Domain Invariant Masked Autoencoders for Self-supervised Learning from Multi-domains

Haiyang Yang[1,4⋆], Shixiang Tang[3,4], Meilin Chen[2,4], Yizhou Wang[2,4], Feng Zhu[4], Lei Bai[3⋆⋆], Rui Zhao[4,5], and Wanli Ouyang[3]

[1]Nanjing University, China [2]Zhejiang University, China
[3]The University of Sydney, Austrilia [4]SenseTime Research
[5]Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China
hyyang@smail.nju.edu.cn, baisanshi@gmail.com, {yizhouwang, merlinis}@zju.edu.cn, stan3906@uni.sydney.edu.au, {zhufeng, zhaorui}@sensetime.com, wanli.ouyang@sydney.edu.au

Abstract. Generalizing learned representations across significantly different visual domains is a fundamental yet crucial ability of the human visual system. While recent self-supervised learning methods have achieved good performances with evaluation set on the same domain as the training set, they will have an undesirable performance decrease when tested on a different domain. Therefore, the self-supervised learning from multiple domains task is proposed to learn domain-invariant features that are not only suitable for evaluation on the same domain as the training set, but also can be generalized to unseen domains. In this paper, we propose a Domain-invariant Masked AutoEncoder (DiMAE) for self-supervised learning from multi-domains, which designs a new pretext task, *i.e.,* the cross-domain reconstruction task, to learn domain-invariant features. The core idea is to augment the input image with style noise from different domains and then reconstruct the image from the embedding of the augmented image, regularizing the encoder to learn domain-invariant features. To accomplish the idea, DiMAE contains two critical designs, 1) content-preserved style mix, which adds style information from other domains to input while persevering the content in a parameter-free manner, and 2) multiple domain-specific decoders, which recovers the corresponding domain style of input to the encoded domain-invariant features for reconstruction. Experiments on PACS and DomainNet illustrate that DiMAE achieves considerable gains compared with recent state-of-the-art methods.

1 Introduction

Recent advances on self-supervised learning (SSL) with the contrastive loss [19, 7, 9, 36] have shown to be effective in easing the burden of manual annotation, and achieved comparable performance with supervised learning methods. When

⋆ The work was done during an internship at SenseTime
⋆⋆ corresponding author

trained on large-scale datasets, *e.g.* ImageNet [12], self-supervised learning methods are capable of learning high-level semantic image representations [43, 14, 46] that are transferable to various downstream tasks without using expensive annotated labels. However, the great success of existing self-supervised learning methods implicitly relies on the assumption that training and testing sets are identically distributed, and thus these methods will suffer an undesirable performance drop when the trained model is tested on other domains [50, 35, 38] that do not exist in the training set.

Self-supervised learning from multi-domain data aims at learning domain invariant representations that are not only suitable for domains in the training set, but also can generalize well to other domains missing in the training set. Existing methods can be generally divided into two categories, *i.e,* self-prediction methods and contrastive-based methods. Early methods for self-supervised learning from multi-domain data append self-prediction tasks to learn domain-invariant features. For example, [15] randomly rotates the input image and regularizes the model to predict the rotation angle [16] to increase the model generalization ability. These self-prediction tasks are sub-optimal solutions, because they are not specifically designed to eliminate the domain bias in the dataset. Contrastive-based methods [23, 45] explicitly eliminate the domain bias by pulling the sample and its nearest neighbor from a different domain close. However, the positive pair retrieved by the nearest neighbor across the domains is much more noisy than that in a single domain, because semantically similar images from different domains may have a large visual difference.

In this paper, we tackle the self-supervised learning from multi-domain data from a different perspective, *i.e.,* generative self-supervised learning, and propose a new **D**omain **i**nvariant **M**asked **A**uto**E**ncoders (**DiMAE**) for learning domain-invariant features from multi-domain data, which is motivated by the recent generative-based self-supervised learning method Masked Auto-Encoders (MAE) [18]. Specifically, MAE eliminates the low-level information by masking large portion of image patches and drives the encoder to extract semantic information by reconstructing pixels from very few neighboring patches [3] with a light-weighted decoder. However, this design does not take the domain gaps into consideration and thus can not generalize well for the self-supervised learning from multi-domain tasks. To close the gap, our proposed DiMAE constructs a cross-domain reconstruction task, which uses *the image with the mixed style from different domains as input for one content encoder to extract domain invariant features and multiple domain-specific decoders to recover the specific domain style for regressing the raw pixel values of masked patches before style mix under an MSE loss*, as shown in Fig. 1. The critical designs and insights behind DiMAE for self-supervised learning from multi-domain data involve:

(1) The **cross-domain reconstruction task** aims at reconstructing the image from the image with other domain styles. DiMAE disentangles the reconstruction into two processes: a content[1] encoder to remove the domain style by

[1] "content" and "style" are terms widely used in style mix. "content" means domain-invariant information, while "style" means domain-specific information.

extracting domain-invariant features, and a domain-specific decoder to recover the style of the reconstruction target domain. By forcing the decoder to learn specific style information, we regularize the encoder to learn domain-invariant features.

(2) The **content preserved style mixing** aims to add style noise of the other domains to one image while preserving the content information. While there exist some popular mixing methods (*e.g,* mixup [44] and cutmix [42]) able to mix domain styles, they also add content noise to the image. Our experiments find that the content noise will lead to a significant performance decrease in our cross-domain reconstruction task. Therefore, we propose a new non-parametric content preserved style mixing method to take advantage of the cross-domain reconstruction and avoid the undesirable performance decrease by content noise.

