Attention Diversification for Domain Generalization

Rang Meng[1,⋆], Xianfeng Li[1,⋆], Weijie Chen[2,1,✉], Shicai Yang[1,✉], Jie Song[2], Xinchao Wang[3], Lei Zhang[4], Mingli Song[2], Di Xie[1], and Shiliang Pu[1]

[1] Hikvision Research Institute, Hangzhou, China
[2] Zhejiang University, Hangzhou, China
[3] National University of Singapore, Singapore
[4] Chongqing University, Chongqing, China
{mengrang, lixianfeng6, chenweijie5, yangshicai, xiedi, pushiliang.hri}@hikvision.com, {sjie, songml}@zju.edu.cn, xinchao@nus.edu.sg, leizhang@cqu.edu.cn

Abstract. Convolutional neural networks (CNNs) have demonstrated gratifying results at learning discriminative features. However, when applied to unseen domains, state-of-the-art models are usually prone to errors due to domain shift. After investigating this issue from the perspective of shortcut learning, we find the devils lie in the fact that models trained on different domains merely bias to different domain-specific features yet overlook diverse task-related features. Under this guidance, a novel *Attention Diversification* framework is proposed, in which Intra-Model and Inter-Model Attention Diversification Regularization are collaborated to reassign appropriate attention to diverse task-related features. Briefly, Intra-Model Attention Diversification Regularization is equipped on the high-level feature maps to achieve in-channel discrimination and cross-channel diversification via forcing different channels to pay their most salient attention to different spatial locations. Besides, Inter-Model Attention Diversification Regularization is proposed to further provide task-related attention diversification and domain-related attention suppression, which is a paradigm of "*simulate, divide and assemble*": simulate domain shift via exploiting multiple domain-specific models, divide attention maps into task-related and domain-related groups, and assemble them within each group respectively to execute regularization. Extensive experiments and analyses are conducted on various benchmarks to demonstrate that our method achieves state-of-the-art performance over other competing methods. Code is available at `https://github.com/hikvision-research/DomainGeneralization`.

Keywords: Domain Generalization, Attention Diversification

1 Introduction

Domain is clarified as the feature space and marginal probability distribution for a specific dataset [2, 3]. And domain shift reveals the discrepancy between

⋆ Equal contribution. ✉ Corresponding authors.

source and target domains [2, 3, 57], which induces the models trained on source domains to perform defectively on an unseen target domain. Domain adaptation (DA) aims to remedy this issue of domain shift for various tasks in cases that target data is available [7, 8, 29, 32, 36, 39, 49, 62, 72, 73]. However, the domain shift is usually agnostic in real-world scenarios since the target data is not available for training. This issue inspires the research area of domain generalization (DG) [1, 22, 27, 28, 30, 34, 41, 43, 45, 47, 51, 52, 54, 74, 75, 78–80], which is aimed to make models trained on seen domains achieve accurate predictions on unseen domainss, i.e., the conditional distribution $P(Y|X)$ is robust with shifted marginal distribution $P(X)$.

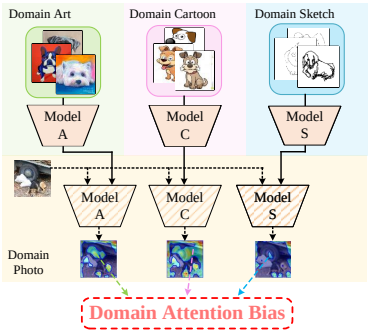

Fig. 1. The visualization of domain attention bias on PACS dataset. Domain-specific models trained on different domains (ACS) pay attention to different regions when they are tested on an unseen domain (P).

Canonical DG focuses on learning a domain-invariant feature distribution $P(F(X))$ across domains for the robustness of conditional distribution $P(Y|F(X))$. In fact, the domain issue can be revisited from the perspective of shortcut learning [15], which indicates that models attempt to find the simplest solution to solve a given task. Models trained on specific domains merely pay attention to salient domain-related features while overlooking other diverse task-related information. When the domain shifts, the discrimination of the biased features will not be held on the unseen domain, leading to the shift of the conditional distribution. This problematic phenomenon is dubbed as "domain attention bias" as shown in Fig. 1.

In this paper, we propose the *Attention Diversi cation* framework, in which the attention mechanism is served as the bridge to achieve the invariance of conditional distribution. In our framework, the proposed Intra-Model Attention Diversification Regularization (Intra-ADR) and Inter-Model Attention Diversification Regularization (Inter-ADR) are collaborated to rearrange appropriate spatial attention to diverse task-related features from coarse to fine. The reasons why the two components are designed in our framework are detailed as follows:

Intra-Model Attention Diversification Regularization. According to *the principle of maximum entropy* [18], when estimating the probability distribution, we should select that distribution which leaves us the largest uncertainty under our constraints, so that we cannot bring any additional assumptions into our computation. That is, when testing the unseen domains, each task-related feature is equally-useful (i.e., the maximum entropy), driving us to propose Intra-ADR, which coarsely recalls overlooked features outside the domain attention bias as much as possible. This is done via forcing different channels to pay attention to different spatial locations, leading all spatial locations to be activated. To this end, in-channel discrimination and cross-channel diversification are facilitated.

Although the Intra-ADR is equipped upon the high-level features, not all spatial regions are consistent with the semantics of the categories. As stated in [15], the background regions mainly involve domain-related features, and some parts of foreground regions are also affected by domain-specific styles [21, 23]. Since the Intra-ADR fails to distinguish features at the finer level into task-related and domain-related ones, the excessive attention is incidentally imposed upon domain-related features, leading to the conditional distribution shift. Thus, an attention diversification paradigm at a finer level is necessary.

Inter-Model Attention Diversification Regularization. To handle the aforementioned issue, features that Intra-ADR coarsely recalls ought to be further refined by Inter-ADR. Thus, the diverse attention for task-related features is encouraged, yet the excessive attention for domain-related ones is suppressed. Inter-ADR is a paradigm of "*simulate, divide and assemble*". Specifically, 1) "*simulate*": we train multiple domain-specific models for each seen domain, and then infer these models on samples from other training domains to simulate domain shift. In addition, the attention maps and predictions for agnostic domains are generated; 2) "*divide*": we divide attention maps from domain-specific models and domain-aggregated model into the task-related and domain-related groups, according to whether the model predictions is consistent with the corresponding ground truth; 3) "*assemble*": attention maps from different models are assembled within each group as the task-related and domain-related inter-model attention maps, respectively. Finally, the attention maps of the domain-aggregated model can be regularized with the task-related and domain-related inter-model attention maps, to diversify task-related attention regions yet suppress domain-related attention regions.

Extensive experiments and analyses are conducted on multiple domain generalization datasets. Our optimization method achieves state-of-the-art results. It is worth emphasizing that our method can bring further performance improvement in conjunction with other DG methods.

