Mixstyle-Entropy: Whole Process Domain Generalization with Causal Intervention and Perturbation

Luyao Tang*[1]
http://lytang63.github.io
Yuxuan Yuan*[1]
yuanyuxuan0908@stu.xmu.edu.cn
Chaoqi Chen[2]
cqchen1994@gmail.com
Xinghao Ding[1]
dxh@xmu.edu.cn
Yue Huang†[1]
yhuang2010@xmu.edu.cn

[1] School of Informatics
Xiamen University
Xiamen, China

[2] Department of Computer Science
The University of Hong Kong
Hong Kong, China

Abstract

Despite the considerable advancements achieved by deep neural networks, their performance tends to degenerate when the test environment diverges from the training ones. Domain generalization (DG) solves this issue by learning representations independent of domain-related information, thus facilitating extrapolation to unseen environments. Existing approaches typically focus on formulating tailored training objectives to extract shared features from the source data. However, the disjointed training and testing procedures may compromise robustness, particularly in the face of unforeseen variations during deployment. In this paper, we propose a novel and holistic framework based on causality, named `Mixstyle-Entropy`, designed to enhance model generalization by incorporating causal intervention during training and causal perturbation during testing. Specifically, during the training phase, we employ entropy-based causal intervention (EnIn) to refine the selection of causal variables. To identify samples with anti-interference causal variables from the target domain, we propose a novel metric, homeostatic score, through causal perturbation (HoPer) to construct a prototype classifier in test time. Experimental results across multiple cross-domain tasks confirm the efficacy of `Mixstyle-Entropy`. Code is available at github.com/lytang63/MixstyleEntropy.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in various computer vision applications, including image classification and object detection tasks. However, Nonethe-

*First two authors contributed equally.
†Corresponding author.
© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

less, their training relies on the assumption that the training and testing datasets are identically and independently distributed (IID) [2]. In real-world scenarios, this assumption is often violated, leading to a marked degradation in the performance of models trained on the source domain [18, 29, 51].

Domain generalization (DG) aims to enhance the generalization capacity of deep neural networks (DNNs) towards out-of-distribution (OOD) data. The challenge of OOD scenarios [23] has been tackled from diverse perspectives encompassing optimization, model architecture, and data manipulation. Optimization-wise, diverse training strategies have sought to cultivate domain-independent feature representations. These strategies include explicit feature alignment [4, 14, 21, 41] and adversarial learning [13, 16]. Concerning model design, enhancing generalization often involves meticulous crafting of DNN structures [34, 46, 55] or employing ensembles of multiple expert models [39, 74]. On the data front, techniques like data augmentation [43, 60, 68] and generation are leveraged to enrich the diversity [5, 56, 73] of training samples. Additionally, approaches based on causal learning [6, 9, 36] and meta-learning [4, 67] have also been explored.

However, mainstream DG strategies often overlook two pivotal limitations. Firstly, while many methods prioritize domain invariance [12, 26, 30, 70], they often disregard spurious correlations within features. For example, enforcing domain invariance can introduce spurious correlations. Mixstyle [75] randomly selects instances and applies linear interpolation to their statistical features, overlooking spurious correlations between semantic content and image style. Secondly, expanding the feature space blindly can lead to intricate decision boundaries. DSU [31] utilizes multivariate Gaussian distributions to construct virtual instances. However, such an approach can compromise existing class embeddings, resulting in reduced inter-class distances and rugged decision boundaries, as shown in Figure 3.

From the perspective of causal learning, the aforementioned approaches share a common objective of extracting domain-invariant variables. A more insightful approach would be: the process of extracting domain-invariant features should involve modeling data generation, extracting causal relationships from observable variables [38], and isolating causal variables used for classification.

Based on the discussion above, we aim to differentiate genuine domain-related variables and intervene in the forward process. To this end, we propose `Mixstyle-Entropy`, a straightforward yet efficacious DG approach. `Mixstyle-Entropy` consists of two key components: Intervention and Perturbation. During training, we extract domain-related statistical features using theoretically derived feature entropy. We perform causal interventions on samples in the embedding space to sever the association between domain-related information and causal variables. This disentanglement enables the isolation of causal variables containing semantic information crucial for precise prediction. Moreover, we integrate causal learning into the model deployment phase through Causal Perturbation, which adjusts embeddings towards the centroid of their respective classes. Additionally, we propose a novel Homeostatic Score, to assess the interference resilience of causal variable branches in a structured causal graph. Finally, we construct a prototype classifier suitable for the testing domain using samples with stable causal relationships and progressively fit the target distribution.

