Robustness to Spurious Correlations via Dynamic Knowledge Transfer

Xiaoling Zhou[1], **Wei Ye**[1*], **Zhemg Lee**[2], **Shikun Zhang**[1*]
[1]Peking University
[2]Tianjin University
xiaolingzhou@stu.pku.edu.cn, wye@pku.edu.cn, zhemglee@tju.edu.cn, zhangsk@pku.edu.cn

Abstract

Spurious correlations pose a significant challenge to the robustness of statistical models, often resulting in unsatisfactory performance when distributional shifts occur between training and testing data. To address this, we propose to transfer knowledge across spuriously correlated categories within the deep feature space. Specifically, samples' deep features are enriched using semantic vectors extracted from both their respective category distributions and those of their spuriously correlated counterparts, enabling the generation of diverse class-specific factual and counterfactual augmented deep features. We then demonstrate the feasibility of optimizing a surrogate robust loss instead of conducting explicit augmentations by considering an infinite number of augmentations. As spurious correlations between samples and classes evolve during training, we develop a reinforcement learning-based training framework called Dynamic Knowledge Transfer (DKT) to facilitate dynamic adjustments in the direction and intensity of knowledge transfer. Within this framework, a target network is trained using the derived robust loss to enhance robustness, while a strategy network generates sample-wise augmentation strategies in a dynamic and automatic way. Extensive experiments validate the effectiveness of the DKT framework in mitigating spurious correlations, achieving state-of-the-art performance across three typical learning scenarios susceptible to such correlations.

1 Introduction

While deep learning models have demonstrated remarkable performance across a wide range of tasks, studies have highlighted that deep models frequently capture spurious correlations between non-causal attributes and classes, posing a threat to the validity and reliability of the models [Wu *et al.*, 2023; Deng *et al.*, 2024; Tian *et al.*, 2025]. For instance, in natural language processing tasks, classifiers may learn that the term "Spielberg" is correlated with positive movie reviews, or that the word "ugly" is associated with toxic comments [Gupta *et al.*, 2022]. Similarly, in computer vision tasks, an image classifier might associate a forest background with the label "bird" [Deng *et al.*, 2024]. The embedding of spurious correlations during training can lead to performance degradation when faced with varying distributions in test data, raising concerns regarding prediction robustness and trustworthiness.

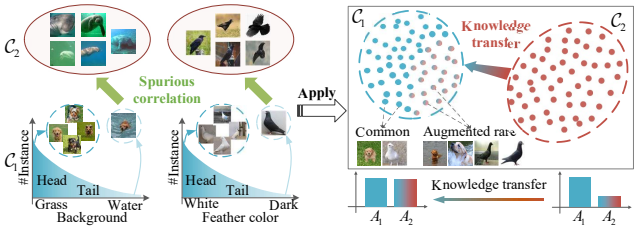

Figure 1: Illustration of spurious correlations induced by imbalanced attributes. Transferring knowledge across categories facilitates the generation of instances with rare attributes, aiding in mitigating such pseudo correlations. Here, A and C refer to attributes and classes, respectively.

Data augmentation has proven to be effective in helping models discern genuinely significant features by diversifying and balancing the training data [Zhou and Wu, 2023; Deng *et al.*, 2024]. Among these techniques, counterfactual data augmentation, which generates counterfactual samples with modified attributes, emerges as one of the most potent approaches for dismantling spurious correlations between non-causal attributes and classes [Wu *et al.*, 2024; Reddy *et al.*, 2023]. However, determining which features should be modified and how to modify them is a complex issue that often relies on human annotation or group information, making the process both costly and time-consuming [Deng *et al.*, 2023]. Additionally, training models with expanded training data inevitably leads to decreased training efficiency [Wang *et al.*, 2025]. Implicit semantic data augmentation [Wang *et al.*, 2019; Zhou *et al.*, 2024] is a technique that transforms samples along semantic directions within the deep feature space. Notably, this method is accomplished solely by optimizing a novel robust loss, ensuring efficiency. Building upon their research, we believe that generating counterfactual samples within the deep feature space could eliminate

*Corresponding authors.

the need for explicit differentiation between causal and non-causal attributes, while also bypassing the necessity to expand the training dataset.