2 Related Works

Domain Generalization. The analysis in [2] proves that the features tend to be general and can be transferred to unseen domains if they are invariant across different domains. Following this research, a sequence of domain alignment methods is proposed, which reduce the feature discrepancy among multiple source domains via aligning domain-invariant features. These methods enable models to generalize well to unseen target domains. Specifically, they use explicit feature alignment by minimizing the maximum mean discrepancy (MMD) [58] or using Instance Normalization (IN) layers [43]. Alternatively, [22, 47] adopt domain adversarial learning for domain alignment, which trains a discriminator to distinguish the domains while training feature extractors to cheat the domain discriminator for learning domain-invariant features. Besides, the ability of generalizing to unseen domains will increase as training data covering more

diverse domains. Several domain diversification attempts had been implemented in previous works: swapping the shape or style information of two images [25], mixing instance-level features of training samples across domains [78], altering the location and scene of objects [46], and simulating the actual environment for generating more training data [56]. In contrast, we investigate the issue of DG inspired by shortcut learning and maximum entropy principle. Besides, we introduce visual attention in our proposed method to boost DG, which is seldom studied in prior works.

Visual Attention. Visual attention has been widely used in deep learning and achieves remarkable advances [59, 69]. It has been exploited in computer vision tasks such as image recognition [9, 10, 33, 48, 61, 71] and object detection among others [6, 16, 35, 66, 67]. CAM [77] provides the attention visualization of feature maps for model interpretable analysis. In essence, visual attention can be interpreted as an allocation mechanism for the model learning resource: it assigns high weights to what the model considers valuable, and vice versa, assigns low weight to what the model considered negligible [70]. Motivated by this mechanism, many computer vision tasks achieve breakthrough. For example, many fine-grained image classification methods learn multi-attention to capture sufficient subtle inter-category differences [14, 53, 68, 76]. Recently, self-attention [13, 20, 64] has emerged to model the long-range dependencies. In the field of transfer learning, Attentional Heterogeneous Transfer (AHT) [40] designed a new heterogeneous transfer learning approach to transfer knowledge from an optimized subset of source domain samples to a target domain. Transferable Attention for Domain Adaptation (TADA) [63] is proposed to use transferable global and local attention with multi-region-level domain discriminators to pick out the images and the transferable areas of the image.

Our work finds that CNN allocates sufficient attention to domain-related features, but insufficient attention to task-related features conversely. Under this consideration, we adopt spatial attention as a bridge to learn diverse transferable features to mitigate domain shifts.

3 Method

Our proposed *Attention Diversi cation* framework is composed of Intra-ADR and Inter-ADR as shown in Fig. 2. Our framework aims to deny shortcut learning, which ignores numerous task-related features. The Intra-ADR and Inter-ADR are collaborated to diversify attention regions for task-related features.

Notations. Given $\mathcal{S}$ training domains $\{\mathcal{D}_d\}_{d=1}^{\mathcal{S}}$, where $\mathcal{D}_d = \{(x_i^d, y_i^d)\}_{i=1}^{N_d}$ with N_d labeled samples covering Z categories. Let $\mathcal{M}$ denote the CNN model used for image classification. Suppose $\mathcal{X}_j^b \in R^{\mathcal{C}^b \times \mathcal{H}^b \times \mathcal{W}^b}$ denote the feature maps output from the b-th block of the model $\mathcal{M}_j$, where $\mathcal{C}^b$, $\mathcal{H}^b$ and $\mathcal{W}^b$ denote the channel number, height and width of $\mathcal{X}_j^b$, and $b \in \{1, ..., B\}$. We denote the domain-specific models and domain-aggregated model as $\{\mathcal{M}_j\}_{j=0}^{\mathcal{S}}$, where $\mathcal{M}_1, ..., \mathcal{M}_{\mathcal{S}}$ represent the former which is trained on the corresponding single

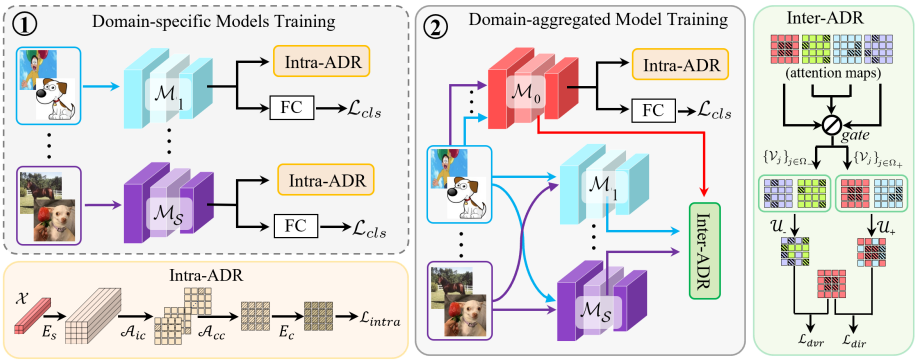

Fig. 2. The pipeline of our proposed *Attention Diversification* framework, which is composed of Intra-ADR and Inter-ADR.

training domain, and $\mathcal{M}_0$ represents the later which is trained on multiple training domains. For the image classification task, the cross-entropy loss is employed as supervision:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\mathcal{M}_j(x_i^d), y_i^d) \tag{1}$$

3.1 Intra-Model Attention Diversification Regularization

In this section, we introduce the design of Intra-ADR, which forces different channels to pay their most salient attention to different spatial locations, leading all spatial locations to be activated. To this end, potential features at all spatial regions are learned as much as possible. Intra-ADR is equipped upon the feature maps $\mathcal{X}^B$ for the last convolutional block.

In-Channel Attention Map. We perform normalization to each channel in $\mathcal{X}^B$ via spatial softmax, and obtain in-channel attention maps for different channels. In doing so, the maximum of the sum of all in-channel attention maps is permanently fixed as 1:

$$\mathcal{A}_{ic}(\mathcal{X}^B_{c,h,w}) = \frac{exp(\mathcal{X}^B_{c,h,w})}{\sum_{h=1}^{\mathcal{H}^B}\sum_{w=1}^{\mathcal{W}^B} exp(\mathcal{X}^B_{c,h,w})} \tag{2}$$

where $\mathcal{A}_{ic}(\cdot)$ is the operation of spatial softmax, and $\mathcal{X}^B_{c,h,w}$ denotes the pixel at the spatial location (h, w) of the c-th channel in $\mathcal{X}^B$. In this way, when the magnitudes of the selected pixels are enhanced, pixels in the remaining spatial location will be suppressed conversely. This means that attention is concentrated on the selected pixels and then we obtain "sparse" in-channel attention maps.

Cross-Channel Attention Map. Inspired by the maxout operation in [17], we enforce Pixel-wise Cross-Channel Maximization upon $\mathcal{A}_{ic}$ to obtain the cross-

channel attention map:

$$\mathcal{A}_{cc}(\mathcal{X}^B_{c,h,w}) = \max_{c=1,2,...,C^B} \mathcal{A}_{ic}(\mathcal{X}^B_{c,h,w}) \tag{3}$$

where $\mathcal{A}_{cc}(\cdot)$ is the Pixel-wise Cross-Channel Maximization, and $\mathcal{A}_{cc}(\mathcal{X}^B_{c,h,w}) \in \mathbb{R}^{\mathcal{H}^B \times \mathcal{W}^B}$ contains the most representative pixels across different channels. Thus, when we maximize the sum of $\mathcal{A}_{cc}(\mathcal{X}^B_{c,h,w})$, all spatial locations are extremely activated to achieve "*dense*" features.