(3) The **multiple domain-specific decoders** aim to recover the corresponding domain style of the target image for reconstruction from the encoded domain-irrelevant features. Although the decoder network design, *e.g.,* such as the number of layers, can determine the semantic level of the learned latent representations as pointed out in MAE [18], we find that a single decoder as used in MAE can not help to regularize the encoder to learn domain-invariant features. To reconstruct the image from a specific domain, the encoder will leak the domain information to guide the decoder to reconstruct the image with the input image's style. This prevents the encoder from learning the domain-invariant features. Therefore, multiple domain-specific decoders are proposed to recover different domain styles by domain-corresponding decoders, which regularizes the encoder to only learn domain-invariant features.

To demonstrate the effectiveness of DiMAE, we conduct experiments on the multi-domain dataset PACS [26] and DomainNet [33], observing consistent performance improvements on both in-domain and cross-domain settings. For the in-domain evaluation, DiMAE outperforms state-of-the-art methods by **+0.8%** on the PACS. On cross-domain testing, we achieve considerable gains over the recent state-of-the-art methods in both linear evaluation and full network fine-tuning. Specifically, in linear evaluation, our method improves the recent state-of-the-art by **+8.07%** on PACS with 1% data fine-tuning fraction. In full network fine-tuning with 100% data, we get an averaged **+13.24%** and **+9.87%** performance gains on PACS and DomainNet, respectively.

The contributions of our work are summarized as three-folds: **(1)** We propose a new generative framework which leverages the cross-domain reconstruction as the pretext to learn domain-invariant features from multi-domain data. **(2)** We propose a new non-parametric style-mix method that can preserve the content information to exploit the cross-domain reconstruction task and avoid performance drop by content noise. **(3)** We modify the single decoder in MAE to multiple domain-specific decoders to regularize the encoder to learn domain-invariant features. We show that our DiMAE outperforms state-of-the-art self-supervised learning baselines on learning representation from multi-domain data.

2 Related Work

2.1 Self-supervised Learning

Self-supervised Learning (SSL) introduces various pretext tasks to learn semantic representations from unlabeled data for a better generalization in downstream tasks. Generally, SSL can be categorized into discriminative [31, 16, 7, 17, 19, 9, 10, 43, 5, 48] and generative methods [32, 24, 25, 18]. Among the former, some early works try to design auxiliary handcrafted prediction tasks to learn semantic representation, such as jigsaw puzzle [31] and rotation prediction [16]. Recently, contrastive approaches [7, 17, 19, 9, 10, 43, 5] emerge as a promising direction for SSL. They consider each instance a different class and promote the instance discrimination by forcing representation of different views of the same image closer and spreading representation of views from different images apart. Although remarkable progress has been achieved, contrastive methods heavily rely on data augmentation [7, 36] and negative sampling [40, 19].

Another recent resurgent line of SSL is generative approaches, many of which train an encoder and decoder for pixel reconstruction. Various pretext tasks have been proposed, such as image inpainting [32] and colorization [24, 25]. Very recently, since the introduction of ViT [13], masked image modeling (MIM) has re-attracted the attention of the community. iGPT [6] proposes to predict the next pixels of a sequence, and BEiT [2] leverages a variational autoencoder (VAE) to encode masked patches. A very relevant work, MAE [18] proposes to train the autoencoder to capture the semantic representation by recovering the input image from very few neighboring patches. Our proposed method, a novel generative approach for SSL, is devoted into a more common scenario, pretraining from multiple domains. As far as we know, we are the first to propose the generative pretraining method for training from multi domain data.

2.2 Domain Generalization

Domain Generalization (DG) considers the transferability to unseen target domains using labeled data from a single or multiple source domains. A common approach is to minimize the distance between source domains for learning domain-invariant representations, among which are minimizing the KL Divergence [39], minimizing maximum mean discrepancy [28] and adversarial learning [30, 34, 1]. Several approaches propose to exploit meta-learning [27] or augmentation [4, 47, 11, 29] to promote the transferability for DG.

Despite the promising advances in recent DG methods, they assume that source domains are annotated. To address this issue, Unsupervised DG (UDG) is proposed as a more general task of training with unlabeled source domains. [15] introduces rotation prediction and mutual information maximization for multi-domain generalization. Derived from contrastive learning, DIUL [45] incorporates domain information into the contrastive loss by a reweighting mechanism considering domain labels. Despite the promising results, these two works carefully design domain-related discriminative pretext tasks and try to strike a compromise between instance and domain discrimination. Our proposed method, in

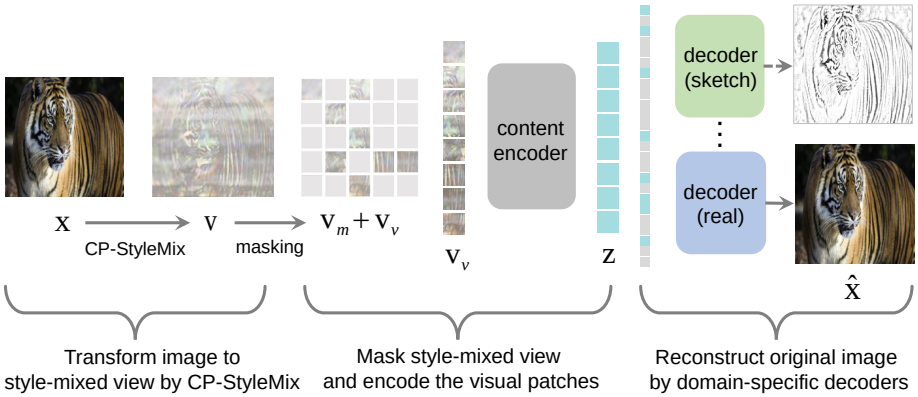


Fig. 1: The pipeline of DiMAE. First, CP-StyleMix transforms the original image $\mathbf{x}$ to its style-mixed view $\mathbf{v}$ by adding style information from other domains without introducing content noise. Second, the style-mixed view $\mathbf{v}$ is divided into visible patches $\mathbf{v}_v$ and masked patches $\mathbf{v}_m$, and the content encoder learns the content representation $\mathbf{z}$ from visible patches. Third, domain-specific decoders learn to reconstruct $\hat{\mathbf{x}}$ by the corresponding decoder.

contrast, is a brand new generative approach for self-supervised learning from multi-domain data, showing strong advantages for UDG setting.

