Mix and Reason: Reasoning over Semantic Topology with Data Mixing for Domain Generalization

Chaoqi Chen[1] **Luyao Tang**[2] **Feng Liu**[3] **Gangming Zhao**[1] **Yue Huang**[2] **Yizhou Yu**[1*]
[1]The University of Hong Kong [2]Xiamen University [3]Deepwise AI Lab
cqchen1994@gmail.com, lytang@stu.xmu.edu.cn, liufeng@deepwise.com
gmzhao@connect.hku.hk, yhuang2010@xmu.edu.cn, yizhouy@acm.org

Abstract

Domain generalization (DG) enables generalizing a learning machine from multiple seen source domains to an unseen target one. The general objective of DG methods is to learn semantic representations that are independent of domain labels, which is theoretically sound but empirically challenged due to the complex mixture of common and domain-specific factors. Although disentangling the representations into two disjoint parts has been gaining momentum in DG, the strong presumption over the data limits its efficacy in many real-world scenarios. In this paper, we propose Mix and Reason (`MiRe`), a new DG framework that learns semantic representations via enforcing the structural invariance of semantic topology. `MiRe` consists of two key components, namely, Category-aware Data Mixing (CDM) and Adaptive Semantic Topology Refinement (ASTR). CDM mixes two images from different domains in virtue of activation maps generated by two complementary classification losses, making the classifier focus on the representations of semantic objects. ASTR introduces relation graphs to represent semantic topology, which is progressively refined via the interactions between local feature aggregation and global cross-domain relational reasoning. Experiments on multiple DG benchmarks validate the effectiveness and robustness of the proposed `MiRe`.

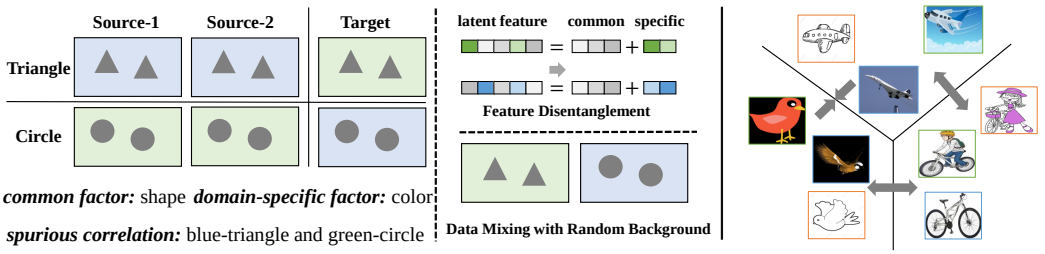
1 Introduction

In real-world scenarios, such as autonomous driving and computer-aided diagnosis, deep neural networks are expected to be trained on data collected from a small number of domains, but with the objective of being deployed across novel domains with different data distributions to reduce annotation costs and meet a broader range of needs. However, out-of-distribution data [27] does not satisfy the basic IID assumption of modern deep learning models and emerges as an inevitable challenge, significantly hindering the deployment of source-trained models in new target domains. Domain generalization (DG) [45, 6, 21], which addresses this challenge by learning invariant representations across multiple source domains, has attract a great deal of attention in the research community.

The objective of DG is to recover latent semantic factors that are independent of domain. Most DG approaches rely on the assumption that the latent representations can be divided into two disjoint components, *i.e.,* common and domain-specific components. Inspired by this, mainstream DG paradigms include data/feature augmentation [66, 73], feature disentanglement [51, 38], invariant risk minimization [3, 2], meta-learning [31, 15], to name a few. By facilitating common knowledge extraction and alleviating domain-specific components, these approaches have made tremendous progress in various tasks, such as image classification [6, 82] and semantic segmentation [35, 49].

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Ideal Case **(b)** Real-World Case

Figure 1: **(a)** In ideal case, we assume that the data has two attributes (shape and color). **(b)** In real-world case, we observe that bird and airplane share more similarity as the design of airplane is motivated by bird, while both of them are dissimilar with bicycle. Thus, the embedding of bird and airplane should be closer and their connecting weight should be enlarged, which are independent of domain and robust to unseen test environments. We call this property as *structural invariance*.

Despite the promise, existing DG approaches may still be confined by two critical limitations. (1) Traditional DG methods [45, 34, 6, 32, 15] strive to enforce invariance between domains but ignore the potential spurious correlations between common and specific components. Take Fig. 1 (a) as an example, the correlation between shape and color may induce the classifier to use background colors as the classification evidence. (2) To statistically avoid the spurious correlations, a body of research [52, 51, 14, 42, 59, 38] proposes to identify the common component through latent factor disentanglement. In this regard, causal graphs [48] provide promising theoretical guarantees. As shown in Fig. 1 (a), previous disentanglement-based methods manifest remarkable success in simulated data, requiring that models have seen some distribution of values for an attribute [69] (*e.g.,* shape and color in DSPRITES [44]), which is too restrictive to be satisfied. In real-world case (cf. Fig. 1 (b)), we cannot (*i*) explicitly know the latent feature is composed of how many factors, (*ii*) identify which factors are semantic-related or not, and (*iii*) how these potential factors interact. Since we have no a *priori* knowledge or assumptions about the attributes (factors), it is prohibitively difficult to disentangle them in an unsupervised manner [40], largely impeding the practical use of these approaches. On the other hand, the internal relations among different classes are neglected, which are crucial for the robustness of learned models to distribution shifts. This finding draws motivation from the human intelligence that when learning new concepts, humans are talented at *comparing* and *reasoning* [24, 13, 5]. In that sense, we aim to endow the classifier with the ability of perceiving and maintaining structural semantic relations rather than ideally seeking *clean* semantic[1] representations.