Spatial-Channel Joint Expanding Module. However, the involving channels in the cross-channel attention map are limited because of the following observation: take ResNet-50 with input size of 224 [19] as an example, the number of spatial location of $\mathcal{X}^B$ ($\mathcal{H}^B \times \mathcal{W}^B = 49$) is far less than that of channels ($\mathcal{C}^B = 2048$). The majority of channels are not involved in regularization, leading to that lots of features cannot get sufficient attention. To remedy this issue, we propose a Spatial-Channel joint Expanding module (SCE) to enlarge both the spatial scope and involved channel number in the cross-channel attention map. SCE consists of two strategies:

- *Spatial Expanding.* The spatial expanding block is composed of a deconvolutional layer, an instance normalization layer and a ReLU activation layer. This is done for two-folder reasons: i) deconvolution can enlarge the resolution of feature maps to offset the gap between channel number and spatial location number; ii) deconvolution can provide more detailed semantic clues. The output of spatial expanding block $\mathbb{X}^B$ can be expressed as:

$$\mathbb{X}^B = E_s(\mathcal{X}^B_{c,h,w}) \tag{4}$$

  where E_s is the spatial expanding block, $\mathbb{X}^B \in \mathbb{R}^{\mathcal{C}^B \times \mathbf{H}^B \times \mathbf{W}^B}$, $\mathbf{H}^B = s * \mathcal{H}^B$ and $\mathbf{W}^B = s * \mathcal{W}^B$. $s > 1$ is a scale factor.
- *Channel Expanding.* We involve more channels in the cross-channel attention map via the Pixel-wise Cross-Channel Top-k selection, which averages the most activated k pixels across channels:

$$\mathbb{A}_{cc}(\mathbb{X}^B_{c,h,w}) = E_c(\mathcal{A}_{ic}(\mathbb{X}^B_{c,h,w})) = \max_{c=1,2,...,\mathcal{C}^B}(k)\ \mathcal{A}_{ic}(\mathbb{X}^B_{c,h,w}) \tag{5}$$

  where k is the number of selected channels, $\mathbb{A}_{cc}(\mathbb{X}_{c,h,w})$ is the output of SCE, and $\max(k)(\cdot)$ is the operation of averaging the most activated k pixels across channels. Note that SCE can make the cross-channel attention map involve $k \cdot s$ times channels compared with the original one.

Intra-Model Regularization. We impose SCE upon the output feature maps from the last convolutional block, and formulate the Intra-ADR term via maximizing the average value of $\mathbb{A}_{cc}(\mathbb{X}^B_{c,h,w})$:

$$\mathcal{L}_{intra} = -\frac{1}{\mathbf{HW}} \sum_{h=1}^{\mathbf{H}} \sum_{w=1}^{\mathbf{W}} \max_{c=1,2,...,\mathcal{C}^B}(k)\ \mathcal{A}_{ic}(\mathbb{X}^B_{c,h,w}) \tag{6}$$

3.2 Inter-Model Attention Diversification Regularization

In this section, we introduce the other component of our proposed framework, i.e., Inter-ADR, which refines the attention assignment of Intra-ADR by a paradigm of "*simulate, divide and assemble*". This is done to diversify task-related features yet suppress the domain-related features. Details are as follows:

Simulate Domain Shift. We train the domain-specific models $\{\mathcal{M}_j\}_{j=1}^{\mathcal{S}}$ on each source domain with $\mathcal{L}_{intra}$ in Eqn. 6 and $\mathcal{L}_{cls}$ in Eqn. 1. We do not only infer each sample on its own domain-specific model but also on other domain-specific models to simulate domain shift. Then we obtain the cross-channel attention maps for the b-th block $\{\{\mathcal{V}_j^b\}_{b=1}^{B}\}_{j=0}^{\mathcal{S}}$ from each domain-specific model in the same manner of Intra-ADR (without spatial expanding block) :

$$\mathcal{V}_j^b = \max_{c=1,2,...,\mathcal{C}^b}(\mathcal{C}^b)\ \mathcal{A}_{ic}(\mathcal{X}_j^b) \tag{7}$$

Divide Attention Maps Across Models. The models prediction $\hat{y}_j$ is utilized as the criterion to divide cross-channel attention maps $\{\{\mathcal{V}_j^b\}_{b=1}^{B}\}_{j=0}^{\mathcal{S}}$ from different models. $\{\{\mathcal{V}_j^b\}_{b=1}^{B}\}_{j=0}^{\mathcal{S}}$ are divided into two groups: if the prediction $\hat{y}_j$ agrees with the corresponding ground truth, $\{\mathcal{V}_j^b\}_{j=1}^{B}$ are viewed as task-related features, otherwise domain-related features. Let $\{\{\mathcal{V}_j^b\}_{b=1}^{B}\}_{j\in\Omega_+}$ and $\{\{\mathcal{V}_j^b\}_{b=1}^{B}\}\}_{j\in\Omega_-}$ denote the two groups respectively:

$$j \in \begin{cases} \Omega_+, \text{ if } \quad \hat{y}_i^j = y_i^d \\ \Omega_-, \text{ otherwise} \end{cases} \quad s.t. \quad j = 0, ..., \mathcal{S} \tag{8}$$

Assemble Attention Maps in Each Group. We assemble the cross-channel attention maps for each group via Pixel-wise Cross-Model Maximization, which is similar to Pixel-wise Cross-Channel Maximization in Eqn. 3:

$$\mathcal{U}_+^b = \max_{j\in\Omega_+} \mathcal{V}_j^b, \quad \mathcal{U}_-^b = \max_{j\in\Omega_-} \mathcal{V}_j^b \tag{9}$$

where $\mathcal{U}_+^b$ and $\mathcal{U}_-^b$ are the task-related and domain-related inter-model attention map, respectively. Thanks to the Pixel-wise Cross-model Maximization, $\mathcal{U}_+^b$ contains appropriate attention regions attributed to correct predictions under domain shift. On the other hand, $\mathcal{U}_-^b$ includes the most salient attention locations, which involve the domain-related features leading to error predictions.

Inter-Model Regularization. After dividing and assembling the attention maps across models, we exploit the inter-model attention map to force the cross-channel attention maps $\{\mathcal{V}_0^b\}_{b=1}^{B}$ of the domain-aggregated model to encourage task-related features yet suppress domain-related features.