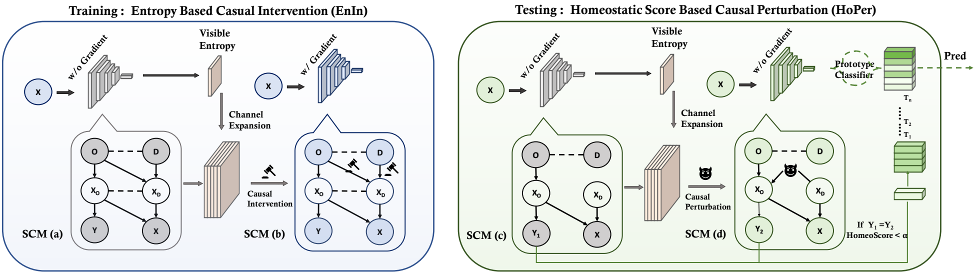


Figure 1: The pipeline of the proposed `Mixstyle-Entropy`. **Training:** Extract visible entropy and perform causal interventions on samples to sever the spurious connections between class and domain. **Testing:** Causal perturbation is applied to the testing samples, and through the HomeoScore, a generalized classifier is built to adapt to the target domain.

2 Related Work

2.1 Causality and Domain Generalization

Causality is often used to establish connections between causal relationships and model generalization [10, 50]. Various techniques such as invariant causal mechanisms [17, 58, 66] and restoring causal features [33, 53] have been proposed to boost OOD generalization. Earlier work [15] ventured causal reasoning into domain adaptation, and others aimed to establish causal links between class labels and samples [17, 27, 37, 49]. MatchDG [38] suggested a domain generalization invariance condition via causal flow between labels. However, many causality and domain generalization solutions rely heavily on restrictive assumptions about causal graphs or structural equations, such as CIRL [36] employs dimensional representations to mimic causal factors, depending on only a highly generic causal structure model without restrictive assumptions.

2.2 Online Parameter Updating

Online parameter updating, a technique typically improving model performance in unfamiliar testing domains, involves testing distribution alignment operations. The chief approaches include Test-time training (TTT) [35, 59] and Test-time adaptation (TTA) [3, 20, 44, 64]. TTT enhances the model through self-supervised testing data tasks like rotation classification [24]. TTA adjusts the model during testing without altering training, with Tent [64] minimizing entropy for parameter updates, and SHOT [32] maximizing mutual information. Some studies [7, 44, 65] explored TTA and continual learning amalgamation.

3 Methodology

3.1 Theoretical Insight

In this section, we explain the theoretical foundations of our work. Similar to the causal relationship constructed by MatchDG [38], we present a Structural Causal Model (SCM) [54] for DG task in Figure 1 SCM (a). For intuition, let's consider an example inspired by

Mixstyle [75]. Each domain (D) displays a unique 'style' for images from different domains while the various objects (O) exhibit a semantic characteristic 'shape'. X_O are the high-level causal features of O which induce the label Y. Note that another high-level proxy feature X_D in SCM is determined by the style of the given image. At the same time, this style information comes from both D and O. Due to the common parent node O, X_O and X_D are correlated, and D will inevitably influence the prediction of X_O. However, we expect that the class Y only relies on the casual feature X_O of the object category. The SCM (a) can be expressed as:

$$\begin{aligned} \text{SCM (a)}: \quad & X_O := f_{X_O}\left(O, U_{X_O}\right) \qquad X_D := f_{X_D}\left(O, D, U_{X_D}\right) \\ & X := f_X\left(X_O, X_D, U_X\right) \quad\; Y := f_Y\left(X_O, U_Y\right). \end{aligned} \tag{1}$$

The variables $\{X_O, X_D, X, Y\}$ are the endogenous variables [25], and they have parent nodes in causal graphs. Additionally, the implicit noise variable U, which includes all relevant background conditions for the deterministic variable, is omitted from the causal graphs. Note that the structural equation is a generalized expression of the variable conditional probability distribution, the truncated factorization [47] and backdoor adjustment [47] techniques remain applicable.

In the task of DG, a model is trained on the source training set $\mathcal{D}_s = \{\mathcal{D}_1, \ldots, \mathcal{D}_S\} (S \geq 2)$, and evaluated on the unseen target set The n-th source domain is denoted as $\mathcal{D}_n = \{(\mathbf{x}_i, y_i, d_i)\}_{i=1}^{N_n}$, where $\mathbf{x}_i \in \mathbb{R}^D$ represents the training data, $y_i = \{1, 2, \ldots, K\}$ is the class label, $d_i = \{1, 2, \ldots, S\}$ denotes the domain label. Source and target domains follow different data distributions in the joint space $\mathcal{X} \times \mathcal{Y}$. Therefore, the optimization function (ERM) [62] can be represented as follows:

$$\arg\min \mathbb{E}_{(\mathbf{x}_i, y_i, d_i) \in \mathcal{D}_s} [\ell(f(\mathbf{x}_i), y_i)], \tag{2}$$

where $\ell(\cdot, \cdot)$ quantifies the inconsistency between between predictions and labels.

To enhance the performance of domain-agnostic feature representations $g(\mathrm{x})$ we employ an entropy-based formulation as done by distribution-matching methods [1]. We follow the methodology in MatchDG [38] and aim to learn a representation that is independent of domain given class label. It can be interpreted as maximizing the entropy of domain given class label, i.e, $g^*(x) = \arg\max H(d|y, g(\mathrm{x}))$, the optimal $g^*(x)$ satisfies $H(d|y, g^*(x)) = H(d|y)$.