Drawing on the aforementioned insights, this study proposes enhancing model robustness by dynamically transferring knowledge across spuriously correlated classes in the deep feature space. Specifically, the deep features of samples are transformed using semantic vectors extracted from both their respective category distributions and those of their spuriously correlated classes. This process generates class-specific factual and counterfactual augmented deep features, aiding in enhancing intra-class data diversity and reducing inter-class spurious correlations. As illustrated in Fig. 1, pigeons with dark feathers are spuriously correlated with the crow class due to the prevalence of dark feathers in the crow class and their rarity in the pigeon class. Conveying knowledge from the crow class to the pigeon class, e.g., generating counterfactual instances of pigeons with dark feathers, helps the model rely more on causal features rather than color information during prediction. By considering an infinite number of augmentations, we theoretically derive a surrogate loss function for the proposed augmentation strategy. Consequently, rather than explicitly augmenting sample features, we can directly minimize this robust loss function, leading to a highly efficient algorithm.

To dynamically adjust the direction and intensity of knowledge transfer across categories during training, we propose a novel reinforcement learning-based training framework that incorporates a target network and a strategy network. The target network is trained using the derived robust loss to enhance its resilience. Meanwhile, as training progresses, the strategy network dynamically adjusts the sample-specific augmentation strategies, encompassing both augmentation distribution and strength, to align with the evolving learning dynamics of the target network. We conduct extensive experiments across three typical learning scenarios susceptible to spurious correlations: subpopulation shift learning, generalized long-tail (GLT) learning, and domain shift learning, encompassing benchmarks from both text and image domains. The results demonstrate that our method consistently attains state-of-the-art (SOTA) performance across various learning scenarios, affirming its efficacy in mitigating the deleterious effects of spurious correlations.

In summary, our main contributions are threefold:

- We undertake a pioneering effort to reduce spurious correlations through dynamic inter-class knowledge transfer. This approach generates diverse factual and counterfactual augmented deep features, achieved exclusively through the optimization of a robust loss function, thereby enhancing efficiency and applicability.
- We present a new reinforcement learning-based training framework, termed dynamic knowledge transfer (DKT), for training classifiers using the derived robust loss function, where the direction and intensity of knowledge transfer are dynamically and automatically determined by a strategy network based on the unique training characteristics of samples.
- We conduct comprehensive experiments across three typical learning scenarios characterized by distribution shifts between training and test data. The results unequivocally demonstrate the effectiveness and broad applicability of our approach.

2 Related Work

Prior research has demonstrated that deep learning models frequently rely on spurious patterns for predictions, consequently exhibiting inadequate generalization and robustness when confronted with unseen environments [Moayeri *et al.*, 2022; Zhao *et al.*, 2025; Veitch *et al.*, 2021; Sun *et al.*, 2021]. For instance, models trained on ImageNet often classify images based on background rather than foreground attributes [Moayeri *et al.*, 2022]. Similarly, Young et al. [2019] demonstrated that deep networks for CT scans, despite high accuracy, often produce explanations outside the relevant regions when visualized with Grad-CAM [Selvaraju *et al.*, 2017]. Additionally, Chew et al. [2024] found that a sentiment classifier might mistakenly learn that the word "performance" is associated with positive reviews, even if the word itself is not commendatory.

To eliminate spurious correlations, various approaches have been proposed, including dataset modifications [Zhao *et al.*, 2024; Wu *et al.*, 2024], causal inference [Tang *et al.*, 2020], model ensembles [Zhang *et al.*, 2021; Wang *et al.*, 2020], regularization techniques [Tang *et al.*, 2022; Krueger *et al.*, 2021; Zhang *et al.*, 2022], and re-training strategies [Zhao *et al.*, 2023; Zhou *et al.*, 2023b]. Among these methods, our work is particularly aligned with counterfactual data augmentation, which involves generating counterfactual samples by modifying attributes to train deep learning models [Wu *et al.*, 2024; Reddy *et al.*, 2023]. For instance, Gupta et al. [2022] mitigated gender bias in the text by swapping gender words (e.g., "he" becomes "she") to augment the data. Xiao et al. [2023] employed masked images, where either semantics-related or semantics-unrelated patches were masked, as counterfactual samples to enhance the robustness of the fine-tuning model. Moreover, Reddy et al. [2023] enabled the generation of counterfactual data by quantifying and eliminating confounding bias. While effective, existing augmentation methods rely on explicit differentiation between causal and non-causal attributes to produce counterfactual data [Gupta *et al.*, 2022; Xiao *et al.*, 2023], which can be resource-intensive and restrictive in their scope. Additionally, these strategies are hand-crafted, thereby lacking flexibility and potentially constraining the generalization and robustness of models [Deng *et al.*, 2023].