Grounded on the above discussions, we propose Mix and Reason (`MiRe`), a simple yet effective DG approach, which enables generalization via data-level augmentation and feature-level regularization. The proposed `MiRe` consists of two key components, Category-aware Data Mixing (CDM) and Adaptive Semantic Topology Refinement (ASTR), which learn semantic representations by enforcing structural invariance of semantic topology across domains in the embedding space. CDM induces the classifier to focus on recognizing foreground objects as their supporting evidence for classification, alleviating the potential influence of domain-specific factors (*e.g.,* scene layout and style). Specifically, foreground region is derived via integrating two complementary network activation maps generated by category and domain classification losses, while background is randomly cropped from images sampled from other domains with Gaussian blur. ASTR introduces the concept of semantic topology, which is instantiated as the relation graphs, to progressively refine the representations (node) and the topological relations (edge) of semantic anchors. Semantic anchor is denoted by the centroids of category-wise features in per domain, largely reducing the intra-domain variations. To be specific, each input instance is compared to the set of semantic anchors and dynamically aggregate features according to the global[2] semantic topology and local adjacent relations. Then, the global features are updated by the aggregated local features. Finally, a bipartite graph network is developed on the top of semantic topology to perform cross-domain relational reasoning and induce structural invariance.

[1]The word 'semantic' and 'common' are used interchangeably, but 'invariance' does not equal to them. Invariance may contain spurious correlation, which can be alleviated but is difficult to be absolutely eliminated.

[2]The word 'global' denotes that the features are computed from the whole domain (dataset), and the word 'local' denotes that the features are computed from mini-batch samples at each iterations.

Experiments on four standard DG benchmarks (PACS, VLCS, Office-Home, and DomainNet) reveal that `MiRe` exceeds the performance of state-of-the-art methods. In addition, CDM is applied as a plug-and-play module for state-of-the-art DG methods (such as MixStyle [82] and EFDM [76]) to boost their performance. The robustness of ASTR is further evaluated on a challenging Chest X-ray benchmark [42] that has explicit spurious correlation.

2 Related Work

Domain Generalization. Generalizing a well-trained model to novel environments with different data distributions is a challenging and long-standing machine learning problem [79, 67]. Current state-of-the-art DG approaches can be roughly categorized into four types. (1) *Feature Alignment.* Bridging the domain gap through statistic matching and adversarial learning [36, 34, 43] provides invariant representations, which are expected to be shared by novel domains. A major defect is that the learned representations is prone to mix common and domain-specific components, potentially resulting in strong bias towards spurious relations. (2) *Feature Disentanglement.* To address the above issue, disentangling latent feature into two disjoint parts has attract a great surge of interest. Prevailing approaches [52, 51, 14, 42, 59, 38] resort to causal graphs [48] to explicitly identify causal and non-causal factors with theoretical guarantees. If this could be perfectly achieved, learned models will generalize well under any circumstances. However, these theoretical results require strong assumptions, *e.g.,* the degree of diversity between causal and non-causal factors and the presence of an estimated causal graph [60], or a priori knowledge about the combinations of latent factors, *e.g.,* distribution of values for a certain attribute. The complex combinations of many real-world cases (*e.g.,* DomainNet [50]) greatly impede the practical uses. In that sense, a principled disentanglement solution is hard to be reached. (3) *Data/Feature Augmentation.* A simple yet effective approach is to augment the diversity of data or feature so as to avoid overfitting to training data [65, 73, 46, 82, 72, 76, 8]. Among them, learning and imposing heuristic style transmission strategies take the dominant positions, while explicitly separating and recombination images remain the boundary to explore. (4) *Meta-Learning.* Some of recent works [31, 37, 32, 15, 39] simulate the distribution shifts between seen and unseen environments by using meta-learning [19], which splits the training data into meta-train and meta-test domains.

Despite a proliferation of DG approaches, Wiles *et al.* [69] reveal that the best DG methods are not consistent over different shifts and datasets in which disentangling offers limited improvements in most real-world scenarios. In addition, Gulrajani and Lopez-Paz [21] cast doubt on the progress that have been made by comparing to a standard empirical risk minimization baseline, showing that conventional domain-invariant methods [58, 20] exhibit robust improvements. In this work, we embrace the strengths of cross-domain invariance by considering a more fine-grained and stable property—*structural invariance*—which is evaluated on a number of challenging benchmarks.

Data Mixing. Mixing data, such as Mixup [75, 64] and its various variants, has shown compelling results in standard supervised and semi-supervised learning. Mixup operations aim to conduct data interpolation via convex combinations of pairs of images and their labels. Instead, our CDM generates diverse training samples by replacing the background of a certain image with a randomly cropped patch from other images but keeps its object label fixed. CutMix [74] replaces a regularly-shaped image region with a patch from another training image and proportionally mix their ground-truth labels. CutPaste [29] cuts an image patch and randomly pastes it at an another image. BackErase [54] uses the object annotations to copy and pastes foreground objects on a background image. Compared to the aforementioned methods that cut a regularly-shaped image patch [74, 29] or rely on ground-truth object labels [54], the proposed CDM simultaneously leverage the category and domain information to derive an irregularly-shaped foreground regions while avoiding fine-grained annotations.