Unlike MatchDG follows the assumption that $x_o \perp\!\!\!\perp d|y$, we consider a more realistic scenario whereby $x_o \not\perp\!\!\!\perp d|y$, i.e., $P(x_o|y)$ varies across the domains. A simple example is that although sharing the same label, the dogs depicted in the sketch exhibit significant differences in their morphological characteristics compared to those in the painting. In this case, considering the correlation between X_O and X_D in SCM (a), we obtain the following:

$$H(g(\mathrm{x}), d|y) - H(d|y) \leqslant H(x_o, x_d, d|y) - H(d|y). \tag{3}$$

If $x_o \perp\!\!\!\perp x_d$, the above equation can be simplified to:

$$H(g(\mathrm{x}), d|y) - H(d|y) = H(x_o, x_d, d|y) - H(d|y) = H(x_o, d|y) - H(d|y) = KL(x_o, d|y). \tag{4}$$

Turning our attention to node X_D under the aforementioned DG setup, x_d changes across domains, i.e., $x_d \not\perp\!\!\!\perp y|d$. Similar to the above analysis, we can also perform a causal analysis on x_d and obtain the formula for entropy. For simplicity, we denote the entropy feature representation as $H(g(\mathrm{x}))$. The relationship between $H(g(\mathrm{x}))$ and variables x_o and x_d can be expressed by the following formula:

$$H(g(\mathrm{x})) \propto KL(x_o, d|y) \quad H(g(\mathrm{x})) \propto KL(x_d, y|d). \tag{5}$$

In general, our goal is to minimize prediction error using the most 'pure' x_o. We achieve this by performing a causal intervention, using the do-operation $do(\cdot)$ [47] on x_d. We aim to remove domain-relevant information from the feature representation similar to CIRL [36] and Mixstyle [75]. Ideally, we can eliminate the arrow pointing to X_D, clip the causal relationship between O and X_D, and remove the association between X_O and X_D in SCM (a). The d-separation and perfect map assumption [48] are employed to obtain the category-related semantic factors for prediction. Two critical aspects need to be considered, i.e., the specific form of $do(\cdot)$ and the assumption of x_d. In CIRL, the intervention is performed on x_d by mixing the amplitude information from other domains after Fourier transformation. However, due to the spurious correlation between O and D, the x'_o of the intervened sample will be introduced into the x''_o of nother samples. Therefore, careful selection of causal intervention variables is crucial for improving the model's generalization performance. Based on the theoretical analysis above, we detail our approach in Section 3.2 and 3.3.

3.2 Training: Entropy Based Casual Intervention

Taking CNN as an example, the model $f(\mathrm{x}) = \mathbf{W}^\top g(\mathrm{x})$ can be decomposed into a feature extractor $g(\cdot)$ and a classifier. $g(\cdot)$ consists of multiple blocks with a stacking depth of m, namely, $g(\cdot) = [g_1(\cdot), g_2(\cdot), \ldots, g_m(\cdot)]$. As the input passes through every block, the dimension of the feature space is expanded progressively until the local feature vector $\mathbf{v_i}$ is obtained for classification. Here, $\mathbf{v_i} \in \mathbb{R}^c$ with indices $i = \{1, 2, \ldots, hw\}$. The final prediction scores $\mathbf{F}$ for K classes are computed before applying the softmax function in the following way:

$$\mathbf{F} = \mathbf{W}^\top \frac{1}{hw} \sum_i \mathbf{v}_i = \frac{1}{hw} \sum_i \mathbf{W}^\top \mathbf{v}_i. \tag{6}$$

We aim to display the entropy of different local feature vectors $\mathbf{v_i}$ using varying weights. We denote $\mathbf{W}^\top \mathbf{v_i}$ as $\hat{\mathbf{F}}_i \in \mathbb{R}^K$. It represents the local class score vector of location i in the feature map. The semantic information contained in each $\mathbf{v}_i$ is extracted before the pooling operation. $\hat{\mathbf{p}}_i = \mathrm{softmax}(\hat{\mathbf{F}}_i)$ is the local class probability at location i. To compute the Shannon entropy [57] of $\hat{\mathbf{p}}_i$, we use $H(\hat{\mathbf{p}}_i) = -\sum_{k=1}^{K} \hat{\mathbf{p}}_i(k) \log \hat{\mathbf{p}}_i(k)$. Then we use a simple normalization function $\mathrm{Normalize}(x) := \frac{x - \min x}{\max x - \min x}$, to map the elements to the range [0,1]. This normalization process provides numerical information about the features, which is utilized to generate the feature entropy mask: $\mathcal{M} = \mathrm{Normalize}(H(\hat{\mathbf{p}}_i))$.