3 Methodology

This study introduces a novel learning framework to diminish spurious correlations by transferring knowledge across categories in the deep semantic space. We start by elucidating the mechanism of the proposed knowledge transfer strategy. Following this, we theoretically derive a robust loss function to implement our strategy efficiently. Finally, we propose a reinforcement learning-based training framework that dynamically determines the direction and intensity of knowledge transfer throughout the training process.

Notation Consider training a deep classifier f, with weights $\boldsymbol{\theta}$, on a training set denoted as $\mathcal{D}^{tr} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where N represents the number of training samples, and $y_i \in \{1, \cdots, \mathcal{C}\}$ denotes the label of sample $\boldsymbol{x}_i$. The deep feature (before logit) learned by f for $\boldsymbol{x}_i$ is represented as a $\mathcal{Z}$-dimensional vector $\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \in \mathbb{R}^{\mathcal{Z}}$. Moreover, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents a multi-variant Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

3.1 Knowledge Transfer Strategy

To alleviate spurious correlations between non-causal attributes and classes, this study proposes transferring knowledge across categories within the deep semantic space. In particular, the deep features of samples undergo transformations along semantic directions derived from the deep feature distributions of their respective categories and those exhibiting spurious correlations. These transformations generate augmented deep features that encompass both factual and counterfactual variations specific to each class, thereby enriching the diversity of training data and helping to eliminate erroneous associations between samples and their non-ground-truth categories.

In line with Wang et al. [2019], we enhance intra-class data diversity by sampling semantic vectors for the deep feature of each sample, $\boldsymbol{z}_i$, from a zero-mean multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{y_i})$. Here, $\boldsymbol{\Sigma}_{y_i}$ represents the class-conditional covariance matrix of class y_i. Additionally, to eliminate spurious correlations across categories, semantic vectors are extracted from the category distributions with spurious associations in the samples. In our approach, the class with the highest predicted probability, excluding the ground-truth class, is treated as the spuriously correlated class. Let the spuriously correlated category for sample $\boldsymbol{x}_i$ be denoted as class c. We perform counterfactual augmentation by sampling semantic vectors from the deep feature distribution of class c, represented as $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. The feature mean $\boldsymbol{\mu}_c$ of each class is computed as:

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \boldsymbol{z}_i, \tag{1}$$

where N_c represents the number of samples in class c. Additionally, the covariance matrix $\boldsymbol{\Sigma}_c$ is calculated as:

$$\boldsymbol{\Sigma}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} (\boldsymbol{z}_i - \boldsymbol{\mu}_c)^T (\boldsymbol{z}_i - \boldsymbol{\mu}_c). \tag{2}$$

In implementation, the mean and covariance matrix are computed online by aggregating statistics from all mini-batches. Specifically, as training progresses, the feature means are estimated as follows:

$$\boldsymbol{\mu}_c^t = \frac{n_c^{t-1} \boldsymbol{\mu}_c^{t-1} + m_c^t \boldsymbol{\mu}_c'^t}{n_c^{t-1} + m_c^t}, \tag{3}$$

where $\boldsymbol{\mu}_c^t$ and $\boldsymbol{\mu}_c'^t$ represent the feature mean estimates of class c at the tth step and within the tth mini-batch, respectively. n_c^{t-1} and m_c^t denote the cumulative count of training samples belonging to class c across all $t-1$ mini-batches and the count of training samples belonging to class c in the tth mini-batch. The covariance matrices are estimated by:

$$\begin{aligned} \boldsymbol{\Sigma}_c^t = & \frac{n_c^{t-1} \boldsymbol{\Sigma}_c^{t-1} + m_c^t \boldsymbol{\Sigma}_c'^t}{n_c^{t-1} + m_c^t} \\ & + \frac{n_c^{t-1} m_c^t \left(\boldsymbol{\mu}_c^{t-1} - \boldsymbol{\mu}_c'^t\right)\left(\boldsymbol{\mu}_c^{t-1} - \boldsymbol{\mu}_c'^t\right)^T}{\left(n_c^{t-1} + m_c^t\right)^2}, \end{aligned} \tag{4}$$

where $\boldsymbol{\Sigma}_c^t$ and $\boldsymbol{\Sigma}_c'^t$ represent the covariance matrix estimates of class c at the tth step and within the tth mini-batch, respectively. Furthermore, $n_c^t = n_c^{t-1} + m_c^t$.