Cross-Domain Invariance. Seeking cross-domain invariance is crucial for out-of-distribution generalization problems, such as Unsupervised Domain Adaptation (UDA) [47] and DG. Traditional methods typically resort to statistical distribution divergence, such as maximum mean discrepancy [41] and second-order moment [58], for measuring domain-wise distribution shifts. More recently, a promising approach is to utilize the concept of class prototype for enforcing semantic consistency across domains [71, 12, 10]. This line of research has also been extensively explored in various downstream tasks. However, we argue that no matter domain- or class-wise invariance cannot guarantee the generalizable representations especially when encountering unseen test environments.

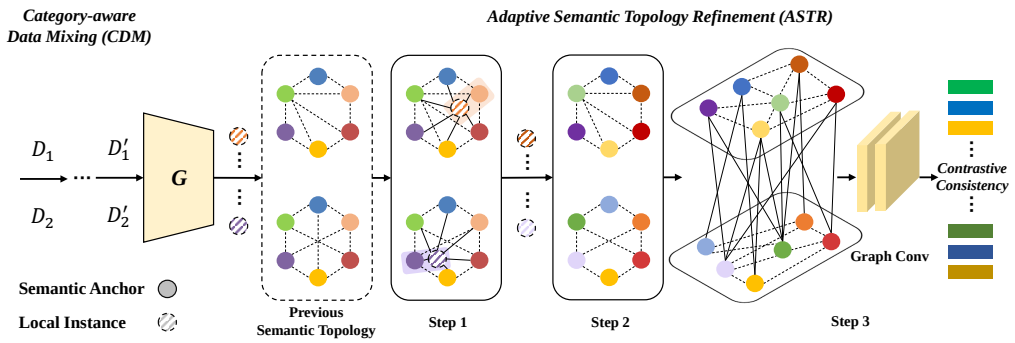


Figure 2: The pipeline of the proposed `MiRe`. Assume that we have two domains $\mathcal{D}_1$ and $\mathcal{D}_2$, CDM augments them to $\mathcal{D}'_1$ and $\mathcal{D}'_2$. In each iteration of ASTR, the embedding feature of each local instance performs feature aggregation based on the semantic topology in previous iteration and their local adjacent relations (**step 1**). Then, global anchor features are updated based on the aggregated local features (**step 2**). Finally, a bipartite graph learning procedure is imposed on the top of semantic topology to perform between-domain relational reasoning and induce structural invariance (**step 3**).

Such an invariance may be susceptible to include some misleading spurious correlation, such as wrongly associating an foreground object with specific background. The justification is that the topological structure of these prototypes is unexplored, making them being sensitive to the change of environments [42, 9, 77]. By contrast, the proposed `MiRe` models the graphical structures of different semantic anchors by means of feature aggregation and message-passing.

3 Methodology

In the task of DG, given that we have access to data from N ($N \geq 2$) source domains $\mathcal{D}_s = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_N\}$ that are sampled from different data distributions defined on the joint space $\mathcal{X} \times \mathcal{Y}$. The n-th domain can be defined as $\mathcal{D}_n = \{(x_i, y_i, d_i)\}_{i=1}^{N_d}$, where $y_i = \{1, 2, ..., K\}$ is the class label, $d_i = \{1, 2, ..., N\}$ is the domain label, and N_d is the number of samples in this domain. An overview of `MiRe` is illustrated in Fig. 2. `MiRe` consists of a CDM stage (Section 3.1) and an ASTR stage (Section 3.2), which jointly enable generalization on data-level and feature-level.

3.1 Category-aware Data Mixing

Most of DG works aim to learn predictive models with certain invariant properties that are independent of domain. As aforementioned, this sort of invariance may be brittle due to the presence of spurious correlation (cf. Fig. 1). To build robust classifier, many recent efforts are devoted to synthesize novel images/features through various augmentation strategies, such as image transformations [65] and style transformations [82, 46, 76]. They typically manipulate the low-level feature statistics (*e.g.,* color and texture) to generate diverse image styles while maintaining the content unchanged. However, we argue that these operations may still be constrained by the spurious correlation as style-agnostic representation is not sufficient to ensure semantic invariance [42].

Inspired by this, we introduce a novel data mixing strategy, which incorporates the categorical information to directly recombine source images. Considering that the category and domain labels of each sample are two orthogonal yet complementary supervision signals, we directly obtain the activation maps generated by these two classification networks for depicting the foreground regions.

An overview of the CDM is illustrated in Fig. 3. Specifically, for an input sample x_i, we first leverage the vanilla Grad-CAM [55] to generate activation maps $\mathcal{M}_c$ (class) and $\mathcal{M}_d$ (domain) by using its label y_i and d_i, respectively. As shown in Fig. 3, the activation map ($\mathcal{M}_d$) induced by domain classification does not focus on the background as our expectation (this phenomenon holds in most DG benchmarks). Thus, we merge $\mathcal{M}_c$ and $\mathcal{M}_d$ to extract the foreground objects as follows,