Then, channel duplication and spatial interpolation are applied in $\mathcal{M}$ to achieve scaling without introducing extra parameters. Taking inspiration from normalization technique [45, 75], the statistics of features from instances to be observable and manipulable x_d to achieve causal intervention. During mini-batch training, features $g(\mathrm{x})$ are randomly sampled from the training data. Subsequently, $\tilde{g}(\mathrm{x})$ is obtained by shuffle, the specific formula of shuffle, $\sigma(\cdot)$, and $\mu(\cdot)$ are shown in the supplementary material. We discussed in Eq. 4 that x_d variables in features with the maximum entropy and domain-related information is the most relevant. Hence, we must filter the statistical features from $\tilde{g}(\mathrm{x})$ to a significant extent to ensure that x_d comes from its domain. The simplest solution is to crop the featuremap and its corresponding entropy into patches of equivalent sizes. Then choose the region with the maximum entropy $\tilde{\mathcal{M}}^{max}_{crop}$ and the minimum entropy $\tilde{\mathcal{M}}^{min}_{crop}$, calculate mixed feature statistics using the following:

$$\begin{aligned}\tilde{\gamma}_{mix} &= \lambda\sigma(g(\mathrm{x})) + (1-\lambda)\sigma(\tilde{\mathcal{M}}^{max}_{crop}\odot\tilde{g}(\mathrm{x})) \quad \tilde{\beta}_{mix} = \lambda\mu(g(\mathrm{x})) + (1-\lambda)\mu(\tilde{\mathcal{M}}^{max}_{crop}\odot\tilde{g}(\mathrm{x}))\\ \gamma_{mix} &= \lambda\sigma(g(\mathrm{x})) + (1-\lambda)\sigma(\mathcal{M}^{min}_{crop}\odot g(\mathrm{x})) \quad \beta_{mix} = \lambda\mu(g(\mathrm{x})) + (1-\lambda)\mu(\mathcal{M}^{min}_{crop}\odot g(\mathrm{x})),\end{aligned} \tag{7}$$

where $\lambda \sim \mathrm{Beta}(\alpha,\alpha)$. The motivation for constructing $\mathcal{M}^{max}_{crop}$ comes from the second half of Eq. 5 to achieve a better causal intervention plan and minimize the inclusion of class-related variables from $\tilde{g}(\mathrm{x})$ in the intervention sample, thereby preventing blurring of the classification boundary in the feature space. The first half of Eq. 5 shows the relationship between entropy, the causal variable x_o, and domain-related information. The optimization goal for the current intervention sample should also consider the relationship $x_o \perp\!\!\!\perp d|y$. In addition, we can strengthen the semantic features of the samples themselves by $\mathcal{M}^{min}_{crop}$. Therefore, the mini-batch EnIn is formulated as follows:

$$\mathrm{EnIn}(g(\mathrm{x})) = \begin{cases}\mathcal{M}\odot g(\mathrm{x}) + \tilde{\gamma}_{\mathrm{mix}}\frac{g(\mathrm{x})-\mu(g(\mathrm{x}))}{\sigma(g(\mathrm{x}))} + \tilde{\beta}_{\mathrm{mix}}\\ \gamma_{\mathrm{mix}}\frac{g(\mathrm{x})-\mu(g(\mathrm{x}))}{\sigma(g(\mathrm{x}))} + \beta_{\mathrm{mix}}.\end{cases} \tag{8}$$

For simplification, during the forward pass of mini-batch training, EnIn is executed with a probability of 0.5 without gradients, and not executed during the test.

3.3 Testing: Homeostatic Score Based Causal Perturbation

Most DG algorithms prioritize the training stage by focusing on extracting domain-independent information from multiple source domains. Nevertheless, it is also crucial to deliberate how to model the features of unlabeled data rationally during the testing phase, to counter the issue of domain shift. Recent efforts [32, 64] utilize backpropagation to train the model on the target domain or construct a prototype classifier and adjust it during testing [8, 20]. However, they neglect two aspects: (1) the samples used for constructing category prototypes may inadequately represent the current category characteristics. (2) the samples affecting the decision boundary are still disregarded.

Using the causal graph constructed above, we build a prototype classifier using causal stable samples. Specifically, we modify the x_d attributes of test samples, introducing more class-related x_o information. This generates a feature representation closer to the category centroid, facilitating a denser cluster of category prototypes. Instead of only filtering high-entropy samples [20], which can limit generalization performance, we introduce the Homeostatic Score (HomeoScore) for sample selection. By comparing the probability difference between original and perturbed test samples, we exclude perturbation-sensitive samples and optimize the decision boundary.