Based on the unique training dynamics of the samples, the augmentation distributions for different samples, which control the direction of knowledge transfer, should vary. In other words, the proportions of semantic and counterfactual augmentations should differ across samples. Consequently, the augmented deep features $\tilde{\boldsymbol{z}}_i$ for $\boldsymbol{z}_i$ are given by

$$\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{z}_i + \alpha_i \boldsymbol{\mu}_c, \beta_i \boldsymbol{\Sigma}_{y_i} + \alpha_i \boldsymbol{\Sigma}_c), \tag{5}$$

where α_i and β_i are two positive coefficients controlling the proportions of counterfactual and factual augmentations, respectively. Their values are automatically and dynamically computed by a deep network, utilizing the distinctive training characteristics of the samples as input, which will be detailed in Section 3.3. Notably, our method can sample arbitrary semantic directions from spuriously correlated classes for counterfactual augmentation. While we cannot guarantee the exclusion of inherent attributes from spurious categories, the limited scope of counterfactual perturbations and the retention of original labels are expected to encourage the model to focus more on other intrinsic attributes of the current class for classification. Additionally, since early epoch estimates are less informative when the network is undertrained, both values are decayed by a factor of t/T, where t and T denote the current and total number of epochs, respectively.

As for the augmentation strength, that is the number of augmented features for each sample, it is assumed to follow $\mathcal{K}_i = \mathcal{K} \times \gamma_i$, where $\mathcal{K}$ is a constant and γ_i refers to the strength factor. The value of γ_i is also dynamically calculated based on the unique training characteristics of sample $\boldsymbol{x}_i$. Consequently, for each deep feature $\boldsymbol{z}_i$, a set of augmented deep features can be obtained in each iteration, represented as $\{\tilde{\boldsymbol{z}}_i^1, \tilde{\boldsymbol{z}}_i^2, \cdots, \tilde{\boldsymbol{z}}_i^{\mathcal{K}_i}\}$.

3.2 Surrogate Robust Loss Function

Given the computational inefficiency of directly using all augmented deep features for training, we address this by exploring the scenario where the number of augmented features for each sample approaches infinity and derive an upper bound for the expected Cross-Entropy (CE) loss. This approach allows us to achieve a highly efficient implementation while maintaining the benefits of augmentation.

The CE loss for all augmented features is as follows:

$$\mathcal{L}_{\mathcal{K}}(\boldsymbol{W}, \boldsymbol{b}) = -\frac{1}{\hat{\mathcal{K}}} \sum_{i=1}^{N} \sum_{j=1}^{\mathcal{K}_i} \log \frac{e^{\boldsymbol{w}_{y_i}^T \tilde{\boldsymbol{z}}_i^j + b_{y_i}}}{\sum_{k=1}^{\mathcal{C}} e^{\boldsymbol{w}_k^T \tilde{\boldsymbol{z}}_i^j + b_k}}, \tag{6}$$

where $\hat{\mathcal{K}} = \sum_{i=1}^{N} \mathcal{K}_i$. $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{\mathcal{C}}]^T \in \mathbb{R}^{\mathcal{C} \times \mathcal{Z}}$ and $\boldsymbol{b} = [b_1, \cdots, b_{\mathcal{C}}]^T \in \mathbb{R}^{\mathcal{C}}$. Each $\boldsymbol{w}_k$ and b_k represent the

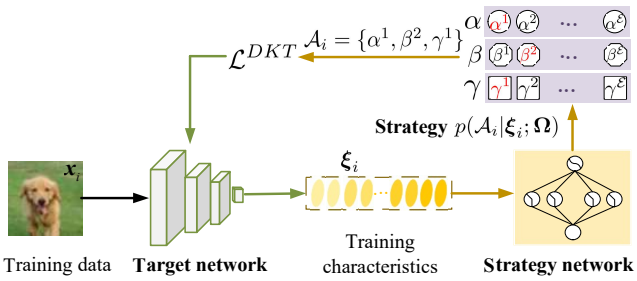

Figure 2: Illustration of the DKT framework. Given a sample, the target network first generates a series of training characteristics, which are then input into the strategy network to determine the sample-wise augmentation strategies (including distributions and strengths). Using these strategies, the target network can be trained with the derived DKT loss $\mathcal{L}^{DKT}$ to enhance its robustness.