$$\mathcal{M}_f := \texttt{threshold}(\mathcal{M}_c + \mathcal{M}_d), \tag{1}$$

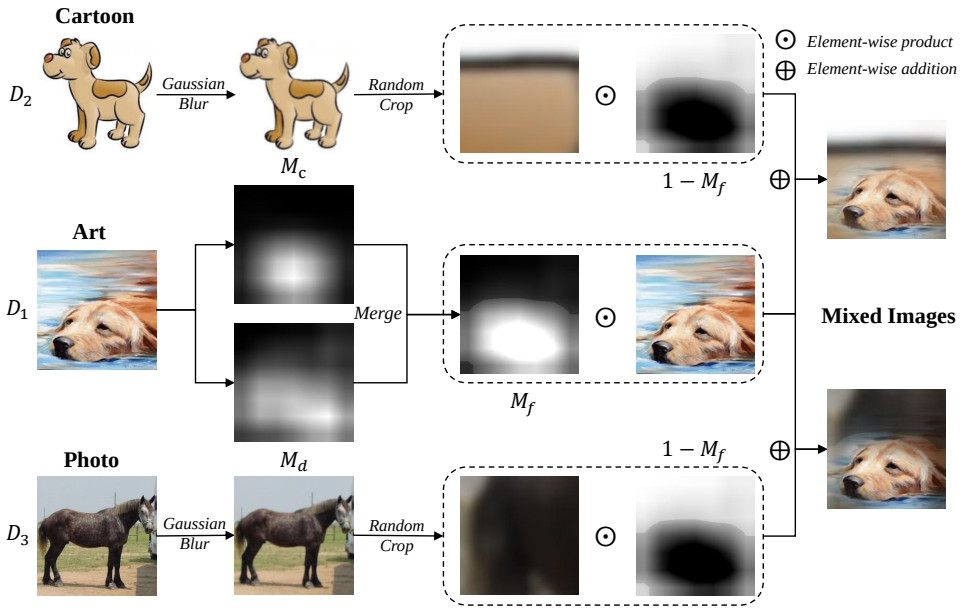


Figure 3: The workflow of CDM. For an image from Domain "Art", **(i)** extract two types of its foreground CAM ($\mathcal{M}_c$ and $\mathcal{M}_d$) and merge them ($\mathcal{M}_f$); **(ii)** randomly select two images from Domain "Cartoon" and Domain "Photo" to construct two backgrounds; **(iii)** apply Gaussian blur to the selected image and then randomly crop a patch from the blurred image; **(iv)** mix the foreground and background images by virtue of the computed $\mathcal{M}_f$.

where $\texttt{threshold}(\cdot)$ aims to reduce the outlier points and is set to 0.2 in all experiments to avoid additional tuning. Intuitively, $\mathcal{M}_d$ is a coarse-level supervision, which is sensitive to the domain-wise variations, while $\mathcal{M}_c$ is a fine-level supervision, which is sensitive to the intra-domain variations.

Next, to simulate the real-world situation that an object may appear in various scenes with distinct backgrounds, we randomly crop a small region of an image sampled from other domains. Before cropping, a Gaussian blur is applied to image x_j to reduce image noise and detail, *i.e.,* $x'_j = \texttt{Gaussian}(x_j)$. A small patch is cropped from x'_j and resized to the same scale, *i.e.,* $x_j^{\texttt{crop}} = \texttt{resize}(\texttt{crop}(x'_j))$. Finally, we synthesize a new image x'_i by substituting the background of x_i with the background of x'_j using the merged Grad-CAM of image x_i as a balancing weight, *i.e.,*

$$x_i^{\texttt{mix}} := \mathcal{M}_f \odot x_i + (1 - \mathcal{M}_f) \odot x_j^{\texttt{crop}}, \tag{2}$$

where $\odot$ is the element-wise product. The above operation is applied to every source image, and we will obtain $(N-1) \cdot |\mathcal{D}_s|$ images in all, where $|\mathcal{D}_s|$ is the number of original source images.

3.2 Adaptive Semantic Topology Refinement

Structural Invariance. After the data mixing, we introduce an Adaptive Semantic Topology Refinement (ASTR) mechanism to adaptively and progressively refine the topological structure of semantic anchors in the feature space, explicitly constraining the structural invariance across domains. The key idea is to endow the semantic anchor with the capability of reasoning over semantic topology globally, where semantic anchor is represented by the prototype features of each class. We argue that semantic topology, which consists of the semantic anchors as well as their interactions, is robust to unseen test environments. Technically, the structural invariance is characterised from two aspects: *cross-domain variation* and *cross-model variation* (historical and present models), implying that the semantic topology should be independent of the spatial and temporal changes.

Overview. We introduce an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{C})$ to represent the semantic topology, where $|\mathcal{V}| = K$ is the set of K nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. The feature matrix is denoted by $\boldsymbol{C} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\} \in \mathbb{R}^{K \times D}$, where $\boldsymbol{c}_k$ denotes a prototype corresponding to the k-th class, D is the feature dimension, and each element in $\boldsymbol{C}$ is associated to a node $\boldsymbol{v} \in \mathcal{V}$. The adjacency

matrix is denoted by $\boldsymbol{A} \in \mathbb{R}^{K\times K}$ where each element $\boldsymbol{A}_{ij}$ denotes the weight of edge $\boldsymbol{e}_{ij} \in \mathcal{V}$. During training, each input sample is first compared to a set of semantic anchors and then adaptively aggregates features from its adjacent anchors to update its features. All updated sample features will be further used to update the corresponding anchor feature. Finally, the structural invariance is optimized based on bipartite graph learning.

Step 0: ASTR first computes the initial global anchor features $\boldsymbol{C}_{(0)}^{\texttt{global}}$ based on the vanilla DeepAll model, which is trained on both original and generated data. For the relation graph, we define its corresponding adjacency matrix $\mathcal{G}^{\texttt{global}}$ as follows,

$$\boldsymbol{A}_{ij}^{(I)} = \exp(-\frac{\|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2^2}{2\sigma^2}), \tag{3}$$

where I indicates the iteration times and σ^2 is the kernel bandwidth.