During testing, we use a memory bank $\mathbb{B} = \{(g'(\mathrm{x}), p')\}$ to retain sample embedding features and predicted logits. We sample and extract feature representations $g(\mathrm{x_t})$ along with their class probabilities p_t and pseudo-label y_t at time t. As these representations are based on the source domain feature extractor, guaranteeing their generalizability is challenging. Therefore, we apply a feature transformation similar to EnIn via Eq. 5, where $\gamma_{mix} = \lambda\sigma(g(\mathrm{x_t})) + (1-\lambda)\sigma(\mathcal{M}^{min}_{crop}\odot g(\mathrm{x_t}))$ and $\beta_{mix} = \lambda\mu(g(\mathrm{x_t})) + (1-\lambda)\mu(\mathcal{M}^{min}_{crop}\odot g(\mathrm{x_t}))$, to obtain the target domain's new representation $g'(\mathrm{x_t})$ using Causal Perturbation (CP). CP boosts class-related information in $g(\mathrm{x_t})$, generating perturbed $p_t{}'$ and $y_t{}'$, and resulting in a tighter cluster of prototypes in the memory bank. Outliers are then filtered using the HomeoScore, defined as the distance between the original and perturbed sample representations:

$$\text{HomeoScore} = \left(\sum_{j=1}^{k} \left| p_t^j - p_t^{j\prime} \right|^2 \right)^{\frac{1}{2}}. \tag{9}$$

Then the prototype of class k could be formulated as :

$$\mathbb{B}_t^k = \begin{cases} \mathbb{B}_{t-1}^k \cup \left\{ \frac{g'(\mathrm{x_{t-1}})}{\|g'(\mathrm{x_{t-1}})\|} \right\} & \text{, if } y'_{(t-1)} = y^k \textit{ and } \text{HomeoScore} < \alpha \\ \mathbb{B}_{t-1}^k & \text{, else} \end{cases} \tag{10}$$

where α is a fixed threshold, which is set to 0.2. Figure 4 illustrates that HomeoScore can better distinguish erroneous pseudo-labels. We refrain from using samples with a larger HomeoScore for updating the memory bank as their pseudo-labels can damage the decision-making of the prototype classifier. The entropy threshold β controls the resource usage and operational efficiency of the memory bank using $\mathbb{B}_t^k = \left\{ g'(\mathrm{x}) \mid g'(\mathrm{x}) \in \mathbb{B}_t^k, H(p') \leq \beta \right\}$. Finally, we define the prototype-based classification output as the softmax over the feature similarities to prototypes for class k :

$$y_j^k = \frac{\exp\left(sim\left(g_j(\mathrm{x_t}), c_k\right)\right)}{\sum_{k'=1}^{|Y|} \exp\left(sim\left(g_j(\mathrm{x_t}), c_{k'}\right)\right)}, \tag{11}$$

where $c^k = \frac{1}{|\mathbb{B}^k|} \sum_{z \in \mathbb{B}_t^k} z$, $sim(\cdot,\cdot)$ denotes cosine similarity.

4 Experiments

To validate the generalization performance of our proposed method. We conduct experiments on image classification, semantic segmentation, and instance retrieval, specifically where the training and testing sets are from different domains. More experiment results and implementation details are presented in the supplementary material. All results are generated through five rounds of experiments with different random seeds.

4.1 Generalization on Multi-Source Domain Classification

Datasets and Implementation Details. We conduct classification experiments based on Dassl [75] and verify the generalization performance on two standard DG datasets PACS [28] and Office-home [63]. PACS consists of four different domains, each domain containing seven object categories. Office-Home spans four domains across sixty five categories.

Results on PACS and Office-Home based on ReNet-18 and ResNet-50 are reported in Table 1 and Table 2, respectively. It can be observed that `Mixstyle-Entropy` reaches the highest average accuracy among all the compared methods on both backbones even without HoPer. For PACS, compared with CIRL [36], `Mixstyle-Entropy` outperforms it by a large margin of 2.2 % and on ResNet-18. Our approach is also superior to the ensemble method I^2-ADR [40] on the parameter-rich ResNet-50, achieving a higher accuracy of 2.8 % improvement. For Office-Home, our approach exhibits a 3.4 % improvement over CIRL on ResNet-18 and achieves a 4.5 % improvement over the previous SOTA [40] on ResNet-50.

	PACS					Office-Home				
Method	Art	Cartoon	Photo	Sketch	Avg.(%)	Art	Clipart	Product	Real	Avg.(%)
Baseline	74.3	76.7	**96.4**	68.7	79.0	58.8	48.3	74.2	76.2	64.4
Mixup [69]	76.8	74.9	95.8	66.6	78.5	58.2	49.3	74.7	76.1	64.6
RSC [19]	78.9	76.9	94.1	76.8	81.7	58.4	47.9	71.6	74.5	63.1
L2A-OT [72]	83.3	78.2	96.2	76.3	82.8	60.6	50.1	74.8	77.0	65.6
Mixstyle [75]	82.3	79.0	96.3	73.8	82.8	58.7	53.4	74.2	75.9	65.5
DSU [31]	83.6	79.6	95.8	77.6	84.1	60.2	54.8	74.1	75.1	66.1
CIRL [36]	86.1	80.6	95.9	82.7	86.3	61.5	55.3	75.1	76.6	67.1
`Mixstyle-Entropy`	**88.5**±0.4	**84.2**±0.3	95.3±0.2	**85.8**±0.5	**88.5**±0.3	**67.0**±0.4	**62.3**±0.4	**75.5**±0.4	**77.3**±0.2	**70.5**±0.3
- HomeoScore	88.0±0.5	83.6±0.4	95.1±0.2	84.8±0.3	87.9±0.3	66.7±0.5	61.8±0.6	74.7±0.4	77.0±0.2	70.1±0.3
- HoPer	86.8±0.3	82.5±0.3	94.9±0.6	82.3±0.2	86.7±0.4	65.7±0.5	60.3±0.3	73.7±0.5	76.6±0.4	69.1±0.4

Table 1: Leave-one-domain-out multi-domain classification on ResNet-18.