3 Domain-Invariant Masked AutoEncoder

3.1 Cross-domain Reconstruction Framework

Different from MAE which learns high-level semantic representations by reconstruction from a highly masked image, our DiMAE learns domain invariant representation by a cross-domain reconstruction task, which aims at recovering images from an image mixed with other domain styles. Specifically, DiMAE consists three modules, including a Content Preserved Style-Mix (CP-StyleMix), a content encoder, and multiple domain-specific decoders. The CP-StyleMix is used to mix the style information from different domains while preserving the domain-irrelevant object content, which generates the input of the cross-domain reconstruction task. The content encoder $\mathcal{F}(*, \theta_{\mathcal{F}})$ are shared by images from all domains, where $\theta_{\mathcal{F}}$ is the parameter of $\mathcal{F}$, and is expected to encode the content and domain-invariant information by denoising the style information. The domain-specific decoders $\mathcal{G}$ in DiMAE are designed to incorporate the style information to the domain-invariant representation for image reconstruction, where $\mathcal{G} = \{\mathcal{G}_1(*, \phi_1), \mathcal{G}_2(*, \phi_2), ..., \mathcal{G}_{N_d}(*, \phi_{N_d})\}$, ϕ_i is the parameter for the i-th domain-specific decoder and N_d is the number of domains in the training set. As shown in Fig. 1, our DiMAE has the following steps:

Step1: *Transform an image* $\mathbf{x}$ *to its style-mixed view* $\mathbf{v}$ *by Content Preserved Style-Mix (Sec. 3.2).* Given an image $\mathbf{x}$, with Content Preserved Style-Mix, we

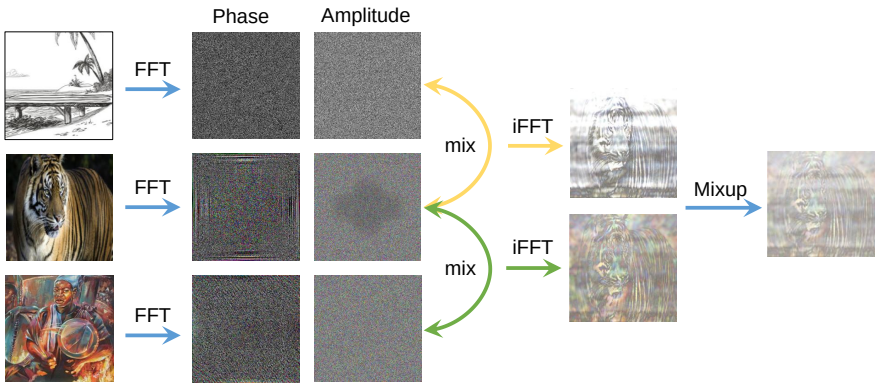


Fig. 2: The pipeline of CP-StyleMix. We mix the Fourier Amplitude of the original image $\mathbf{x}$ and two images from other domains to generate content preserved and style-transferred images, and mix them to generate the style-mixed view $\mathbf{v}$.

mix the style from other domains to the image $\mathbf{x}$ while preserving the content in $\mathbf{x}$ to generate its style-mixed view $\mathbf{v}$.

Step2: *Transform the style-mixed view* $\mathbf{v}$ *to content representation* $\mathbf{z}$ *(Sec. 3.3).* We randomly divides $\mathbf{v}$ into visible patches $\mathbf{v}_v$ and masked patches $\mathbf{v}_m$, and extract content representation $\mathbf{z}$ by encoding the visible patches $\mathbf{v}_v$ by $\mathcal{F}(*, \theta_{\mathcal{F}})$.

Step3: *Reconstruct the image* $\hat{\mathbf{x}}$ *by content representation* $\mathbf{z}$ *with the domain-specific decoders (Sec. 3.4).* Given content representation $\mathbf{z}$ and multiple domain-specific decoders $\mathcal{G} = \{\mathcal{G}_1(*, \phi_1), \mathcal{G}_2(*, \phi_2), ..., \mathcal{G}_{N_d}(*, \phi_{N_d})\}$, we reconstruct the image $\hat{\mathbf{x}}$ by $\mathcal{G}_i$, where $\mathcal{G}_i$ is the decoder of the i-th domain.

Step4: *Backward propagation using the MSE loss (Sec. 3.5).* Given the reconstructed image $\hat{\mathbf{x}}$ and the original image $\mathbf{x}$, the parameter $\theta_{\mathcal{F}}$ in $\mathcal{F}(*, \theta_{\mathcal{F}})$ and the parameters $\phi_1, \cdots, \phi_{N_d}$ in $\mathcal{G}(*, \phi_1)$, ..., $\mathcal{G}(*, \phi_{N_d})$ are learned by MSE loss.