Algorithm 1: Attention Diversification Training Schema

input : training data $\{\mathcal{D}\}_{d=1}^{\mathcal{S}}$, domain-specific models $\{\mathcal{M}_j\}_{j=1}^{\mathcal{S}}$, domain-aggregated model $\mathcal{M}_0$
output: trained domain-aggregated model $\mathcal{M}_0$;

for *d in* $\{1, ..., \mathcal{S}\}$ **do**
  Train domain-specific models $\mathcal{M}_{j=d}$ on the domain $\mathcal{D}_d$ with cross-entropy $\mathcal{L}_{cls}^d$ in Eqn. (1) and Intra-ADR losses $\mathcal{L}_{intra}^d$ in Eqn. (6) ;
end
for $\{x_i^d, y_i^d\}$ *in* $\{\mathcal{D}\}_{d=1}^{\mathcal{S}}$ **do**
  Generate predictions and cross-channel attention maps for domain-specific and domain-aggregated models: $\{\hat{y}_i^j = \mathcal{M}_j(x_i^d)\}_{j=0}^{\mathcal{S}}\}$;$\{\{\mathcal{V}_j^b\}_{b=1}^B\}_{j=0}^{\mathcal{S}}\}$;
  Divide cross-channel attention maps from multiple models into two group based on whether the prediction agrees with the ground truth using Eqn. (8): $\{\{\mathcal{V}_j^b\}_{b=1}^B\}_{j\in\Omega_+}, \{\{\mathcal{V}_j^b\}_{b=1}^B\}_{j\in\Omega_-} \leftarrow \{\{\mathcal{V}_j^b\}_{b=1}^B\}_{j=0}^{\mathcal{S}}\}$;
  Generate task-related and domain-related inter-model attention maps $\{\mathcal{U}_+^b\}_{b=1}^B$ and $\{\mathcal{U}_-^b\}_{b=1}^B$ using Eqn. (9);
  Calculate $\mathcal{L}_{dir}$ and $\mathcal{L}_{dvr}$ using Eqn. (10) and Eqn. (11);
  Train domain-aggregated model $\mathcal{M}_0$ with $\mathcal{L}_{total}$ in Eqn. (13)
end

On the one hand, we minimize the Euclidean distance between $\{\mathcal{V}_0\}_{b=1}^B$ and $\{\mathcal{U}_+^b\}_{b=1}^B$ to enhance attention regions involving task-related features:

$$\mathcal{L}_{dir} = \sum_{b=1}^{B} \|(\mathcal{V}_0^b - \mathcal{U}_+^b)\|_2 \tag{10}$$

On the other hand, we ought to suppress the attention regions involving domain-related features, which are accidentally included during Intra-ADR. This is done through maximizing the Euclidean distance between $\{\mathcal{V}_0\}_{b=1}^B$ and $\{\mathcal{U}_-^b\}_{b=1}^B$:

$$\mathcal{L}_{dvr} = -\sum_{b=1}^{B} ||(\mathcal{V}_0^b - \mathcal{U}_-^b)||_2 \tag{11}$$

Therefore, the Inter-ADR term can be expressed as:

$$\mathcal{L}_{inter} = \lambda_{dir} \cdot \mathcal{L}_{dir} + \lambda_{dvr} \cdot \mathcal{L}_{dvr} \tag{12}$$

where λ_{dir} and λ_{dvr} are the hyperparameters to balance the two losses.

Training Scheme. Our framework is trained in a two-stage manner, which includes domain-specific models training and domain-aggregated-model training as shown in Algorithem 1. In the first stage, only Eqn. 6 is used for attention diversification. In the second stage, both Eqn. 6 and Eqn. 12 are involved into attention diversification training. The total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{intra} \cdot \mathcal{L}_{intra} + \lambda_{dir} \cdot \mathcal{L}_{dir} + \lambda_{dvr} \cdot \mathcal{L}_{dvr} \tag{13}$$

Table 1. Leave-one-domain-out generalization results on PACS dataset.

Methods	References	Art	Cartoon	Photo	Sketch	Avg.	Art	Cartoon	Photo	Sketch	Avg.
		ResNet-18					*ResNet-50*				
Baseline	-	79.0	74.3	94.9	71.4	79.9	86.2	78.7	97.6	70.6	83.2
MetaReg [1]	NeurIPS'18	83.7	77.2	95.5	70.3	81.7	87.2	79.2	97.6	70.3	83.6
MASF [12]	NeurIPS'19	80.2	77.1	94.9	71.6	81.0	82.8	80.4	95.0	72.2	82.6
Epi-FCR [26]	ICCV'19	82.1	77.0	93.9	73.0	81.5	-	-	-	-	-
JiGen [4]	CVPR'19	79.4	75.2	96.0	71.3	80.5	-	-	-	-	-
DMG [5]	ECCV'20	76.9	80.4	93.4	75.2	81.5	82.6	78.1	94.5	78.3	83.4
RSC [21]	ECCV'20	84.4	80.3	95.9	80.8	85.1	87.8	82.1	97.9	83.3	87.9
MixStyle [78]	ICLR'21	84.1	78.8	96.1	75.9	83.7	-	-	-	-	-
SelfReg [24]	ICCV'21	82.3	78.4	**96.2**	77.5	83.6	87.9	79.4	96.8	78.3	85.6
DAML [50]	CVPR'21	83.0	74.1	95.6	78.1	82.7	-	-	-	-	-
SagNet [42]	CVPR'21	83.6	77.7	95.5	76.3	83.3	81.1	75.4	95.7	77.2	82.3
FACT [65]	CVPR'21	85.4	78.4	95.2	79.2	84.5	**89.6**	81.7	96.8	84.4	88.1
Intra-ADR	Ours	82.4	79.4	95.3	82.3	84.9	87.7	81.2	97.1	83.8	87.5
I^2-ADR	Ours	82.9	80.8	95.0	83.5	85.6	88.5	83.2	95.2	**85.8**	88.2
MixStyle + Intra-ADR	Ours	**86.0**	80.3	96.0	84.4	86.7	88.6	83.2	98.0	85.2	88.7
MixStyle + I^2-ADR	Ours	85.3	**81.2**	95.4	**86.1**	**87.0**	87.7	**84.5**	**98.2**	85.6	**89.2**

4 Experiments

In this section, we demonstrate the effectiveness of our *Attention Diversi cation* framework on three mainstream DG benchmarks. For convenience, we abbreviate our *Attention Diversi cation* framework to I^2-ADR.

4.1 Experimental Setup

Datasets. **PACS** [27] is a common DG benchmark that contains 9991 images of 7 categories from 4 different domains, i.e., Art (A), Cartoon (C), Photo (P), and Sketch (S). **Office-home** [60] contains images sharing 65 categories from 4 different domains around 15,579 images, i.e., Art (Ar), Clipart (Cl), Product (Pr), and Real-World (Rw). **DomainNet** [44] is a very large DG dataset, consisting of about 600K images with 345 categories from 6 different domains, i.e., Clipart (C), Infograph (I), Painting (P), Quickdraw (Q), Real (R), Sketch (S).