weight vector and bias, respectively, associated with the final fully connected layer for class k. We then investigate the scenario of augmenting an infinite number of times on the deep feature of each sample. As $\mathcal{K}$ in $\mathcal{K}_i$ approaches infinity, the expected CE loss is expressed as:

$$\mathcal{L}_\infty = -\sum_{i=1}^{N} \gamma_i \mathrm{E}_{\tilde{\boldsymbol{z}}_i}[\log \frac{e^{\boldsymbol{w}_{y_i}^T \tilde{\boldsymbol{z}}_i + b_{y_i}}}{\sum_{j=1}^{\mathcal{C}} e^{\boldsymbol{w}_j^T \tilde{\boldsymbol{z}}_i + b_j}}]. \tag{7}$$

However, accurately calculating Eq. (7) is computationally challenging, and thus we aim to derive a more efficient surrogate loss for it. Let $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ denote the mean and covariance matrix for the spuriously correlated class of sample $\boldsymbol{x}_i$. Due to the concave nature of the logarithmic function $\log(\cdot)$, with Jensen's inequality, $\mathrm{E}[\log X] \leq \log \mathrm{E}[X]$, we derive an upper bound for the expected loss in Eq. (7):

$$\mathcal{L}_\infty \leq \sum_{i=1}^{N} \gamma_i \log(\sum_{j=1}^{\mathcal{C}} \mathrm{E}_{\tilde{\boldsymbol{z}}_i}[e^{(\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)\tilde{\boldsymbol{z}}_i + (b_j - b_{y_i})}]). \tag{8}$$

Given that $\tilde{\boldsymbol{z}}_i$ follows a Gaussian distribution characterized by $\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{z}_i + \alpha_i \hat{\boldsymbol{\mu}}_i, \beta_i \boldsymbol{\Sigma}_{y_i} + \alpha_i \hat{\boldsymbol{\Sigma}}_i)$, we have $(\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)\tilde{\boldsymbol{z}}_i + (b_j - b_{y_i}) \sim \mathcal{N}\big((\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)(\boldsymbol{z}_i + \alpha_i \hat{\boldsymbol{\mu}}_i) + (b_j - b_{y_i}), (\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)(\beta_i \boldsymbol{\Sigma}_{y_i} + \alpha_i \hat{\boldsymbol{\Sigma}}_i)(\boldsymbol{w}_j - \boldsymbol{w}_{y_i})\big)$. Then, leveraging the moment-generating function $\mathrm{E}\left[e^{tX}\right] = e^{t\mu + \frac{1}{2}\sigma^2 t^2}$, the upper bound in Eq. (8) can be written as

$$\mathcal{L}_\infty \leq -\sum_{i=1}^{N} \gamma_i \log \frac{e^{\mathcal{F}_i^{y_i}}}{\sum_{j=1}^{\mathcal{C}} e^{\mathcal{F}_i^j}} := \mathcal{L}^{DKT}, \tag{9}$$

where $\mathcal{F}_i^j = \boldsymbol{w}_j^T(\boldsymbol{z}_i + \alpha_i \hat{\boldsymbol{\mu}}_i) + b_j + \frac{1}{2}(\boldsymbol{w}_j^T - \boldsymbol{w}_{y_i}^T)(\beta_i \boldsymbol{\Sigma}_{y_i} + \alpha_i \hat{\boldsymbol{\Sigma}}_i)(\boldsymbol{w}_j - \boldsymbol{w}_{y_i})$.

Consequently, the above deviation provides a surrogate loss for our proposed knowledge transfer strategy. Instead of explicitly performing the augmentation process, we can directly optimize the derived upper bound ($\mathcal{L}^{DKT}$) during classifier training, thereby enhancing efficiency.

3.3 Reinforcement Learning-Based Framework

The augmentation distribution and strength for each sample dictate the direction and intensity of knowledge transfer. As determining the optimal augmentation strategies is inherently a parameter selection problem without gold labels, we employ reinforcement learning to address it. The workflow of our proposed DKT framework is illustrated in Fig. 2.

Pipeline of the DKT Framework

Our framework comprises a target network and a strategy network. The target network undergoes training using the derived DKT loss to bolster its robustness. Meanwhile, the strategy network receives a series of training characteristics extracted from the target network as input and generates sample-specific augmentation strategies, encompassing augmentation distributions and strengths. Given that the input to the strategy network is tabular data, we design it as a two-layer Multilayer Perceptron. Accordingly, as training progresses, the strategy network dynamically adjusts augmentation strategies to align with the evolving learning dynamics of the target network.