3.2 Content Preserved Style-Mix

Content Preserved Style-Mix (CP-StyleMix) aims at mixing styles into an image while preserving the content information. This is a critical part for the cross-domain reconstruction tasks. Inspired by [41], the style information and the content information can be disentangled in the Fourier space. The content information is encoded in the phase of the Fourier signals, and the style information is encoded in the amplitude of the Fourier signals. We propose to first mix the style of the i-th domain to the image $\mathbf{x}$, generating its style views $\{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{N_d}\}$, where N_d is the number of domains. Then we mix these style views by the typical Mixup method [42], generating the final style-mixed view $\mathbf{v}$.

Specifically, for mixing in the Fourier space, given an image $\mathbf{x}$ from j-th domain and a randomly selected image $\mathbf{x}_{aux}$ from the i-th domain ($i \neq j$), the view $\mathbf{v}_i$ of image $\mathbf{x}$ can be formulated as

$$\mathbf{v}_i = \mathcal{K}^{-1}(\mathcal{K}^A_{mix}, \mathcal{K}^P(\mathbf{x})), \tag{1}$$

Table 1: Comparison between existing augmentation methods and CP-StyleMix.

Method	Venue	Content preserved	No extra training
CutMix [42]	ICCV'2019		✓
MixUp [44]	ICLR'2018		✓
StyleMix [21]	CVPR'2021	✓	
CycleGan+Mix [49]	ICCV'2017	✓	
CP-StyleMix(ours)	-	✓	✓

where $\mathcal{K}^A_{mix} = \lambda\mathcal{K}^A(\mathbf{x}_{aux}) + (1-\lambda)\mathcal{K}^A(\mathbf{x})$, $\mathcal{K}^{-1}$ is Fourier inversion, and $\mathcal{K}^A$, $\mathcal{K}^P$ returns the amplitude and phase of Fourier transformation, respectively. Then we implement the second step of mix on the image space by Mixup [44] process. Mathematically, the Mixup process can be formulated as

$$\mathbf{v} = \sum_{i=1}^{N_d} \mu_i \mathbf{v}_i, \tag{2}$$

where μ_i is the weight of different views, $\sum_{i=1}^{N_d} \mu_i = 1$, $\mu_j = 0$. Different from the Fourier style transfer proposed by [41], which do not have style mix, we mix different styles in both Fourier space and image space, leading to more diverse style information.

Discussion. Theoretically, as summarized in Tab. 1, there are various methods to mix the style information from other domains to the input image, including CutMix [42], MixUp [44], StyleMix [21], and CycleGan+Mix[49]. Our content preserved style mix is better than these methods in two critical aspects. First, our CP-StyleMix can preserve content information compared to CutMix and Mixup, which also mix contents. Detailed experiments and analysis in Sec. 4.3 illustrates that compared with content-pereserved methods, the mixture of content with Mixup and CutMix would significantly decrease the performance in reconstruction tasks by -10.47% and -9.71%, respectively. Second, our CP-StyleMix is non-parametric and does not need extra data. StyleMix [21] and CycleGan+Mix[49] can preserve the content information, but they require to train the transfer module by extra data, which will lead to unfair comparison with existing methods [15, 45].

3.3 Content Encoder

The content Encoder, *i.e.,* $\mathcal{F}(*, \theta_{\mathcal{F}})$, is designed to extract the domain-invariant content representations from the style-mixed view $\mathbf{v}$. Similar to MAE [18], our content encoder also follows the vision transformer design, which extracts content representations only by visible patches. Specifically, given a style-mixed view $\mathbf{v}$, we randomly divide the image patches into visible patches $\mathbf{v}_v$ with the probability p, leaving the remaining patches as the masked patches $\mathbf{v}_m$. The content representation $\mathbf{z}$ is then extracted by $\mathbf{v}_v$ using the content encoder, *i.e,*

$$\mathbf{z} = \mathcal{F}(\mathbf{v}_v, \theta_{\mathcal{F}}). \tag{3}$$

3.4 Domain Specific Decoders

Domain specific decoders are the critical designs in our proposed DiMAE. Besides the target of the decoder in MAE that is to reconstruct the semantic meaning of the masked patches, Domain specific decoders are expected to additionally reconstruct the domain style of the masked patches. To achieve this, we design a domain-specific decoder to each domain in the training set. Specifically, the domain specific decoders are defined as $\mathcal{G} = \{\mathcal{G}_1(*, \phi_1), \mathcal{G}_2(*, \phi_2), ..., \mathcal{G}_{N_d}(*, \phi_{N_d})\}$, where N_d is the number domains in the training set, $\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_{N_d}$ share the same architectural design, and ϕ_i is the parameter of the i-th domain-specific decoder $\mathcal{G}_i$. Given content representation $\mathbf{z}$, to reconstruct the patches in the i-th domain, we feed both the content representation $\mathbf{z}$ and the learnable masked tokens [18] into the i-th domain specific decoder $\mathcal{G}_i$, *i.e.,*

$$\hat{\mathbf{v}}_m^i = \mathcal{G}_i(\mathbf{z}, \mathbf{q}_m^i), \tag{4}$$

where $i \in [1, N_d]$ denotes the domain index, and the $\mathbf{q}_m^i$ denotes the masked tokens in the i-th domain-specific decoder.

Discussion. As pointed in MAE [18], the decoder design plays a key role in determining the semantic level of the learnt latent features. However, we argue that the domain-invariant features can not be learnt by changing the single decoder designs probably because of the style conflict in different domains. Instead, we propose to use multiple domain-specific decoders to learn the domain-invariant features. Specifically, we use a shared content encoder to learn the domain-invariant features, and expect the domain-specific decoder to recover the specific style information for the cross-domain reconstruction.

