consistency loss is optimized jointly with the classification loss, using the following total loss

$$\mathcal{L}_{tot} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{sc} \tag{6}$$

where hyper-parameter λ controls the trade-off between the two loss terms.

4. Experimental setup

4.1. Datasets and evaluation

We validate our method using five popular DG benchmarks for classification, PACS (Li et al., 2017), VLCS (Fang et al., 2013), OfficeHome (Venkateswara et al., 2017), DomainNet (Peng et al., 2019), and Digits-DG (Zhou et al., 2020b). PACS contains 9991 images of 7 classes belonging to four different domains, $d \in$ {Photo, Art, Cartton, Sketch}. VLCS (Fang et al., 2013) is comprised of four different domains, $d \in$ {Caltech101, LabelMe, SUN09, VOC2007}, five different classes, and 10,729 different photos. OfficeHome (Venkateswara et al., 2017) includes four domains, $d \in$ {Art, Clipart, Product, Real}, 65 classes, and a total of 15,588 photos. DomainNet has 6 domains, $d \in$ {Clipart, Infograph, Painting, Quickdraw, Real, Sketch}, 345 classes and 586,575 photos. Digits-DG consists of four different domains, namely MNIST (LeCun et al., 1998), MNIST-M (Ganin and Lempitsky, 2015), SVHN (Netzer et al., 2011) and SYN (Ganin and Lempitsky, 2015). In this dataset, images of different domains vary significantly in terms of font style, color and background, making it a highly challenging benchmark for out-of-distribution scenarios.

We follow a leave-out-one-domain strategy to evaluate performance for these datasets, where three domains are selected for training and the remaining one is used for testing. The final results correspond to the average accuracy calculated across the four different test domains.

4.2. Implementation details

For our experiments, depending on the dataset, we use different model architectures and data augmentations. Specifically, for PACS and Digits-DG, we employ a ResNet-18 model, while for the more challenging VLCS, OfficeHome, and DomainNet datasets, due to their pronounced domain shifts and size, we opt for ResNet-50 as our backbone. The feature stylization block is incorporated after the initial two residual blocks of the encoder. For decoding, we employ residual blocks similar to the encoder but use transposed convolution layers in lieu of down-sampling layers. At test time, the decoder is omitted, retaining only the encoder and classification head for inference.

During training, for all datasets, we apply random cropping and horizontal flipping as transformations. Additionally, for VLCS, OfficeHome, and DomainNet, we used ColorJitter, RandomGrayScaling, and ColorNormalizing transformations, following by the approaches in DomainBed (Gulrajani and Lopez-Paz, 2020). All our models are optimized using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 0.0005, allocating 20% of the training data for validation.

5. Results

We start by analyzing the proposed method by performing ablation studies and evaluating the impact of varying its hyper-parameters. We then provide visualization examples showing the ability of our method to preserve structural information across different feature-based augmentations. Finally, we compare our method against state-of-the-art approaches for DG and show its superior performance.

5.1. Ablation and parameter impact studies

We conduct an ablation study on the PACS dataset, using the ResNet-18 backbone, to evaluate the respective contribution to performance of the structural consistency loss and feature stylization components of our model. Four ablation variants are compared: the Baseline model where these two components are disabled, the model using only feature stylization ($\lambda = 0$), the model with the edge reconstruction task but no feature stylization ($p = 0$), and the proposed model combining both components.

As reported in Table 1, both the feature stylization and the edge reconstruction task yield significant improvements compared to the Baseline, when used by themselves. Using only feature stylization increases the average accuracy by 3.81% on PACS. Similarly, adding edge reconstruction without any feature-based augmentation raises average accuracy by 2.73%.

Not surprisingly, we observe that improvements brought by edge reconstruction is highest for domains with strong structural information, such as the Cartoon (improvement of 4.01%) and Sketch (improvement of 10.72%) domains. Additionally, improvements compared to using only feature stylization or edge reconstruction are statistically significant with $p < 0.05$ based on a paired t-test.