Step 1: At each training iteration, a batch of samples $\mathcal{B}$ is projected into the embedding space via the feature extractor G, *i.e.,* $\boldsymbol{f}_i = G(x_i), x_i \in \mathcal{B}$. For the sake of simplicity, we omit the subscript i. In the I-th iteration, for an image x that belongs to class k, we perform feature aggregation as follows,

$$\boldsymbol{f}' = \phi\boldsymbol{f} + (1-\phi)\sum_{j=1}^{K} w_j \cdot \boldsymbol{c}_j^{\texttt{global}}, \quad w_j = \frac{\exp((\boldsymbol{A}_{kj}^{(I-1)} + \cos(\boldsymbol{f}, \boldsymbol{c}_j))/2)}{\sum_j \exp((\boldsymbol{A}_{kj}^{(I-1)} + \cos(\boldsymbol{f}, \boldsymbol{c}_j))/2)}, \tag{4}$$

where $j = \{1, 2, ..., K\}$, $\cos(\cdot,\cdot)$ denotes cosine distance, and $\phi \in (0, 1)$ is a positive constant. $\cos(\boldsymbol{f}, \boldsymbol{c}_j)$ *locally* measures the relations between each instance and all semantic anchors, while $\boldsymbol{A}_{kj}^{(I-1)}$ provides the *global relations* among different semantic anchors. By doing so, each instance will be aware of the holistic semantic topology and dynamically conduct aggregation.

Step 2: Then, we update the global anchor features based on $\mathcal{B}$ in a moving average manner,

$$c_{j(I)}^{\texttt{global}} = \phi\, c_{j(I)}^{\texttt{local}} + (1-\phi)\, c_{j(I-1)}^{\texttt{global}}, \tag{5}$$

where $c_{j(I)}^{\texttt{local}}$ stands for the local anchor features computed from all $\boldsymbol{f}'$. Meanwhile, the adjacency matrix will be updated accordingly.

Step 3: Finally, we optimize the cross-domain and cross-model consistencies to achieve structural invariance. A naive approach is to compare each anchor in one domain/model to all anchors in the other domain/model. However, the structural relations of each anchor within the graph are neglected. On the other hand, the constructed graph may still contain some spurious edges, leading to unreliable inter-graph node correspondence reasoning. Remedying these issues, we introduce a Bipartite Graph Convolutional Network (BGCN) to perform cross-domain relational reasoning based on a bipartite graph $\mathcal{G}^{\texttt{bg}}(\mathcal{V}_{s'}, \mathcal{V}_{s''}, \mathcal{E}, \boldsymbol{A}_{ij}^{\texttt{bg}})$, where s' and s'' can be any two source domains. The node affinities between $\mathcal{V}_{s'}$ and $\mathcal{V}_{s''}$ are simultaneously constrained by cross-domain and cross-model variations. Specifically, for a node v_i in $\mathcal{V}_{s'}$ and a node v_j in $\mathcal{V}_{s''}$, we have,

$$\boldsymbol{A}_{ij}^{\texttt{bg}} = \frac{\cos(\boldsymbol{c}_{i(I-1)}^{\texttt{global}}, \boldsymbol{c}_{i(I)}^{\texttt{global}}) + \cos(\boldsymbol{c}_{j(I-1)}^{\texttt{global}}, \boldsymbol{c}_{j(I)}^{\texttt{global}})}{2} \cdot \cos(\boldsymbol{c}_{i(I)}^{\texttt{global}}, \boldsymbol{c}_{j(I)}^{\texttt{global}}), \tag{6}$$

By doing so, the influence of those unreliable intra- and inter-graph correspondence were largely reduced. Then, we conduct bipartite graph convolution [26, 11] over $\mathcal{G}^{\texttt{bg}}$ by stacking multiple layers. Meanwhile, we introduce a contrastive consistency regularization on the top of BGCN,

$$\mathcal{L}_{\texttt{CCR}} = \sum_{k=1}^{K} \Phi(\boldsymbol{z}_{\mathcal{V}_{s'}}^{k}, \boldsymbol{z}_{\mathcal{V}_{s''}}^{k}) + \sum_{m\neq n} (\max\{0, \xi - \Phi(\boldsymbol{z}_{\mathcal{V}_{s'}}^{m}, \boldsymbol{z}_{\mathcal{V}_{s''}}^{n})\}), \tag{7}$$

where $\boldsymbol{z}$ is the embedded vector through BGCN, Φ is the squared Euclidean distance, and ξ is the margin term and is set to 2 in experiments. To this end, the overall training objective of `MiRe` is,

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}}\ \mathcal{L}_{\text{cls}} + \lambda \cdot \mathcal{L}_{\texttt{CCR}}, \tag{8}$$

where $\mathcal{L}_{\text{cls}}$ is the standard cross-entropy loss and λ is the trade-off parameter.

Table 1: Generalization results (%) on PACS benchmark.

Source	Target	DeepAll -	JiGen [6]	CrossGrad [57]	MSAF [15]	ER [78]	Metareg [4]	RSC [23]	MixStyle [82]	SagNet [46]	EFDM [76]	MiRe (Ours)
C,P,S	Art	77.0±0.3	79.4	79.8	80.3	80.7	83.7	78.9	83.1	83.6	83.9	84.6±0.5
A,P,S	Cartoon	74.8±0.5	75.3	76.8	77.2	76.4	77.2	76.9	78.6	77.7	79.4	79.5±0.4
A,C,S	Photo	95.8±0.1	96.0	96.0	95.0	96.7	95.5	94.1	95.9	95.5	96.8	96.8±0.2
A,C,P	Sketch	70.0±0.6	71.6	70.2	71.7	71.8	70.3	76.8	74.2	76.3	75.0	78.4±1.0
Average		79.4	80.5	80.7	81.1	81.4	81.7	81.7	82.9	83.3	83.9	**84.8**

Table 2: Generalization results (%) on VLCS benchmark.