3.5 Objective Function

The objective function constrains the error between predicted patches and target patches, which drives the model to recover the original image $\mathbf{x}$ using very few mixed-styled neighboring patches. Specifically, given the image $\mathbf{x}$ from the j-th domain, the objective function can be formulated as

$$\mathcal{L} = (\hat{\mathbf{v}}^j - \mathbf{x})^2, \tag{5}$$

where $\hat{\mathbf{v}}^j$ is the reconstructed image by $\mathcal{G}_j$, $\mathbf{x}$ is the original image.

4 Experiment

4.1 Experimental Setup

Dataset. To validate our approach, we conduct extensive experiments with two generalization settings, namely in-domain and cross-domain, which detailed in Sec. 4.2. Two benchmark datasets are adopted to carry through these two settings. PACS [26] is a widely used benchmark for domain generalization. It consists of four domains, including Photo (1,670 images), Art Painting (2,048

Table 2: Results of In-domain top-1 linear evaluation accuracies on PACS dataset. Results style: **best**, second best.

Training Domain	(Photo, Art, Cartoon, Sketch)				
Method	Photo	Art	Cartoon	Sketch	Avg.
MoCo V3	70.6	39.4	64.8	54.4	57.3
MAE	83.5	53.4	74.2	**73.8**	71.2
DeepAll+MI,RotNet	81.6	55.5	68.5	63.4	67.3
DeepAll+MI,AET	80.9	56.9	69.6	67.9	68.8
DiMAE (ours)	**84.7**	**57.2**	**76.3**	69.8	**72.0**

images), Cartoon (2,344 images), Sketch (3,929 images) and each domain contains seven categories. DomainNet [33] is the largest, most diverse and recent cross-domain benchmark. Six domains are included: Real, Painting, Sketch, Clipart, Infograph and Quickdraw, with 345 object classes and 586, 575 examples.

For In-domain evaluations, we use all training subset in all domains for self-supervised learning, and then use the validation subset of each domain for evaluation. For cross-domain generalization, following DIUL [45], we select Painting, Real, Sketch as source domains and Clipart, Infograph, Quickdraw as target domains for DomainNet [33]. We select 20 classes out of 345 categories for both training and testing, exactly following the setting in [45]. For PACS, we follow the common setting in domain generalization [30, 34, 1] where three domains are selected as source domains, and the remaining domain is target domain.

Implementation details. In our implementation, we use ViT-small [2] as the backbone unless otherwise specified. The learning rate for pretraining is 0.5×10^{-4} and then decays with a cosine decay schedule. The weight decay is set to 0.05 and the batch size is set to $256 \times N_d$, where N_d is the number of domains in the training set. All methods are pretrained for 1000 epochs, which is consistent with the implementations in [45]. For finetuning, we follow the exact training schedule as that in [45]. Following [23], we use an ImageNet pretraining.

4.2 Experimental Results

In-Domain Evaluation. In-Domain Evaluation is proposed by [15], and aims to evaluate the performance of the self-supervised learning methods in the domains that appear in the training set. We exactly follow the protocol of [15]. Specifically, we learn the backbone on the training subset of all domains on PACS in a self-supervised manner, and then linearly train a classifier for each domain using the training subset of each domain with the backbone fixed, respectively. We evaluate our model on the validation subset in each domain, and report the averaged results by 10 runs. The experimental results are summarized

[2] We do not use the widely-used ResNet18 [20] as the backbone, because DiMAE is exactly a generative method, in which Convolutaional networks are not applicable. We choose the ViT-small model for comparison because the number of their model parameters is similar.

Table 3: Results of the cross-domain generalization on DomainNet. All of the models are trained on Painting, Real, Sketch domains of DomainNet and tested on the other three domains. The title of each column indicates the name of the domain used as target. All the models are pretrained for 1000 epoches before finetuned on the labeled data. Results style: **best**, second best.

	Label Fraction 1%					Label Fraction 5%				
method	Clipart	Infograph	Quickdraw	Overall	Avg.	Clipart	Infograph	Quickdraw	Overall	Avg.
ERM	6.54	2.96	5.00	4.75	4.83	10.21	7.08	5.34	6.81	7.54
MoCo V2 [9]	18.85	10.57	6.32	10.05	11.92	28.13	13.79	9.67	14.56	17.20
SimCLR V2 [8]	23.51	15.42	5.29	11.80	14.74	34.03	17.17	10.88	17.32	20.69
BYOL [17]	6.21	3.48	4.27	4.45	4.65	9.60	5.09	6.02	6.49	6.90
AdCo [22]	16.16	12.26	5.65	9.57	11.36	30.77	18.65	7.75	15.44	19.06
MAE	22.38	12.62	10.50	13.51	15.17	32.60	15.28	13.43	17.85	20.44
DIUL	18.53	10.62	12.65	13.29	13.93	39.32	**19.09**	10.50	18.73	22.97
DiMAE (ours)	**26.52**	**15.47**	**15.47**	**17.72**	**19.15**	**42.31**	18.87	**15.00**	**21.68**	**25.39**
	Label Fraction 10%					Label Fraction 100%				
method	Clipart	Infograph	Quickdraw	Overall	Avg.	Clipart	Infograph	Quickdraw	Overall	Avg.
ERM	15.10	9.39	7.11	9.36	10.53	52.79	23.72	19.05	27.19	31.85
MoCo V2	32.46	18.54	8.05	15.92	19.69	64.18	27.44	25.26	33.76	38.96
SimCLR V2	37.11	19.87	12.33	19.45	23.10	68.72	27.60	30.56	37.47	42.29
BYOL	14.55	8.71	5.95	8.46	9.74	54.44	23.70	20.42	28.23	32.86
AdCo	32.25	17.96	11.56	17.53	20.59	62.84	26.69	26.26	33.80	38.60
MAE	51.86	24.81	23.94	29.87	33.54	59.21	28.53	23.27	32.06	37.00
DIUL	35.15	20.88	15.69	21.08	23.91	72.79	32.01	33.75	41.19	46.18
DiMAE (ours)	**70.78**	**38.06**	**27.39**	**39.20**	**45.41**	**83.87**	**44.99**	**39.30**	**49.96**	**56.05**