Next, we study the impact on performance of two important hyper-parameters, λ and p, respectively controlling the weight of the structural consistency loss of Eq. (5) and the ratio of samples on which feature stylization is applied. Table 2 shows our method's accuracy for the Sketch domain of PACS using different values of hyper-parameters λ and p. As can be seen, our method gives a good accuracy for a wide range of values, but generally works well with lower λ and higher p values.

Table 3
Comparison of our method with other state-of-the-art (SOTA) methods across three datasets: PACS (using Resnet-18 as the backbone), and VLCS and DomainNet (using Resnet-50 as the backbone). The table also provides the average performance across the three datasets. For other methods, results for PACS, are sourced from Kim et al. (2021); results for VLCS, OfficeHome and DomainNet, they are sourced from Cha et al. (2021).

Algorithm	PACS	VLCS	OfficeHome	DomainNet	Avg.
InfoDrop[a] (Shi et al., 2020)	82.2	–	–	–	–
EISNet[a] (Wang et al., 2020)	82.2	–	–	–	–
L2A-OT[a] (Zhou et al., 2020b)	82.8	–	–	–	–
DSON[a] (Seo et al., 2020)	85.1	–	–	–	–
pAdaIN[a] (Nuriel et al., 2021)	82.5	–	–	–	–
FSDCL[a] (Jeon et al., 2021)	85.9	–	–	–	–
DMG (Chattopadhyay et al., 2020)	81.5	–	–	43.6	–
MetaReg (Balaji et al., 2018)	81.7	–	–	43.6	–
mDSDI (Bui et al., 2021)	–	79.0	69.2	42.8	–
MMD (Li et al., 2018b)	84.6	77.5	66.4	23.4	63.0
Mixstyle (Zhou et al., 2021b)	83.7	77.9	60.4	34.0	64.0
IRM (Arjovsky et al., 2019)	83.5	78.6	64.3	33.9	65.1
GroupDRO (Sagawa et al., 2019)	84.4	76.7	66.0	33.3	65.1
ARM (Zhang et al., 2021)	85.1	77.6	64.8	35.5	65.8
VREx (Krueger et al., 2021)	84.9	78.3	66.4	33.6	65.8
CDANN (Li et al., 2018d)	82.6	77.5	65.7	38.3	66.0
DANN (Ganin et al., 2016)	83.6	78.6	65.9	38.3	66.6
RSC (Huang et al., 2020)	85.2	77.1	65.5	38.9	66.7
MTL (Blanchard et al., 2021)	84.6	77.2	66.4	40.6	67.2
Mixup (Yan et al., 2020)	84.6	77.4	68.1	39.2	67.3
MLDG (Li et al., 2018a)	84.9	77.2	66.8	41.2	67.5
ERM (Vapnik, 1998)	85.5	77.3	66.5	40.9	67.6
SagNet (Nam et al., 2021)	86.3	77.8	68.1	40.3	68.1
CORAL (Sun and Saenko, 2016)	86.2	78.8	68.7	41.5	68.8
SelfReg[b] (Kim et al., 2021)	**86.5**	77.8	67.9	43.1	68.8
SWAD (Cha et al., 2021)	82.9	79.1	70.6	46.5	69.8
DNA[b] (Chu et al., 2022)	83.1	79.0	**71.2**	**47.2**	70.1
Ours	86.1	**80.7**	70.2	46.1	**71.0**

[a] Methods marked are sourced from Jeon et al. (2021).
[b] Results for methods marked are sourced from their original paper.

Table 4
Comparison to the state-of-art on the Digits-DG dataset, reporting the mean accuracy and standard deviation across three runs.
Source: Results of other methods are taken from Zhou et al. (2021b).