Implementation Details. Our framework is trained from ImageNet [11] pretrained models. We utilize an SGD optimizer, batch size of 64 and weight-decay of 0.0004 with 150 epochs for optimization. The initial learning rate is set 0.008 and adjusted by a cosine annealing schedule. Following the standard augmentation protocol in [4], we train our framework with horizontal flipping, random cropping, color jittering, and grayscale conversion. We follow the standard data splits and leave-one-domain-out evaluation protocol as the prior work [4]. The best models are selected based on the validation split of training domains. The accuracy for test domains is reported and averaged over three runs. We mainly use ResNet-18/50 as the backbones. Note that the same backbone is adopted among the domain-aggregated and the domain-specific models. We set the hyper-parameters, λ_{intra}, λ_{dir}, and λ_{dvr} as 0.005, 2 and 1 for all datasets, respectively. Our framework is implemented with PyTorch on NVIDIA Tesla V100 GPUs.

Table 2. Office-Home under ResNet-18.

Methods	Ar	Cl	Pr	Rw	Avg.
ResNet-18					
Baseline	57.8	52.7	73.5	74.8	64.7
RSC [21]	58.4	47.9	71.6	74.5	63.1
MixStyle [78]	58.7	53.4	74.2	75.9	65.5
SagNet [42]	60.2	45.4	70.4	73.4	62.3
FACT [65]	60.3	54.9	74.5	**76.6**	66.6
Intra-ADR	64.5	54.0	73.9	74.7	66.8
I^2-ADR	66.4	53.3	74.9	75.3	67.5
MixStyle + Intra-ADR	65.9	55.3	74.3	75.1	67.7
MixStyle + I^2-ADR	**66.8**	**56.8**	**75.3**	75.7	**68.7**

Table 3. Office-Home under ResNet-50.

Methods	Ar	Cl	Pr	Rw	Avg.
ResNet-50					
Baseline	61.3	52.4	75.8	76.6	66.5
MLDG [28]	61.5	53.2	75.0	77.5	66.8
RSC [21]	50.7	51.4	74.8	75.1	65.5
SelfReg [24]	63.6	53.1	76.9	78.1	67.9
SagNet [42]	63.4	54.8	75.8	78.3	68.1
Intra-ADR	67.3	54.1	78.8	78.8	69.8
I^2-ADR	70.3	55.1	80.7	79.2	71.4
MixStyle + Intra-ADR	69.5	55.9	80.6	80.4	71.4
MixStyle + I^2-ADR	**71.1**	**56.9**	**81.8**	**80.5**	**72.5**

4.2 Comparison with State-of-The-Arts

Results on PACS. Our framework achieves SOTA results on PACS dataset with both ResNet-18 and ResNet-50. In Table 1, the average performance of our framework achieves 85.6% and 88.2% with ResNet-18 and ResNet-50, respectively. Our framework provides impressive improvements of 5.7% and 5.0% compared with the corresponding baselines. Compared with other SOTA results, our framework surpasses other competing DG methods. Note that one component of our framework for recalling overlooked features in shortcut learning, Intra-ADR, can be surprisingly superior to most of DG methods. Moreover, Inter-ADR can further lift the DG performance of Intra-ADR and other competing DG methods.

Results on Office-Home. From Table 2 and Table 3, it can be observed that the baseline has a strong performance on Office-Home. Many previous DG methods cannot improve or perform worse than the baseline. Nevertheless, our framework achieves 67.5% and 71.4% with ResNet-18 and ResNet-50, respectively. Moreover, the proposed I^2-ADR surpasses the majority of other related methods, including the latest MixStyle [78], RSC [21], and FACT [65]. Notably, the Intra-ADR can achieve SOTA results of 66.8% and 69.8% on Office-Home with ResNet-18 and ResNet-50, respectively. Results on Office-Home justify the impact of each component of our framework.

Results on DomainNet. DomainNet is a very challenging large-scale dataset. The comparisons between our framework and other DG methods are reported in Table 4. The number of data in DomainNet is much larger than other DG benchmarks, leading to be very challenging to use ResNet-18 as the backbone. Fortunately, our framework using ResNet-18 achieves competing results on DomainNet. In addition, the performance of Intra-ADR and I^2-ADR using ResNet-50 are among the top ones. We notice that the performance of our framework exceeds that of SelfReg [24] by 1.6% and DMG [5] by 0.4%, respectively. This again verifies the superiority of our framework.

Table 4. Leave-one-domain-out generalization results on DomainNet dataset.

Methods	References	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
ResNet-18								
Baseline	-	57.1	17.6	43.2	13.8	54.9	39.4	37.6
MetaReg [1]	NeurIPS'18	53.7	**21.1**	**45.3**	10.6	**58.5**	42.3	38.6
DMG [5]	ECCV'20	**60.1**	18.8	44.5	**14.2**	54.7	41.7	**39.0**
Intra-ADR	Ours	57.3±0.1	14.9±0.3	42.8±0.2	12.2±0.4	52.9±0.5	46.0±0.2	37.7
I^2-ADR	Ours	57.3±0.3	15.2±0.3	44.1±0.1	12.1±0.4	53.9±0.6	46.7±0.2	38.2
MixStyle + Intra-ADR	Ours	57.4±0.2	15.3±0.1	43.3±0.2	12.3±0.4	53.5±0.3	46.5±0.2	38.1
MixStyle + I^2-ADR	Ours	57.4±0.4	15.7±0.2	44.7±0.1	12.3±0.4	54.4±0.2	**47.4±0.1**	38.7
ResNet-50								
Baseline	-	62.2	19.9	45.5	13.8	57.5	44.4	40.5
MetaReg [1]	NeurIPS'18	59.8	**25.6**	**50.2**	11.5	**64.6**	50.1	43.6
MLDG [28]	AAAI'18	59.1±0.2	19.1±0.3	45.8±0.7	13.4±0.3	59.6±0.2	50.2±0.4	41.2
C-DANN [31]	ECCV'18	54.6±0.4	17.3±0.1	43.7±0.9	12.1±0.7	56.2±0.4	45.9±0.5	38.3
RSC [21]	ECCV'20	55.0±1.2	18.3±0.5	44.4±0.6	12.2±0.2	55.7±0.7	47.8±0.9	38.9
DMG [5]	ECCV'20	**65.2**	22.2	50.0	**15.7**	59.6	49.0	43.6
SagNet [42]	CVPR'21	57.7±0.3	19.0±0.2	45.3±0.3	12.7±0.5	58.1±0.5	48.8±0.2	40.3
SelfReg [24]	ICCV'21	60.7±0.1	21.6±0.1	49.4±0.2	12.7±0.1	60.7±0.1	51.7±0.1	42.8
Intra-ADR	Ours	63.6±0.1	20.0±0.1	49.4±0.1	14.8±0.3	60.0±0.4	**54.4±0.1**	43.7
I^2-ADR	Ours	64.4±0.2	20.2±0.6	49.2±0.5	15.0±0.2	61.6±0.4	53.3±0.1	44.0
MixStyle + Intra-ADR	Ours	63.9±0.1	20.1±0.5	49.4±0.2	15.0±0.4	60.4±0.3	**54.4±0.1**	43.9
MixStyle + I^2-ADR	Ours	64.1±0.1	20.4±0.2	49.2±0.4	15.1±0.2	61.3±0.4	54.3±0.4	**44.1**

Table 5. Ablation studies on the three components contained in I^2-ADR.