Source	Target	DeepAll -	DBA-DG [30]	MMD-AAE [34]	MLDG [31]	Epi-FCR [32]	JiGen [6]	MASF [15]	MMLD [43]	SFA-A [33]	MiRe (Ours)
L,C,S	VOC	71.8±0.6	70.0	67.7	67.7	67.1	70.6	69.1	72.0	70.4	70.3±0.3
V,C,S	LabelMe	61.1±0.4	63.5	62.6	61.3	64.3	60.9	64.9	58.8	62.0	63.6±0.7
V,L,S	Caltech	95.8±0.2	93.6	94.4	94.4	94.1	96.9	94.8	96.7	97.2	96.2±0.4
V,L,C	Sun	62.5±0.6	61.3	64.4	64.4	65.9	64.3	67.6	68.1	66.2	69.4±0.7
Average		72.8	72.1	72.3	72.3	72.9	73.2	74.1	73.9	74.0	**74.9**

Table 3: Generalization results (%) on Office-Home benchmark.

Source	Target	DeepAll -	MMD-AAE [34]	CrossGrad [57]	JiGen [6]	DDAIG [80]	DOSN [56]	RSC [23]	SagNet [46]	MixStyle [82]	MiRe (Ours)
C,P,R	Art	59.4±0.4	59.9	58.4	53.0	59.2	59.4	58.4	60.2	58.7	60.2±0.8
A,P,R	Clipart	48.0±1.1	47.3	49.4	47.5	52.3	45.7	47.9	45.4	53.4	53.2±0.9
A,C,R	Product	72.7±0.5	72.1	73.9	71.5	74.6	71.8	71.6	70.4	74.2	75.1±0.6
A,C,P	Real	75.3±0.4	74.8	75.8	72.8	76.0	74.7	74.5	73.4	75.9	76.4±0.6
Average		63.9	62.7	64.4	61.2	65.5	62.9	63.1	62.4	65.5	**66.2**

4 Experiments

4.1 Setup

Datasets. We fully verify the generalization performance of `MiRe` on four standard DG benchmarks: **PACS** [30], **VLCS** [17], **Office-Home** [63], and **DomainNet** [50]. (1) PACS is the most-widely used DG benchmark, which has 9,991 images of 7 classes from four kinds of environment: *Photo*, *Art Painting*, *Cartoon*, and *Sketch*. (2) VLCS contains 10,729 images of 5 classes from four photographic domains: *PASCAL VOC 2007* [16], *LabelMe* [53], *Caltech* [18], and *Sun* [70]. (3) Office-Home is composed of 15,500 images of 65 classes from four domains (*Artistic*, *Clipart*, *Product*, and *Real World*). The images are collected from office and home environments. (4) DomainNet is a more recent large-scale dataset for multi-source domain adaptation and DG. It spans 0.6 million images of 345 classes from six domains (*Clipart*, *Infograph*, *Quickdraw*, *Painting*, *Real* and *Sketch*).

Model Selection. To facilitate fair comparisons, we strictly follow the model selection strategy used by baseline methods. In general, we follow the leave-one-domain-out setting, *i.e.,* one domain is chosen as the held out domain and the remaining domains are seen as source domains. For PACS, we follow the original train and val splits established by [30]. On VLCS, according to the prior methods [6, 43], we split 30% of the source samples as validation datasets. For Office-Home, we employ the train and val splits established in [30, 81, 82]. On DomainNet, we follow the train and val splits used in [7], *i.e.,* randomly divide the train split from [50] in a 90-10% proportion to obtain train and val splits. Then, we select the model that shows the best performance in the validation dataset.

Implementation Details. For PACS, Office-Home, and DomainNet, we use ResNet-18 [22] pre-trained on the ImageNet as the backbone feature extractor. To include more baselines for comparison, we use AlexNet [28] for VLCS. On DomainNet, we also provide results of ResNet-50. The networks are trained using SGD with momentum of 0.9 and weight decay of 5e-4 for 100 epochs. For PACS, VLCS, and Office-Home, the batch size is set to 16. For DomainNet, the batch size is set to 256. For CDM, the cropped area ratio (background image) is empirically set to $\frac{1}{8}$. For Eq. (4), we set $\phi = 0.5$. For Eq. (8), we set $\lambda = 0.1$. We evaluate the out-of-domain generalization performance in all object

Table 4: Generalization results (%) on DomainNet benchmark.

	Method	Clipart	Infograph	Quickdraw	Painting	Real	Sketch	Avg
ResNet-18	DeepAll	57.2	17.7	43.2	13.9	54.9	39.4	37.7
	Multi-Headed [7]	55.4	17.5	40.8	11.2	52.9	38.6	36.1
	MetaReg [4]	53.6	21.0	45.2	10.6	58.4	42.3	38.5
	DMG [7]	60.0	18.7	44.5	14.1	54.7	41.7	39.0
	`MiRe`	62.5	23.0	46.5	15.7	59.1	44.4	**41.9**
ResNet-50	DeepAll	61.8±0.6	20.2±0.7	45.0±1.3	14.3±1.0	57.3±0.7	44.8±0.6	40.6
	Multi-Headed [7]	61.7	21.2	46.8	13.8	58.4	45.4	41.2
	MetaReg [4]	59.7	25.5	50.1	11.5	64.5	50.0	43.6
	DMG [7]	65.2	22.1	50.0	15.6	59.6	49.0	43.6
	`MiRe`	64.7±1.0	27.8±1.2	53.1±0.8	16.4±1.5	64.1±0.6	52.3±0.9	**46.4**

classes and report the means and standard derivations over 10 runs with different random seeds. We implement all experiments based on the PyTorch framework.