	PACS					Office-Home				
Method	Art	Cartoon	Photo	Sketch	Avg.(%)	Art	Clipart	Product	Real	Avg.(%)
Baseline	86.2	78.7	97.6	70.6	83.2	61.3	52.4	75.8	76.6	66.5
RSC [19]	87.8	82.1	**97.9**	83.3	87.9	50.7	51.4	74.8	75.1	65.5
SelfReg [22]	87.9	79.4	96.8	78.3	85.6	63.6	53.1	76.9	78.1	67.9
SagNet [42]	81.1	75.4	95.7	77.2	82.3	63.4	54.8	75.8	78.3	68.1
I^2-ADR [40]	88.5	83.2	95.2	85.8	88.2	70.3	55.1	80.7	79.2	71.4
`Mixstyle-Entropy`	**90.9**±0.4	**86.6**±0.6	95.9±0.3	**86.5**±0.7	**90.0**±0.5	**74.4**±0.4	**66.8**±0.5	**80.6**±0.4	**81.6**±0.4	**75.9**±0.4
- HomeoScore	90.3±0.5	85.8±0.5	95.7±0.2	86.1±0.4	89.5±0.3	73.9±0.5	66.2±0.4	79.7±0.3	80.9±0.5	75.1±0.4
- HoPer	89.9±0.5	85.2±0.5	95.5±0.3	85.7±0.4	89.1±0.4	73.1±0.6	65.3±0.4	78.9±0.3	80.1±0.3	74.4±0.4

Table 2: Leave-one-domain-out multi-domain classification on ResNet-50.

Highlights. `Mixstyle-Entropy` does not require any additional training parameters. Compared to previous methods that expand the embedding space, such as Mixstyle [75] and DSU [36], our method consistently surpasses them in terms of average accuracy by 5.3 % and 4.4 %, respectively. Even after removing the HoPer module, our approach still achieves SOTA performance, surpassing them by 3.7 % and 2.8%. Notably, `Mixstyle-Entropy` can be seamlessly integrated into any existing DG framework.

4.2 Generalization on Semantic Segmentation

Datasets and Implementation Details. Semantic segmentation impacts autonomous driving, yet dynamic driving scenes can hamper it. We assess `Mixstyle-Entropy` generalizability on the gaming-based GTA5 [52] image dataset and the Cityscapes [11] dataset featuring German urban street scenes. Our implementation aligns with prior work DSU [36].

Results on Semantic Segmentation are shown in Table 3. It emphasizes that appropriate feature space expansion is vital in pixel-level classification. Overstepping other categories' embedding spaces could lead to incorrect predictions, illustrated in Figure 2. In sunny conditions, vehicle shadows can lead to car misclassifications. Only EnIn accurately segments the car region by breaking this false association. The failure to differentiate sidewalks from road surfaces in pixel predictions shows the danger of indiscreet feature space interpolation, particularly in autonomous driving cases. Only our method correctly segments such scenes, providing clear differentiation.

Method	mIoU(%)	mAcc(%)
ERM	36.0	49.5
pAdaIN [45]	38.1	51.1
Mixstyle [75]	38.7	51.2
DSU [31]	40.7	**53.8**
EnIn (Ours)	**42.6**	53.7

Table 3: Results of semantic segmentation from GAT5 to Cityscapes.

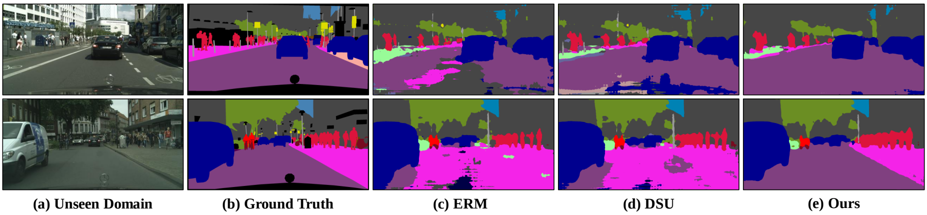


Figure 2: The visualization on unseen domain Cityscapes with the model trained on GTA5.