Method	$\mathcal{L}_{intra}$	$\mathcal{L}_{dir}$	$\mathcal{L}_{dvr}$	Art	Cartoon	Photo	Sketch	Avg.
I^2-ADR	✓	-	-	82.4	79.4	**95.3**	82.3	84.9
	-	✓	✓	82.3	80.0	95.1	82.6	85.0
	✓	✓	-	82.7	80.5	95.0	83.2	85.4
	✓	-	✓	82.5	80.2	95.1	82.9	85.2
	✓	✓	✓	**82.9**	**80.8**	95.0	**83.5**	**85.6**

Table 6. Ablation studies on two strategies in SCE module.

Method	E_s	E_c	Art	Cartoon	Photo	Sketch	Avg.
Intra-ADR	-	-	81.3	77.3	94.7	78.8	83.0
	-	✓	80.0	77.2	**96.0**	80.9	83.5
	✓	-	81.9	79.3	95.5	79.3	84.0
	✓	✓	**82.4**	**79.4**	95.3	**82.3**	**84.9**

4.3 Ablation Studies

In this section, we carry out various ablation studies to dissect the effectiveness of our proposed *Attention Diversification* framework. All ablation studies are conducted on PACS dataset with ResNet-18.

Analysis for SCE Module. We conduct ablation studies on SCE module to analyze the effectiveness of each component in SCE module. There are two critical strategies, Spatial Expanding (E_s) and Channel Expanding (E_c) designed to facilitate the effectiveness of the Intra-ADR on the high-level features. As shown in Table 6, we can observe that both E_s and E_c can improve the ability of generalization across domains, and the gains of E_s are slightly better than E_c. Note that the performance w/ E_c on (A, C) is indeed worse than that w/o E_c (baseline Intra-ADR), but the performance w/ E_c is better than the baseline on average. On the other hand, we analyze the effect of the two crucial hyper-parameters in SCE Module, the scale factor s in E_s and the selected channels number k in E_c. As shown in Table 10, the average performance increases as the channel number is expanded. Besides, The scale factor also has a significant impact on the effectiveness of Intra-ADR. We set $k = 10$, $s = 2$ as the default setting for all experiments.

Table 7. Ablation studies on the equipped positions of Intra-ADR.

Methods	Art	Cartoon	Photo	Sketch	Avg.
Intra-ADR (res1)	80.8	78.4	94.7	80.1	83.5
Intra-ADR (res2)	81.0	78.6	94.9	80.7	83.8
Intra-ADR (res3)	81.8	78.8	95.5	81.1	84.3
Intra-ADR (res4)	**82.4**	**79.4**	**95.3**	**82.3**	**84.9**

Table 8. Ablation studies on the equipped positions of Inter-ADR.

Methods	Art	Cartoon	Photo	Sketch	Avg.
Intra-ADR	82.4	79.4	95.3	82.3	84.9
+ Inter-ADR (res1)	**83.0**	79.3	**95.5**	82.9	85.2
+ Inter-ADR (res12)	82.6	80.0	95.3	83.0	85.2
+ Inter-ADR (res123)	82.7	80.5	95.1	83.2	85.4
+ Inter-ADR (res1234)	82.9	**80.8**	95.0	**83.5**	**85.6**

Table 9. The Seen Domain performance of our proposed method.

Methods	Backbone	C&P&S	A&P&S	A&C&S	A&C&P
Baseline	ResNet-18	**96.4**	95.6	95.0	95.6
Intra-ADR		96.2	**96.5**	94.9	96.3
I^2-ADR		96.3	96.3	**95.1**	**96.6**
Baseline	ResNet-50	96.9	**96.8**	95.6	97.3
Intra-ADR		97.4	96.7	95.5	**97.4**
I^2-ADR		**97.6**	96.7	**97.2**	97.1

Table 10. Ablation studies on the scale factor s and selected channels number k.

(k, s)	Art	Cartoon	Photo	Sketch	Avg.
(2, 2)	82.0	78.1	95.1	81.0	84.1
(2, 4)	82.2	77.9	**96.0**	81.0	84.3
(4, 2)	82.1	78.2	95.0	81.2	84.2
(4, 4)	82.4	77.9	**96.0**	81.0	84.5
(10, 2)	**82.4**	**79.4**	95.3	**82.3**	**84.9**
(10, 4)	**82.4**	79.2	95.6	82.2	**84.9**

Analysis for Different Losses in I^2-ADR. We conduct ablation studies to investigate the effectiveness of the three losses in I^2-ADR. As shown in Table 5, the first row denotes the results of Intra-ADR, the second row denotes the results of Inter-ADR. We can observe that $\mathcal{L}_{dir}$ + $\mathcal{L}_{dvr}$ contributes to the impressive improvement of 0.7% on the average performance compared with Intra-ADR. After removing $\mathcal{L}_{dir}$ which diversifies task-related attention regions, there is a drop of 0.4% compared with $\mathcal{L}_{dir}$ + $\mathcal{L}_{dvr}$, but still an improvement of 0.3% compared with Intra-ADR. Besides, after removing $\mathcal{L}_{dvr}$ that suppresses the domain-related attention regions, $\mathcal{L}_{dir}$ solely surpasses the Intra-ADR by 0.5%, but losses 0.2% compared with $\mathcal{L}_{dir}$ + $\mathcal{L}_{dvr}$.

Analysis for The Positions of Intra-ADR and Inter-ADR. Extensive works have been discussed that different layers of CNNs have different effects on the information flow [37, 38, 55]. Here we also analyze the equipped positions of the proposed two modules, Intra-ADR and Inter-ADR. Let's denote 4 bottleneck stages of a standard ResNet backbone as *res1-4*. For instance, *res1* means the outputted feature maps of the first bottleneck stage. As shown in Table 7, Intra-ADR is limited when equipped upon the low- and middle-level feature maps, but provides an impressive improvement when upon the high-level. Besides, the hierarchical Inter-ADR achieves a significant impact on the average performance. Thereby, Intra-ADR is ewuipped on the highest layer to diversify task-related features instead of introducing too many domain-related features. Inter-ADR is equipped upon multi-level layers to facilitate information flow with a mechanism of distinguishing the task- and domain-related features.

4.4 Discussions and Visualization

Performance on Seen Domains. In this section, we report the performance of our framework on the seen domains from PACS dataset. As shown in Table

Table 11. Performance comparison on single-source DG. We train our methods with a single source domain and evaluate with other remaining target domains.