in Tab. 2. DiMAE outperforms MoCo V3 and MAE by **+14.7%** and **+0.8%**, respectively, showing the superior of in-domain instance discrimination ability against the previous methods. Furthermore, when we compare the baseline generative method, *i.e.,* MAE, with contrastive learning methods, *i.e.,* MoCoV3, we infer that the reconstruction task can learn better representations of the domains that appear in the training set.

Cross-Domain Generalization. Cross-Domain Generalization is firstly proposed by DIUL [45], which evaluates the generalization ability of the self-supervised learning methods to the domains that are missing in the training set. We exactly follow the cross-domain generalization evaluation process in DIUL [45], which is divided into three steps. First, we train our model on source domains in the unsupervised manner. Then, we will use a small number of labeled training examples of the validation subset in the source domains to finetune the classifier or the whole backbone. In detail, when the fraction of labeled finetuning data is lower than 10% of the whole validation subset in the source domains, we only finetune the linear classifier for all the methods. When the fraction of labeled finetuning data is larger than 10% of the whole validation subset in the source domains, we finetune the whole network, including the backbone and the classifier. Last, we can evaluate the model on the target domains.

The results are presented in Tab. 3 (DomainNet) and Tab. 4 (PACS). In this setting, our DiMAE achieves a better performance than previous works on most tasks and gets significant gains over DIUL and other SSL methods on overall and

Table 4: Results of the cross-domain generalization setting on PACS. Given the experiment for each target domain is run respectively, there is no overall accuracy across domains. Thus we report the average accuracy and the accuracy for each domain. The title of each column indicates the name of the domain used as target. All the models are pretrained for 1000 epochs before finetuned on the labeled data. Results style: **best**, second best.

	Label Fraction 1%					Label Fraction 5%				
method	Photo	Art.	Cartoon	Sketch	Avg.	Photo	Art.	Cartoon	Sketch	Avg.
MoCo V2	22.97	15.58	23.65	25.27	21.87	37.39	25.57	28.11	31.16	30.56
SimCLR V2	30.94	17.43	**30.16**	25.20	25.93	**54.67**	35.92	35.31	36.84	40.68
BYOL	11.20	14.53	16.21	10.01	12.99	26.55	17.79	21.87	19.65	21.47
AdCo	26.13	17.11	22.96	23.37	22.39	37.65	28.21	28.52	30.35	31.18
MAE	30.72	23.54	20.78	24.52	24.89	32.69	24.61	27.35	30.44	28.77
DIUL	27.78	19.82	27.51	29.54	26.16	44.61	39.25	**36.41**	36.53	39.20
DiMAE (ours)	**48.86**	**31.73**	25.83	**32.50**	**34.23**	50.00	**41.25**	34.40	**38.00**	**40.91**
	Label Fraction 10%					Label Fraction 100%				
method	Photo	Art.	Cartoon	Sketch	Avg.	Photo	Art.	Cartoon	Sketch	Avg.
MoCo V2	44.19	25.85	33.53	24.97	32.14	59.86	28.58	48.89	34.79	43.03
SimCLR V2	54.65	37.65	46.00	28.25	41.64	67.45	43.60	54.48	34.73	50.06
BYOL	27.01	25.94	20.98	19.69	23.40	41.42	23.73	30.02	18.78	28.49
AdCo	46.51	30.21	31.45	22.96	32.78	58.59	29.81	50.19	30.45	42.26
MAE	35.89	25.59	33.28	32.39	31.79	36.84	25.24	32.25	34.45	32.20
DIUL	53.37	39.91	46.41	30.17	42.47	68.66	41.53	56.89	37.51	51.15
DiMAE (ours)	**77.87**	**59.77**	**57.72**	**39.25**	**58.65**	**78.99**	**63.23**	**59.44**	**55.89**	**64.39**

average accuracy[3]. Compared with contrastive learning based methods, such as MoCo V2, SimCLR V2, BYOL, AdCo, our generative based methods improves the cross-domain generalization tasks by **+3.98%** and **+2.42%** for DomainNet and **+8.07%** and **+0.23%** for PACS on 1% and 5% fraction setting respectively, which is tested by linear evaluation. Our DiMAE also improves other states-of-the-art methods by **+11.87%** and **+9.87%** for DomainNet, **+16.18%** and **+13.24%** for PACS on 10% and 100% fraction setting, respectively, when the whole backbone are finetuned. The significant improvement to contrastive learning based methods illustrate our proposed DiMAE can learn more domain-invariant features in the self-supervised learning from multiple domain data.

4.3 Ablation Study

To investigate the effectiveness of each component of our proposed DiMAE, We ablate our DiMAE on the Cross-Domain Generalization task. Specifically, we train Vit-Tiny [37] for 100 epoches on the combination of Painting, Real, and Sketch training set in DomainNet, and evaluate the model using the linear evaluation protocol on Clipart.