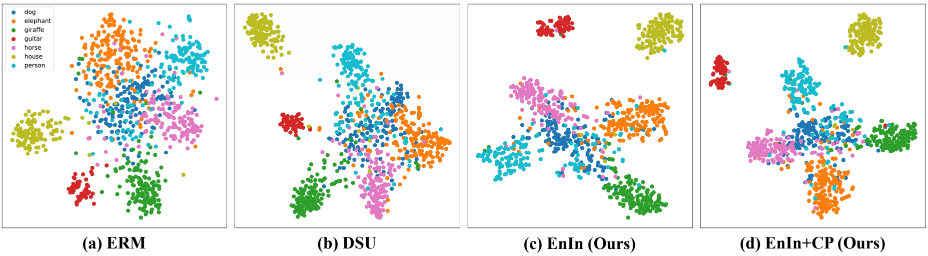


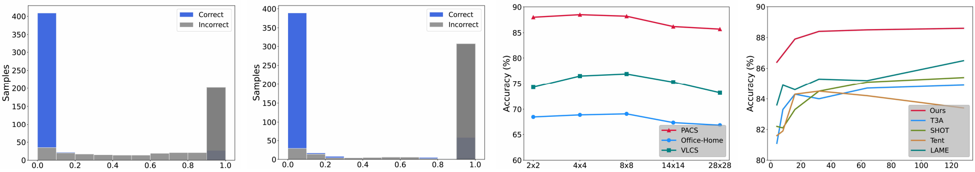
Figure 3: The t-SNE visualization of extracted deep features using different methods on PACS dataset (target domain: cartoon). The different colors stands for different classes.

5 Analysis and Ablations

Ablation Study. `Mixstyle-Entropy`, designed as a plug-and-play module with minimal hyperparameters, hinges on two components: EnIn and HoPer. HomeoScore filtering in HoPer improves average accuracy by 0.6% through providing superior samples, as seen in Tables 1 and 2. HomeoScore outperforms traditional Entropy in pseudo-label filtering (Figure 4). Incorrectly predicted test domain samples possess higher HomeoScore, hence, filtering these optimizes the prototype classifier decision boundary. Despite HoPer exclusion during testing, `Mixstyle-Entropy` surpasses most previous works.

Effects of Causal Perturbation. During the testing phase, we apply causal perturbation (CP) to the samples to further inject class-related information. To better demonstrate the sample distribution in the feature space, we employ t-SNE [61] to visualize the embeddings of the test domain samples, as shown in Figure 3. DSU [36] using multidimensional Gaussian distributions to simulate the distributions, leads to decreased inter-class distances and blurry decision boundaries. In contrast, EnIn implicitly measures the inter-cluster and intra-cluster distances through the feature entropy representation, resulting in an expansion of inter-cluster distances. After undergoing CP, the intra-cluster distances decrease, leading to more compact clusters and clear classification boundaries for the model.

Effects of Patch size and Test Batch Size The selection of $\mathcal{M}$ is based on patching the feature maps. It was found that a ratio of $\frac{1}{4}$ or $\frac{1}{8}$ resulted in better performance on all three datasets in Figure 4. Further reducing the ratio led to a significant decline in model performance due to insufficient statistical features to represent x_o. Additionally, we compared `Mixstyle-Entropy` with other TTA proposals. It was observed that the batch size during testing has a significant impact on the direction of stochastic gradient descent, they



(a) Entropy (b) HomeoScore (c) Number of Patches (d) Test Batch Size

Figure 4: (a) and (b) are pseudo-label selection capability between entropy and HomenScore. (c) is effects of patch sizes. (d) is the influence of test batch size in different schemes.

performed poorly with small batch sizes. In contrast, `Mixstyle-Entropy` consistently demonstrated excellent performance. This also highlights the explicit reduction of domain divergence [71] achieved by a parameter-free classifier.

6 Conclusions

In this paper, we propose the `Mixstyle-Entropy` framework, guided by causal learning, to achieve OOD generalization. The core idea is to decouple domain- and class-related variables using feature representation entropy and to disrupt the relationship between the causal variables and the domain through causal intervention. Furthermore, we unify the training and testing processes and construct a classifier adapted to the target domain using variables stabilized by causal relationships. A limitation that needs to be addressed in future work is the extension of our framework to more backbones, such as vision Transformer and multilayer perceptron. The comprehensive experiments on various benchmarks demonstrate the effectiveness and superiority of our proposal. `Mixstyle-Entropy` achieves state-of-the-art performance in a plug-and-play manner, without additional training parameters.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 82172033, Grant 52105126, Grant 82272071, and Grant 62271430; and in part by the Dreams Foundation of Jianghuai Advance Technology Center; and in part by the Open Fund of the National Key Laboratory of Infrared Detection Technologies.

References

[1] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 315–331, 2020.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[3] Malik Boudiaf, Romain Müller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8334–8343. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00816. URL https://doi.org/10.1109/CVPR52688.2022.00816.

[4] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7119–7129, 2022.

[5] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3677–3694, 2022.

[6] Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. *Advances in Neural Information Processing Systems*, 35:33302–33315, 2022.