	Target Domain																			
Source	**Baseline**					**RSC** [21] (ECCV'20)					**SelfReg** [24] (ICCV'21)					**Intra-ADR** (Ours)				
Domain	A	C	P	S	Avg.	A	C	P	S	Avg.	A	C	P	S	Avg.	A	C	P	S	Avg.
A	-	61.3	96.1	52.3	69.9	-	62.5	96.3	53.2	70.7	-	65.2	96.6	55.9	72.6	-	64.8	94.4	64.3	**74.5**
C	64.1	-	81.8	75.8	73.9	69.0	-	85.9	70.4	75.1	72.1	-	87.5	70.1	**76.6**	66.7	-	83.8	74.9	75.1
P	66.1	29.5	-	32.3	41.6	66.3	26.5	-	32.1	41.6	67.7	29.0	-	33.7	43.5	67.8	40.3	-	39.5	**49.2**
S	38.6	60.5	48.0	-	48.3	38.0	56.4	47.4	-	47.3	37.2	54.0	46.1	-	45.8	42.7	61.5	46.6	-	**50.3**
Avg.	56.2	50.4	74.6	53.4	58.6	57.8	48.5	76.5	51.9	58.7	59.0	47.4	**76.7**	53.3	59.6	**59.1**	**55.5**	74.9	**59.6**	**62.3**

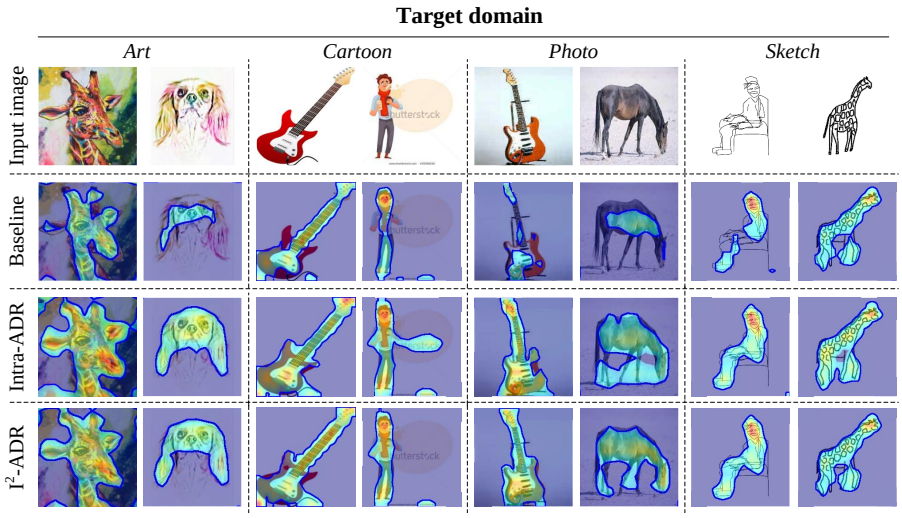

Fig. 3. Attention visualization on the testing domains of PACS with ResNet-18.

9, "A","C","P","S" in the first row represent the classification accuracy on seen domains, including Art, Cartoon, Photo and Sketch, respectively. The performance of our framework, whether Intra-ADR or I^2-ADR, surpasses that of the baseline on almost all sub-tasks using ResNet-18 and ResNet-50, respectively. This verifies that our framework also improves the in-domain generalization.

Single-Source Domain Generalization. We further evaluate our framework on single-source DG tasks. Since the Inter-ADR is not suitable for single-source DG tasks, we only report the results of Intra-ADR. Results are reported as the average accuracy among single source-target pairs. As shown in Table 11, the performance of Intra-ADR is among the top ones. This indicates that the Intra-ADR can handle both the multiple-source and single-source DG tasks, and demonstrate that diverse features effectively can avoid shortcut learning.

Orthogonality to Other DG Methods. Our method can also boost the performance of other DG methods. As shown in Table 1-4, a new SOTA performance is achieved by combining our framework with MixStyle [78] and is superior to other competing DG works by a significant margin.

Attention Visualization. We visualize the attention maps to verify our motivation and the effectiveness of our framework. The attention maps on samples from testing split of the 4 domains in PACS are shown in Fig. 3. The hotter colors denote the more salient attention value, while the cooler colors represent the lower value. To compare the differences of attention regions between the baseline and our framework more clearly, we retain the top normalized attention values ($>= 0.7$). We can see that the proposed Intra-ADR *de-facto* pays sufficient attention to diverse spatial locations, including the task-related regions and some domain-related features. Fortunately, Inter-ADR can suppress the domain-related regions and enhance the task-related regions.

Limitations. As shown in the last row of Fig. 3, there still exist some risks to maintain/enhance domain-related features in some cases. Although Inter-ADR is utilized to suppress domain-related features brought by baseline and Intra-ADR, which exploits the prediction to determine task- and domain-related features, the domain-related features will be maintained/enhanced once the corresponding cross-domain prediction is consistent with the ground-truth. Nevertheless, the proposed Intra-ADR and Inter-ADR boost the DG performance on average. The existing limitations are left as the future works.

5 Conclusion

Investigated from the perspective of shortcut learning, the models trained on different domains will pay attention to different salient features, aka domain attention bias. However, the principle of maximum entropy hints that every task-related feature is equally-useful potentially when encountering unseen domains. This novel insight enlightens us to remedy the issue of DG via Attention Diversification, in which we organically unify the Intra-ADR and Inter-ADR into our framework: we first utilize Intra-ADR to coarsely recall task-related features in the highest layer as much as possible, and then exploit Inter-ADR to delicately distinguish domain- and task-related features in multiple intermediate layers for further suppression and enhancement, respectively.

Acknowledgements

This work was sponsored by National Natural Science Foundation of China (62106220, U20B2066), Hikvision Open Fund (CCF-HIKVISION OF 20210002), NUS Faculty Research Committee (WBS: A-0009440-00-00), and MOE Academic Research Fund (AcRF) Tier-1 FRC Research Grant of Singapore (WBS: A-0009456-00-00).

References

1. Balaji, Y., Sankaranarayanan, S., Chellappa, R.: Metareg: Towards domain generalization using meta-regularization. In: NeurIPS (2018)
2. Ben-David, S., Blitzer, J., Crammer, K., Pereira, F.: Analysis of representations for domain adaptation. In: NIPS (2007)
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1), 151–175 (2010)
4. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2019)
5. Chattopadhyay, P., Balaji, Y., Hoffman, J.: Learning to balance specificity and invariance for in and out of domain generalization. In: ECCV (2020)
6. Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., Pu, S., Song, M., Zhuang, Y.: Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14381–14390 (2022)
7. Chen, M., Chen, W., Yang, S., Song, J., Wang, X., Zhang, L., Yan, Y., Qi, D., Zhuang, Y., Xie, D., et al.: Learning domain adaptive object detection with probabilistic teacher. In: ICML (2022)
8. Chen, W., Lin, L., Yang, S., Xie, D., Pu, S., Zhuang, Y., Ren, W.: Self-supervised noisy label learning for source-free unsupervised domain adaptation. arXiv preprint arXiv:2102.11614 (2021)
9. Chen, W., Xie, D., Zhang, Y., Pu, S.: All you need is a few shifts: Designing efficient convolutional neural networks for image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7241–7250 (2019)
10. Chen, W., Zhang, Y., Xie, D., Pu, S.: A layer decomposition-recomposition framework for neuron pruning towards accurate lightweight networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3355–3362 (2019)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
12. Dou, Q., Castro, D.C., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. In: NeurIPS (2019)
13. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
14. Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.: Channel interaction networks for fine-grained image categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10818–10825 (2020)
15. Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665–673 (2020)
16. Gong, K., Li, B., Zhang, J., Wang, T., Huang, J., Mi, M.B., Feng, J., Wang, X.: Posetriplet: Co-evolving 3d human pose estimation, imitation, and hallucination under self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
17. Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. In: ICML (2013)
18. Guiasu, S., Shenitzer, A.: The principle of maximum entropy. The mathematical intelligencer **7**(1), 42–48 (1985)