	Accuracy (%)				
Method	MNIST	MNIST-M	SVHN	SYN	Avg
Baseline	95.8	58.8	61.7	78.6	73.7
JiGen (Carlucci et al., 2019a)	96.5	61.4	63.7	74.0	73.9
CCSA (Motiian et al., 2017)	95.2	58.2	65.5	79.1	74.5
MMD-AAE (Li et al., 2018b)	96.5	58.4	65.0	78.4	74.6
CrossGrad (Shankar et al., 2018)	96.7	61.1	65.3	80.2	75.8
L2A-OT (Zhou et al., 2020b)	96.7	63.9	68.6	83.2	78.1
MixStyle (Zhou et al., 2021b)	96.5	63.5	64.7	81.2	76.5
SWAD[a] (Cha et al., 2021)	97.30 ±0.17	61.36 ±0.76	63.81 ±0.94	87.24 ±0.31	77.43
DNA[a] (Chu et al., 2022)	97.46 ±0.05	62.41 ±0.63	62.77 ±1.16	87.75 ±0.92	77.60
Ours	**98.10 ±0.16**	**65.12 ±0.22**	**71.29 ±0.13**	**91.24 ±0.06**	**81.44**

[a] Methods were evaluated over three runs using the same backbone as in the original implementation of MixStyle.

5.2. Edge reconstruction analysis

We demonstrate that our consistency loss helps preserve structural information by comparing the true edge map of a non-stylized image to the reconstructed output of the same image after feature stylization. As shown in Fig. 2, the quality of reconstruction increases with higher values of hyper-parameter λ, since more importance is then given to the edge reconstruction task. Although small differences in the reconstructed edge map are observed across different augmentations, the outputs are globally consistent.

5.3. Comparison to the state-of-the-art

We present a detailed comparison of our method with leading domain generalization approaches in Table 3 for datasets PACS, VLCS, OfficeHome, DomainNet, and their average. Additionally, our results on the Digits-DG dataset are provided in Table 4.

For the Digits-DG dataset, we adopted the experimental setup from Zhou et al. (2021b) to ensure a fair comparison.

Our results indicate that while our approach may not always rank first for every dataset, it demonstrates stable and strong performance across varied domain shifts. For example, our method handles changes in style in datasets like PACS and DomainNet and also could adjusts to different environments and contexts in datasets such as VLCS and OfficeHome. As a result, our method achieves the best average accuracy when compared to other models.

In our comparison, we included various domain generalization techniques ranging from data augmentation methods like L2A-OT (Zhou et al., 2020b), pAdaIn (Nuriel et al., 2021), and Mixstyle (Zhou et al., 2021b) to regularization approaches like RSC (Huang et al., 2020).

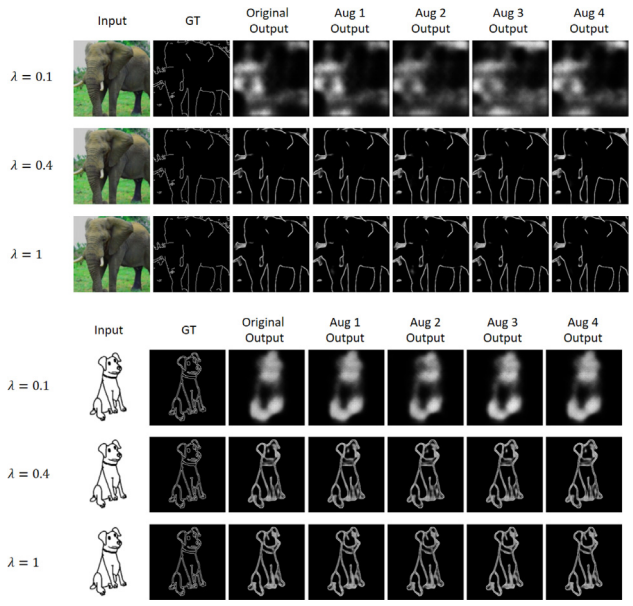

Fig. 2. Comparison of decoder output for original input and 4 random feature stylized input when model trained with different λ and $p = 0.5$ on Sketch as target domain.

These comparisons further highlight the consistent performance of our method across different challenges and benchmarks.