4.2 Baselines

We compare `MiRe` against state-of-the-art DG methods, which are summarized as follows. **DeepAll:** Training a classification model through simply merging all source domains. **Feature Alignment:** MMD-AAE [34] and EFDM [76]. **Feature Disentanglement:** DOSN [56], DMG [7] and SagNet [46]. **Data Augmentation:** JiGen [6] and DDAIG [80]. **Feature Augmentation:** CrossGrad [57], MixStyle [82] and SFA-A [33]. **Meta-Learning:** MASF [15] and Metareg [4]. **Other SOTA:** RSC [23] designs a model-agnostic training strategy, and ER [78] ensures the conditional invariance via an entropy regularization term. For most of the compared approaches, we cite the classification accuracy reported in their original papers. In PACS, the results of 'RSC' and 'MixStyle' are obtained with their official codes by following the leave-one-domain-out setting.

4.3 Results

PACS and VLCS. In Tab. 1 and Tab. 2, we display the quantitative comparisons with baseline methods on PACS and VLCS benchmarks. As can be seen, `MiRe` outperforms state-of-the-art DG methods by large margins, revealing the effectiveness of the proposed structural invariance in real-world data. In particular, `MiRe` improves over the best baseline method [76] (CVPR 2022) by 0.9% on the PACS dataset. For the hard generalization task A, C, S $\rightarrow$ Sketch, our method substantially outperforms the suboptimal result by 1.6%.

Office-Home and DomainNet. Tab. 3 and Tab. 4 report the results on Office-Home and DomainNet benchmarks, which have larger data volume in contrast to PACS and VLCS. The proposed `MiRe` substantially and consistently exceeds all comparison methods, showing the scalability of our `MiRe` on large-scale datasets with distinct classes.

4.4 Empirical Analyses and Ablations

Ablation Study. We investigate the role of each part of `MiRe` by conducting in-depth ablation studies on four benchmarks. The results are listed in Tab. 5. We highlight two sets of key observations. (1) CDM w/o $(1 - \mathcal{M}_d)$ represents that we replace $\mathcal{M}_d$ by $(1 - \mathcal{M}_d)$ in Eq. (1). The results testify that the domain classification evidence is beneficial for depicting the foregrounds. In addition, $\mathcal{M}_c$, $\mathcal{M}_d$, and Gaussian blur operation are reasonably designed, providing high-quality augmentation results. Only with CDM could provides very competitive results. (2) Introducing relation graphs to model semantic topology plays an important role in encouraging the so-called structural invariance especially on the challenging benchmark, such as DomainNet.

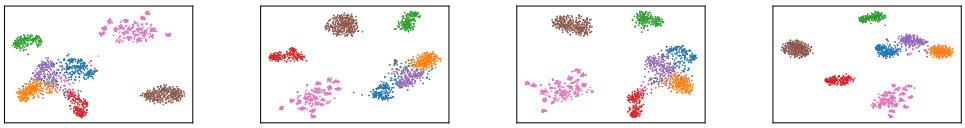
The Effects of CDM. Since CDM does not requires any specific model architectures and training procedures, we apply it to three existing methods, *i.e.,* ER [78], MixStyle [82] and EFDM [76], to further enhance their classification performance. The results are shown in Tab. 6, where CDM consistently improves over the baseline methods.

Table 5: Ablation of MiRe on four DG benchmarks (%).

Method	PACS	VLCS	Office-Home	DomainNet	Average
DeepAll	79.4	72.8	63.9	37.7	63.5
MiRe w/o CDM	83.5	74.0	65.1	39.3	65.5
CDM w/o $\mathcal{M}_c$	82.6	72.8	65.0	39.4	65.0
CDM w/o $\mathcal{M}_d$	84.4	74.4	65.7	41.0	66.4
CDM w/o Gaussian Blur	83.9	74.2	64.6	40.6	65.8
CDM w/ $(1 - \mathcal{M}_d)$	83.0	73.4	64.9	38.8	65.0
MiRe w/o ASTR	82.3	71.8	64.7	38.9	64.4
ASTR w/o Cross-Domain Invariance	84.0	74.5	65.6	40.6	66.2
ASTR w/o Cross-Model Invariance	84.3	74.6	65.7	41.0	66.4
ASTR w/o Graph Structure	82.8	73.3	64.8	40.1	65.3
ASTR w/o Feature Aggregation	84.2	74.1	65.7	40.4	66.1
MiRe	84.8	74.9	66.2	41.9	67.0

Table 6: The effects of CDM on PACS benchmark.

Source	Target	DeepAll	CDM	+ ER (81.4)	+ MixStyle (82.9)	+ EFDM (83.9)
C,P,S	Art	77.0	80.8	80.4	82.9	83.8
A,P,S	Cartoon	74.8	78.5	77.8	78.2	81.3
A,C,S	Photo	95.8	94.8	94.4	96.0	95.5
A,C,P	Sketch	70.0	75.0	77.4	77.6	78.2
Average		79.4	82.3	82.5 (+1.1)	83.7 (+0.8)	84.7 (+0.8)



(a) DeepAll (b) MixStyle (c) EFDM (d) MiRe (Ours)

Figure 4: The t-SNE visualization of extracted deep features using different methods on PACS dataset. The different colors stands for different classes.