19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3588–3597 (2018)
21. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: ECCV (2020)
22. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: CVPR (2020)
23. Jing, Y., Liu, X., Ding, Y., Wang, X., Ding, E., Song, M., Wen, S.: Dynamic instance normalization for arbitrary style transfer. In: AAAI (2020)
24. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9619–9628 (October 2021)
25. Li, B., Wu, F., Lim, S., Belongie, S., Weinberger, K.Q.: On feature normalization and data augmentation. In: CVPR (2021)
26. Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y., Hospedales, T.M.: Episodic training for domain generalization. In: ICCV (2019)
27. 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)
28. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Learning to generalize: Meta-learning for domain generalization. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
29. Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., Zhuang, Y.: A free lunch for unsupervised domain adaptive object detection without source data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8474–8481 (2021)
30. 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)
31. Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., Tao, D.: Deep domain generalization via conditional invariant adversarial networks. In: ECCV (2018)
32. Li, Z., Zhao, L., Chen, W., Yang, S., Xie, D., Pu, S.: Target-aware auto-augmentation for unsupervised domain adaptive object detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3848–3852. IEEE (2022)
33. Lin, L., Liang, L., Jin, L., Chen, W.: Attribute-aware convolutional neural networks for facial beauty prediction. In: IJCAI. pp. 847–853 (2019)
34. Lin, L., Xie, H., Yang, Z., Sun, Z., Liu, W., Yu, Y., Chen, W., Yang, S., Xie, D.: Semi-supervised domain generalization in real world: New benchmark and strong baseline. arXiv preprint arXiv:2111.10221 (2021)
35. Liu, H., Yang, Y., Wang, X.: Overcoming catastrophic forgetting in graph neural networks. In: AAAI Conference on Artificial Intelligence (2021)
36. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML. pp. 97–105 (2015)
37. Matthew Zeiler, D., Rob, F.: Visualizing and understanding convolutional neural networks. ECCV (2014)
38. Meng, R., Chen, W., Xie, D., Zhang, Y., Pu, S.: Neural inheritance relation guided one-shot layer assignment search. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 5158–5165 (2020)

39. Meng, R., Chen, W., Yang, S., Song, J., Lin, L., Xie, D., Pu, S., Wang, X., Song, M., Zhuang, Y.: Slimmable domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7141–7150 (2022)
40. Moon, S., Carbonell, J.G.: Completely heterogeneous transfer learning with attention - what and wahr not to transfer. In: IJCAI (2017)
41. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: International Conference on Machine Learning. pp. 10–18. PMLR (2013)
42. Nam, H., Lee, H., Park, J., Yoon, W., Yoo, D.: Reducing domain gap by reducing style bias. In: CVPR (2021)
43. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: ECCV (2018)
44. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: ICCV (2019)
45. Peng, X., Saenko, K.: Synthetic to real adaptation with generaitve correlation alignment networks. In: WACV (2018)
46. Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., Birchfield, S.: Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In: ICRA (2019)
47. Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S.: Correlation-aware adversarial domain adaptation and generalization. Pattern Recognition **100** (2020)
48. Ren, S., Zhou, D., He, S., Feng, J., Wang, X.: Shunted self-attention via multi-scale token aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
49. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
50. Shu, Y., Cao, Z., Wang, C., Wang, J., Long, M.: Open domain generalization with domain-augmented meta-learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9624–9633 (2021)
51. Sicilia, A., Zhao, X., Hwang, S.J.: Domain adversarial neural networks for domain generalization: When it works and how to improve. arXiv:2102.03924 (2021)
52. Song, J., Shen, C., Lei, J., Zeng, A.X., Ou, K., Tao, D., Song, M.: Selective zero-shot classification with augmented attributes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 468–483 (2018)
53. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Springer, Cham (2018)
54. Sun, Z., Shen, Z., Lin, L., Yu, Y., Yang, Z., Yang, S., Chen, W.: Dynamic domain generalization. In: IJCAI (2022)
55. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. In: 2015 ieee information theory workshop (itw). pp. 1–5. IEEE (2015)
56. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IROS (2017)
57. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
58. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv:1412.3474 (2014)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
60. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)

61. Wang, F., Qian, M., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)
62. Wang, M., Wang, W., Li, B., Zhang, X., Lan, L., Tan, H., Liang, T., Yu, W., Luo, Z.: Interbn: Channel fusion for adversarial unsupervised domain adaptation. In: Proceedings of the 29th ACM international conference on multimedia. pp. 3691–3700 (2021)
63. Wang, X., Li, L., Ye, W., Long, M., Wang, J.: Transferable attention for domain adaptation. In: AAAI (2019)
64. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
65. 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)
66. Xue, Y., Mao, J., Niu, M., Xu, H., Mi, M.B., Zhang, W., Wang, X., Wang, X.: Point2seq: Detecting 3d objects as sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
67. Yang, X., Ye, J., Wang, X.: Factorizing knowledge in neural networks. In: European Conference on Computer Vision (2022)
68. Yang, Y., Feng, Z., Song, M., Wang, X.: Factorizable graph convolutional networks. In: Conference on Neural Information Processing Systems (2020)
69. Yang, Y., Qiu, J., Song, M., Tao, D., Wang, X.: Distilling knowledge from graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
70. Yang, Y., Qiu, J., Song, M., Tao, D., Wang, X.: Learning propagation rules for attribution map generation. In: European Conference on Computer Vision (2020)
71. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
72. Yuan, P., Chen, W., Yang, S., Xuan, Y., Xie, D., Zhuang, Y., Pu, S.: Simulation-and-mining: Towards accurate source-free unsupervised domain adaptive object detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3843–3847. IEEE (2022)
73. Zhao, Y., Zhong, Z., Luo, Z., Lee, G.H., Sebe, N.: Source-free open compound domain adaptation in semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology (2022)
74. Zhao, Y., Zhong, Z., Yang, F., Luo, Z., Lin, Y., Li, S., Sebe, N.: Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
75. Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., Lee, G.H.: Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
76. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. IEEE (2019)
77. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
78. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. In: ICLR (2021)

79. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Ding, S., Ma, L.: Adaptive mixture of experts learning for generalizable face anti-spoofing. In: Proceedings of the 30th ACM International Conference on Multimedia (2022)
80. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Sheng, K., Ding, S., Ma, L.: Generative domain adaptation for face anti-spoofing. In: European Conference on Computer Vision. Springer (2022)