[3] Overall and Avg. indicate the overall accuracy of all the test data and the arithmetic mean of the accuracy of 3 domains, respectively. Note that they are different because the capacities of different domains are not equal.

Table 5: Comparison of using content-preserved methods, content-mix methods, and no augmentation. Aug. is short for augmentation.

Content-preserved		Content-mix		No aug.
CP-StyleMix	CP-StyleCut	Mixup	CutMix	
48.56	47.21	38.09	37.50	36.85

Table 6: Comparison of style transfer [41], CP-StyleMix and CP-StyleCut. Aug. is short for augmentation.

Content-preserved Augmentation	Top-1
Style transfer [41]	46.11
CP-StyleMix	**48.56**
CP-StyleCut	47.21

Effectiveness of Preserving Contents in Style Mix. To demonstrate the importance of preserving contents in style mix, we ablate the content-preserved and content-mix augmentation methods for DiMAE, which is presented in Tab. 5. Specifically, we choose CP-StyleMix for content-preserved methods and Mixup and CutMix for content-mixed methods. Additionally, to fairly compare with CutMix, we replace the Mixup step in CP-StyleMix with CutMix, creating a competing method called Content Preserved StyleCut (CP-StyleCut). We conclude that preserving the content information is critical for reconstruction tasks. Specifically, we observe that content-mix methods, *i.e.,* Mixup and CutMix, bring at most +**1.24**% performance improvement compared with no augmentation. However, content preserved style mix methods, *i.e,* CP-StyleMixp and CP-StyleCut, can further improve the content-mix style-mix augmentations, *i.e.,* Mixup and CutMix, by +**10.47**% and +**9.71**%. The large performance gap between content-preserved and content-mix augmentations methods indicates the importance of preserving contents in the reconstruction tasks.

Effectiveness of Mixing Style Information. To illustrate the importance of mixing style information in our proposed DiMAE (Eq. 2), we ablate the mixing step by comparing the experiments where we use the mixed-style view $\mathbf{v}$ in Eq. 2, and the view $\mathbf{v}_i$ before mixing. Here, $\mathbf{v}_i$ is the i-th style view after style transfer (Eq. 1) before Mixup (Eq. 2). As shown in Tab. 6, after applying Mixup and CutMix on the views after style transfer, the performance further increases by +**2.45**% and +**1.10**%. The consistent improvement indicates that adding more style noise can effectively help the encoder to learn domain-invariant features.

Effectiveness of Multiple Domain-specific Decoders. A novel design of our proposed DiMAE is the domain-specific decoders, which reconstruct corresponding domain-specific images using the encoded latent representation. We ablate this design with all other factors fixed. Experimental results are illustrated in Tab. 7, showing the linear evaluation performance when the single decoder and Domain Specific Decoders are applied. We observe that the methods using domain-specific decoders improve the methods using the single decoder by +**10.47**% and +**9.71**% when images are augmented by CP-StyleMix and

Table 7: Comparison of single decoder and Domain Specific Decoders. Domain Specific Decoders achieve significant performance improvement with CP-StyelMix and CP-StyleCut.

Augmentations	Single Decoder	Domain Specific Decoders
CP-StyleMix	38.09	**48.56**
CP-StyleCut	37.50	**47.21**

Table 8: Comparison of different depth of Domain Specific Decoders.

Depth	Single Decoder	Multi Decoders
1	37.46	44.93
2	37.81	45.35
4	38.01	46.62
8	**38.09**	**48.56**
12	37.96	46.11

CP-StyleCut. The significant performance gap verifies the importance of using domain-specific decodoers in our proposed DiMAE. To explain the performance gap, we argue that this is because domain-specific decoders help to decouple the different style information from different domains to the corresponding decoders, regularizing the encoder to only learn domain-invariant features.

Designs in the single decoder and multiple domain-specific decoders. Tab. 8 varies the decoder depth, from which we have two findings. First, we find the depth of the decoder is also important in our task, because a sufficiently deep decoder can improves the performance by 0.63% and 3.63% in single and multiple decoders design, respectively. Second, the performance gain in multi-decoders design (**+3.63%**) is much larger than in single-decoder design (**+0.63%**), because the depth of decoders can influence the semantic level of the learned feature, but can not help to regularize the encoder to learn domain-invariant features, which is crucial in our self-supervised learning from multi-domain data task.

4.4 Visualization

Feature Distribution Visualization. Qualitatively, Fig. 3 visualizes the feature distribution of MoCo V3, MAE and DiMAE by t-SNE, on the combination of Painting, Real, and Sketch training set in DomainNet. We observe that the features of DiMAE between three domains are significantly better mixed than the others. This suggests that compared with MoCo V3 and MAE, DiMAE is able to capture better domain-invariant representations.

Reconstruction Visualization. We visualize reconstruction results of DiMAE using ViT-base in Fig. 4. The results demonstrate that, in our DiMAE, the encoder removes the domain style and multiple decoders learn specific style information. Specifically, DiMAE eliminates the style noise on visible patches as no messy style information appears in reconstructions. Second, DiMAE provides

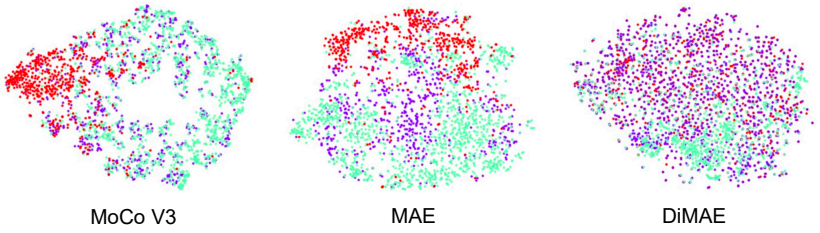


Fig. 3: Visualization of the feature distribution of MoCo V3, MAE and DiMAE.