6. Conclusion

In this paper, we proposed a novel approach for domain generalization based on structure-aware feature stylization. Our approach enables the network to simulate domain shift in the latent representation by mixing instance-wise statistics of features in the encoder. It preserves structural information across different feature-based augmentations using a consistency loss that imposes reconstructed edge maps of stylized images to be similar to the true edges of the original, non-stylized ones. Experimental results showed that our multi-task setup regularizes the network by exploiting domain-invariant cues related to structure in the representation. Consequently, the classifier trained with our structure-aware stylization framework can better generalize to unseen domains. Our results also demonstrated the outstanding performance of our method compared to state-of-art approaches for DG, in particular for large domain shifts where preserving structural information is crucial.

CRediT authorship contribution statement

Milad Cheraghalikhani: Conceptualization, Methodology, Software, Writing – original draft. **Mehrdad Noori:** Conceptualization, Methodology, Software, Writing – original draft. **David Osowiechi:** Software, Writing – review & editing. **Gustavo A. Vargas Hakim:** Software, Visualization, Writing – review & editing. **Ismail Ben Ayed:** Supervision. **Christian Desrosiers:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

we have shared the link to our code repository in supplementary material document.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2024.104016.

References

Albuquerque, I., Naik, N., Li, J., Keskar, N., Socher, R., 2020. Improving out-of-distribution generalization via multi-task self-supervised pretraining. arXiv:2003.13525 [cs].

Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D., 2019. Invariant risk minimization. arXiv preprint arXiv:1907.02893.

Balaji, Y., Sankaranarayanan, S., Chellappa, R., 2018. Metareg: Towards domain generalization using meta-regularization. Adv. Neural Inf. Process. Syst. 31.

Blanchard, G., Deshmukh, A.A., Dogan, Ü., Lee, G., Scott, C., 2021. Domain generalization by marginal transfer learning. J. Mach. Learn. Res. 22 (1), 46–100.

Blanchard, G., Lee, G., Scott, C., 2011. Generalizing from several related classification tasks to a new unlabeled sample. In: Advances in Neural Information Processing Systems, vol. 24, Curran Associates, Inc..

Bucci, S., D'Innocente, A., Liao, Y., Carlucci, F.M., Caputo, B., Tommasi, T., 2021. Self-supervised learning across domains. arXiv:2007.12368 [cs].

Bui, M.-H., Tran, T., Tran, A., Phung, D., 2021. Exploiting domain-specific features to enhance domain generalization. Adv. Neural Inf. Process. Syst. 34, 21189–21201.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. (6), 679–698.

Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T., 2019a. Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Carlucci, F.M., Russo, P., Tommasi, T., Caputo, B., 2019b. Hallucinating agnostic images to generalize across domains. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop. ICCVW, pp. 3227–3234. http://dx.doi.org/10.1109/ICCVW.2019.00403.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., Park, S., 2021. SWAD: Domain generalization by seeking flat minima. Adv. Neural Inf. Process. Syst. 34, 22405–22418.

Chattopadhyay, P., Balaji, Y., Hoffman, J., 2020. Learning to balance specificity and invariance for in and out of domain generalization. In: European Conference on Computer Vision. Springer, pp. 301–318.

Chu, X., Jin, Y., Zhu, W., Wang, Y., Wang, X., Zhang, S., Mei, H., 2022. DNA: Domain generalization with diversified neural averaging. In: International Conference on Machine Learning. PMLR, pp. 4010–4034.

Fang, C., Xu, Y., Rockmore, D.N., 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1657–1664.

Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning. PMLR, pp. 1180–1189.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (19), 1–35.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 33, 21271–21284.

Gulrajani, I., Lopez-Paz, D., 2020. In search of lost domain generalization. arXiv preprint arXiv:2007.01434.

Hendrycks, D., Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. arXiv:1903.12261 [cs, stat].

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. arXiv:1703.06868 [cs].

Huang, Z., Wang, H., Xing, E.P., Huang, D., 2020. Self-challenging improves cross-domain generalization. In: European Conference on Computer Vision. Springer, pp. 124–140.