[7] Chaoqi Chen, Luyao Tang, Yue Huang, Xiaoguang Han, and Yizhou Yu. Coda: generalizing to open and unseen domains with compaction and disambiguation. *Advances in Neural Information Processing Systems*, 36, 2023.

[8] Chaoqi Chen, Luyao Tang, Leitian Tao, Hong-Yu Zhou, Yue Huang, Xiaoguang Han, and Yizhou Yu. Activate and reject: towards safe domain generalization under category shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11552–11563, 2023.

[9] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and J. Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:6614–6630, 2020.

[10] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 6447–6458, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2974788b53f73e7950e8aa49f3a306db-Abstract.html.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1180–1189. JMLR.org, 2015.

[14] Muhammad Ghifary, W. Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2551–2559, 2015.

[15] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.

[16] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

[17] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Mach. Learn.*, 110(2):303–348, 2021. URL https://doi.org/10.1007/s10994-020-05924-1.

[18] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019*.

[19] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 124–140. Springer, 2020.

[20] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2427–2440, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1415fe9fea0fa1e45dddcff5682239a0-Abstract.html.

[21] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3140–3149, 2020.

[22] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2020.

[24] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International conference on learning representations (ICLR)*, 2018.

[25] Arie W Kruglanski. The endogenous-exogenous partition in attribution theory. *Psychological Review*, 82(6):387, 1975.

[26] Chenxin Li, Yunlong Zhang, Zhehan Liang, Wenao Ma, Yue Huang, and Xinghao Ding. Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 61–65. IEEE, 2021.

[27] Chenxin Li, Xin Lin, Yijin Mao, Wei Lin, Qi Qi, Xinghao Ding, Yue Huang, Dong Liang, and Yizhou Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in biology and medicine*, 141: 105144, 2022.

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[29] Hanxuan Li, Bin Fu, Ruiping Wang, and Xilin Chen. Point2real: Bridging the gap between point cloud and realistic image for open-world 3d recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3055–3063, 2024.

[30] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5400–5409. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00566. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Li_Domain_Generalization_With_CVPR_2018_paper.html.

[31] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Lingyu Duan. Uncertainty modeling for out-of-distribution generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. ICLR, 2022.

[32] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[33] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6804–6814. PMLR, 2021. URL http://proceedings.mlr.press/v139/liu21h.html.

[34] Jiawei Liu, Zhipeng Huang, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Debiased batch normalization via gaussian process for generalizable person re-identification. In *AAAI*, pages 1729–1737. AAAI Press, 2022.

[35] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: when does self-supervised test-time training fail or thrive? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 21808–21820, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b618c3210e934362ac261db280128c22-Abstract.html.

[36] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, 2022.

[37] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

[38] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7313–7324. PMLR, 2021. URL http://proceedings.mlr.press/v139/mahajan21b.html.

[39] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: Domain generalization through source-specific nets. *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1353–1357, 2018.

[40] Rang Meng, Xianfeng Li, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Mingli Song, Di Xie, and Shiliang Pu. Attention diversification for domain generalization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 322–340. Springer, 2022.

[41] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5716–5726, 2017.

[42] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[43] Narges Honarvar Nazari and Adriana Kovashka. Domain generalization using shape representation. In *ECCV Workshops*, 2020.

[44] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16888–16905. PMLR, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

[45] Oren Nuriel, Sagie Benaim, and Lior Wolf. Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9482–9491, 2021.

[46] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. *ECCV*, 2018.

[47] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Penguin Books Limited, 2018. ISBN 9780241242643. URL https://books.google.de/books?id=EmY8DwAAQBAJ.

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

[49] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[50] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[51] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019. URL http://proceedings.mlr.press/v97/recht19a.html.

[52] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[53] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1): 1309–1342, 2018.

[54] Bernhard Schölkopf. *Causality for Machine Learning*, page 765–804. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501755.

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

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

[57] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[58] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 3118–3127. PMLR, 2019. URL http://proceedings.mlr.press/v89/subbaswamy19a.html.

[59] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 2020. URL http://proceedings.mlr.press/v119/sun20b.html.

[60] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, V. Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1082–10828, 2018.

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

[62] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991.

[63] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[64] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. ICLR, 2021.

[65] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

[66] Yunqi Wang, Furui Liu, Zhitang Chen, Yik-Chung Wu, Jianye Hao, Guangyong Chen, and Pheng-Ann Heng. Contrastive-ace: Domain generalization through alignment of causal mechanisms. *IEEE Transactions on Image Processing*, 32:235–250, 2022.

[67] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Toalign: Task-oriented alignment for unsupervised domain adaptation. In *Neural Information Processing Systems*, 2021.

[68] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto L. Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2100–2110, 2019.

[69] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*. ICLR, 2018.

[70] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8035–8045, 2022.

[71] Yi-Fan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. *arXiv preprint arXiv:2304.12566*, 2023.

[72] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 561–578. ECCV, 2020.

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

[74] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2020.

[75] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. ICLR, 2021.