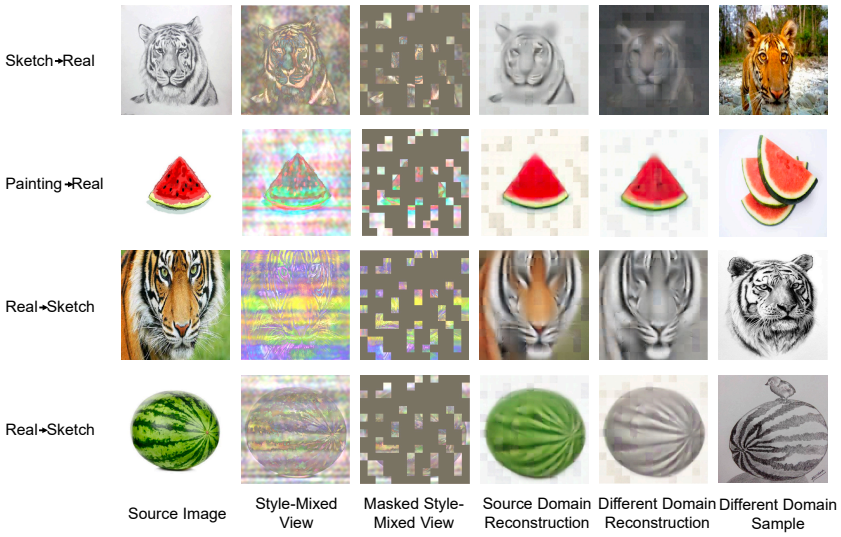


Fig. 4: Reconstruction visualization of different decoders. Sketch→Real denotes using Sketch as source domain and Real as the a different domain to reconstruct.

complete reconstructions with specific domain styles. Third, we also observe that it is quite hard for DiMAE to recover colors perfectly from sketch inputs.

5 Conclusions

In this paper, we propose a novel Domain invariant Masked AutoEncoder (DiMAE) to tackle the self-supervised learning from multi-domain data. Our DiMAE constructs a new cross-domain reconstruction task with a proposed content preserved style mix and multiple decoder designs to learn domain-invariant features. The content preserved style mix aims to mix style information from different domains, while preserving the image content. The multiple decoders are proposed to regularize the encoder to extract domain-invariant features. Extensive experiments validate the effectiveness of DiMAE.

References

1. Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T.H., Mitliagkas, I.: Generalizing to unseen domains via distribution matching. arXiv preprint arXiv:1911.00804 (2019)
2. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Cao, S., Xu, P., Clifton, D.A.: How to understand masked autoencoders. arXiv preprint arXiv:2202.03670 (2022)
4. 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)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
6. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020)
9. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
10. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9640–9649 (2021)
11. Dai, Y., Li, X., Liu, J., Tong, Z., Duan, L.Y.: Generalizable person re-identification with relevance-aware mixture of experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16145–16154 (June 2021)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5414–5423 (2021)
15. Feng, Z., Xu, C., Tao, D.: Self-supervised representation learning from multi-domain data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3245–3255 (2019)
16. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
17. 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.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)

18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
20. 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)
21. Hong, M., Choi, J., Kim, G.: Stylemix: Separating content and style for enhanced data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14862–14870 (2021)
22. Hu, Q., Wang, X., Hu, W., Qi, G.J.: Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1074–1083 (2021)
23. Kim, D., Saito, K., Oh, T.H., Plummer, B.A., Sclaroff, S., Saenko, K.: Cds: Cross-domain self-supervised pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9123–9132 (2021)
24. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: European conference on computer vision. pp. 577–593. Springer (2016)
25. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6874–6883 (2017)
26. 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)
27. 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)
28. 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)
29. Li, X., Dai, Y., Ge, Y., Liu, J., Shan, Y., DUAN, L.: Uncertainty modeling for out-of-distribution generalization. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=6HN7LHyzGgC
30. 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)
31. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
32. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
33. 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)
34. Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S.: Correlation-aware adversarial domain adaptation and generalization. Pattern Recognition **100**, 107124 (2020)

35. Sariyildiz, M.B., Kalantidis, Y., Larlus, D., Alahari, K.: Concept generalization in visual representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9629–9639 (2021)
36. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? arXiv preprint arXiv:2005.10243 (2020)
37. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
38. Wang, Y., Tang, S., Zhu, F., Bai, L., Zhao, R., Qi, D., Ouyang, W.: Revisiting the transferability of supervised pretraining: an mlp perspective. arXiv preprint arXiv:2112.00496 (2021)
39. Wang, Z., Loog, M., van Gemert, J.: Respecting domain relations: Hypothesis invariance for domain generalization. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9756–9763. IEEE (2021)
40. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
41. Xu, Q., Zhang, R., Zhang, Y., Wang, Y., Tian, Q.: A fourier-based framework for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14383–14392 (2021)
42. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
43. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
44. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
45. Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., Liu, H.: Domain-irrelevant representation learning for unsupervised domain generalization. arXiv preprint arXiv:2107.06219 (2021)
46. Zhao, N., Wu, Z., Lau, R.W., Lin, S.: What makes instance discrimination good for transfer learning? arXiv preprint arXiv:2006.06606 (2020)
47. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13025–13032 (2020)
48. Zhu, J., Tang, S., Chen, D., Yu, S., Liu, Y., Rong, M., Yang, A., Wang, X.: Complementary relation contrastive distillation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
49. 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)
50. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)