The Effects of ASTR. To testify the robustness of ASTR under the presence of spurious correlation, we additionally perform experiments on three Chest X-ray datasets: NIH [68], ChexPert [25] and RSNA [1]. We follow the same settings as in [42] to create spurious correlation. The empirical comparisons are shown in Tab. 7, where the results of baseline methods (*i.e.,* ERM [62], IRM [3], CSD [51] and MatchDG [42]) are cited from [42].

Table 7: Results (%) on Chest X-Rays data.

Method	RSNA	ChexPert	NIH
ERM	55.1	60.9	53.4
IRM	57.0	63.3	54.6
CSD	58.6	64.4	54.7
MatchDG	58.2	59.0	53.2
ASTR	**63.6**	**65.0**	**56.4**

Feature Visualization. We use the t-SNE algorithm [61] to visualize the deep representations in ResNet-18 when *Photo* (PACS benchmark) is regarded as target domain. We compare the proposed MiRe with DeepAll, MixStyle [82], and EFDM [76] in Fig. 4. As expected, MiRe exhibits better clustering (class) patterns due to the refinement of semantic topology during training.

5 Conclusion

In this paper, we proposed the MiRe framework to achieve out-of-distribution generalization. The key idea of our method is to endow the predictive models with the capability of reasoning over the semantic topology. MiRe instantiates this objective with the incorporation of two elaborate modules, *i.e.,* CDM and ASTR. CDM augments the original data to reduce potential bias towards spurious correlations, while ASTR builds relation graphs to represent semantic topology, which is updated through feature aggregation and message-passing. Experiments on standard benchmarks verified the effectiveness of MiRe. Additional empirical analysis reveals the individual effect of each component.

Acknowledgements

This work was supported by Hong Kong Research Grants Council through Research Impact Fund (Grant R-5001-18).

References

[1] Kaggle: Rsna pneumonia detection challenge. 2018.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *ICML*, pages 145–155, 2020.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, pages 998–1008, 2018.

[5] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[6] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019.

[7] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, pages 301–318, 2020.

[8] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *CVPR*, pages 7119–7129, 2022.

[9] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *ICCV*, pages 2703–2712, 2021.

[10] Chaoqi Chen, Jiongcheng Li, Hong-Yu Zhou, Xiaoguang Han, Yue Huang, Xinghao Ding, and Yizhou Yu. Relation matters: Foreground-aware graph-based relational reasoning for domain adaptive object detection. *TPAMI*, 2022.

[11] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *arXiv preprint arXiv:2209.13232*, 2022.

[12] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, pages 627–636, 2019.

[13] Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

[14] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE TPAMI*, 2021.

[15] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, pages 6447–6458, 2019.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[17] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, pages 1657–1664, 2013.

[18] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR-W*, pages 178–178, 2004.

[19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

[20] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016.

[21] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[23] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140, 2020.

[24] Wilhelm Humboldt. On language: On the diversity of human language construction and its influence on the mental development of the human species. 1999.

[25] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, pages 590–597, 2019.

[26] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[27] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, pages 5637–5664, 2021.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[29] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, 2021.

[30] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[32] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, pages 1446–1455, 2019.

[33] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *ICCV*, pages 8886–8895, 2021.

[34] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. Domain invariant and class discriminative feature learning for visual domain adaptation. *IEEE Transactions on Image Processing*, 27(9):4260–4273, 2018.

[35] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. In *ICLR*, 2022.

[36] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018.

[37] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pages 3915–3924, 2019.

[38] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *NeurIPS*, 34, 2021.

[39] Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *MICCAI*, pages 475–485. Springer, 2020.

[40] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, pages 4114–4124, 2019.

[41] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[42] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *ICML*, pages 7313–7324, 2021.

[43] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020.

[44] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. Disentanglement testing sprites dataset. In *https://github.com/deepmind/dsprites-dataset/*, 2017.

[45] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5715–5725, 2017.

[46] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, pages 8690–8699, 2021.

[47] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[48] Judea Pearl. *Causality*. Cambridge university press, 2009.

[49] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *CVPR*, 2022.

[50] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pages 1406–1415, 2019.

[51] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, pages 7728–7738, 2020.

[52] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *JMLR*, 19(1):1309–1342, 2018.

[53] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[54] Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world. In *ECCV*, 2022.

[55] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[56] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, pages 68–83, 2020.

[57] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *ICLR*, 2018.

[58] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016.

[59] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. *NeurIPS*, 34, 2021.

[60] Takeshi Teshima and Masashi Sugiyama. Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation. In *UAI*, pages 86–96, 2021.

[61] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[62] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[63] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017.

[64] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, pages 6438–6447, 2019.

[65] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *ICCV*, pages 7980–7989, 2019.

[66] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5334–5344, 2018.

[67] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *TKDE*, 2022.

[68] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017.

[69] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Taylan Cemgil, et al. A fine-grained analysis on distribution shift. 2021.

[70] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[71] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *ICML*, pages 5419–5428, 2018.

[72] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021.

[73] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *ICLR*, 2021.

[74] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019.

[75] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[76] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, 2022.

[77] Gangming Zhao, Quanlong Feng, Chaoqi Chen, Zhen Zhou, and Yizhou Yu. Diagnose like a radiologist: Hybrid neuro-probabilistic reasoning for attribute-based medical image diagnosis. *TPAMI*, 2021.

[78] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *NeurIPS*, 33:16096–16107, 2020.

[79] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *TPAMI*, 2022.

[80] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020.

[81] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *ECCV*, pages 561–578, 2020.

[82] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]