Ilse, M., Tomczak, J.M., Louizos, C., Welling, M., 2019. DIVA: Domain invariant variational autoencoders. arXiv:1905.10427 [cs, stat].

Jeon, S., Hong, K., Lee, P., Lee, J., Byun, H., 2021. Feature stylization and domain-aware contrastive learning for domain generalization. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 22–31.

Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J., 2021. Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9619–9628.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A., 2021. Out-of-distribution generalization via risk extrapolation (rex). In: International Conference on Machine Learning. PMLR, pp. 5815–5826.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Li, Y., Gong, M., Tian, X., Liu, T., Tao, D., 2018c. Domain generalization via conditional invariant representation. arXiv:1807.08479 [cs, stat].

Li, H., Pan, S.J., Wang, S., Kot, A.C., 2018b. Domain generalization with adversarial feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5400–5409.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D., 2018d. Deep domain generalization via conditional invariant adversarial networks. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 624–639.

Li, D., Yang, Y., Song, Y.-Z., Hospedales, T.M., 2017. Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5542–5550.

Li, D., Yang, Y., Song, Y.-Z., Hospedales, T., 2018a. Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32.

Mancini, M., Akata, Z., Ricci, E., Caputo, B., 2020. Towards recognizing unseen categories in unseen domains. arXiv:2007.12256 [cs].

Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G., 2017. Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5715–5725.

Muandet, K., Balduzzi, D., Schölkopf, B., 2013. Domain generalization via invariant feature representation. arXiv:1301.2115 [cs, stat].

Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D., 2021. Reducing domain gap by reducing style bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8690–8699.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., 2011. Reading digits in natural images with unsupervised feature learning.

Nuriel, O., Benaim, S., Wolf, L., 2021. Permuted AdaIN: reducing the bias towards global statistics in image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9482–9491.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B., 2019. Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1406–1415.

Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P., 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731.

Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., Han, B., 2020. Learning to optimize domain specific normalization for domain generalization. In: European Conference on Computer Vision. Springer, pp. 68–83.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S., 2018. Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745.

Shi, B., Zhang, D., Dai, Q., Zhu, Z., Mu, Y., Wang, J., 2020. Informative dropout for robust representation learning: A shape-bias perspective. In: International Conference on Machine Learning. PMLR, pp. 8828–8839.

Somavarapu, N., Ma, C.-Y., Kira, Z., 2020. Frustratingly simple domain generalization via image stylization. arXiv:2006.11207 [cs].

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248.

Sun, B., Saenko, K., 2016. Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, pp. 443–450.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience.

Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S., 2017. Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5018–5027.

Wang, S., Yu, L., Li, C., Fu, C.-W., Heng, P.-A., 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In: European Conference on Computer Vision. Springer, pp. 159–176.

Yan, S., Song, H., Li, N., Zou, L., Ren, L., 2020. Improve unsupervised domain adaptation with mixup training. arXiv preprint arXiv:2001.00677.

Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B., 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. pp. 2100–2110.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., Finn, C., 2021. Adaptive risk minimization: Learning to adapt to domain shift. Adv. Neural Inf. Process. Syst. 34, 23664–23678.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C., 2021a. Domain generalization in vision: A survey. arXiv preprint arXiv:2103.02503.

Zhou, K., Yang, Y., Hospedales, T., Xiang, T., 2020a. Deep domain-adversarial image generation for domain generalisation. Proc. AAAI Conf. Artif. Intell. (ISSN: 2374-3468) 34 (07), 13025–13032. http://dx.doi.org/10.1609/aaai.v34i07.7003.

Zhou, K., Yang, Y., Hospedales, T., Xiang, T., 2020b. Learning to generate novel domains for domain generalization. In: European Conference on Computer Vision. Springer, pp. 561–578.

Zhou, K., Yang, Y., Qiao, Y., Xiang, T., 2021b. Domain generalization with MixStyle. arXiv:2104.02008 [cs].