Let $\mathcal{A}_i = \{\alpha_i, \beta_i, \gamma_i\} \in \mathcal{A}$ represent an augmentation strategy for sample $\boldsymbol{x}_i$, where α_i and β_i jointly determine the augmentation distribution, while γ_i determines the augmentation strength. Here, $\mathcal{A}$ represents the value space of augmentation strategies. Each parameter has $\mathcal{E}$ options, which are encoded by a one-hot vector. For example, γ_i takes values from the set $\{\gamma^1, \gamma^2, \cdots, \gamma^{\mathcal{E}}\}$, providing $\mathcal{E}$ options (γ^1 to $\gamma^{\mathcal{E}}$) such as $\{0.1, 0.2, \cdots, 1\}$ for the augmentation strength of each sample. If $\mathcal{A}_i = \{0.1, 0.2, 0.1\}$, then the augmentation distribution for $\boldsymbol{z}_i$ is $\mathcal{N}(0.1\boldsymbol{\mu}_c, 0.2\boldsymbol{\Sigma}_{y_i} + 0.1\boldsymbol{\Sigma}_c)$, and its augmentation strength is $0.1\mathcal{K}$. Nevertheless, instead of explicitly augmenting deep features, the parameters of augmentation strategies are employed to compute the derived DKT loss $\mathcal{L}^{DKT}$ in our framework. Accordingly, during the training process, the strategy network captures the conditional distribution $p(\mathcal{A}_i|\boldsymbol{\xi}_i;\boldsymbol{\Omega})$ where $\boldsymbol{\xi}_i$ denotes the training characteristics of sample $\boldsymbol{x}_i$, detailed in subsequent subsections, and $\boldsymbol{\Omega}$ represents the parameters of the strategy network.

Training Characteristics Extraction

In each iteration, a series of training characteristics are extracted from the classifier for each sample, which serves to capture the degree of spurious correlations between the sample and various classes, along with its learning difficulty [Zhou and Wu, 2023; Lin *et al.*, 2024; Zhou *et al.*, 2023a]. These characteristics are then input into the strategy network to facilitate the generation of sample-specific augmentation strategies. The characteristics that reflect the sample-class correlation include the prediction distribution and the cosine similarity between the sample feature and the classifier weights for each class. Each of these two characteristics has $\mathcal{C}$ dimensions. Furthermore, to gauge the learning difficulty of samples comprehensively, we consider metrics including loss, loss gradient, uncertainty, margin, forgetfulness, and class proportion, all of which are scalar values. As a result, a characteristic vector $\boldsymbol{\xi}_i \in \mathbb{R}^{2\times\mathcal{C}+6}$ is derived for each sample at each iteration.

Optimization Process

We introduce a reinforcement learning-based algorithm for the alternating optimization of parameters in both the target ($\boldsymbol{\theta}$) and strategy ($\boldsymbol{\Omega}$) networks. With $\boldsymbol{\Omega}$ fixed, the optimization

Dataset	CelebA		CMNIST		Waterbirds		CivilComments	
Metric	Avg. (↑)	Worst (↑)	Avg. (↑)	Worst (↑)	Avg. (↑)	Worst (↑)	Avg. (↑)	Worst (↑)
ERM	**94.88**	47.76	27.79	0.11	**97.00**	63.73	<u>92.22</u>	56.13
CORAL [Li *et al.*, 2018]	93.82	76.91	71.79	69.48	90.32	79.83	88.68	65.57
IRM [Arjovsky *et al.*, 2019]	94.01	77.82	72.07	70.33	87.45	75.64	88.82	66.30
GroupDRO [Sagawa *et al.*, 2020]	92.11	87.23	72.29	68.58	91.76	<u>90.62</u>	89.94	70.01
DomainMix [Xu *et al.*, 2020]	93.44	65.59	51.41	48.07	76.45	53.11	90.87	63.62
IB-IRM [Ahuja *et al.*, 2021]	93.62	85.03	72.25	70.73	88.52	76.51	89.14	65.38
V-REx [Krueger *et al.*, 2021]	92.24	86.79	71.77	70.25	88.03	73.61	90.22	64.90
LISA [Yao *et al.*, 2022]	92.44	89.36	74.08	73.36	91.84	89.28	89.20	72.63
CNC [Zhang *et al.*, 2022]	89.92	88.86	<u>90.88</u>	<u>77.25</u>	90.81	88.53	81.73	68.78
C-GAN [Reddy *et al.*, 2023]	93.04	87.62	78.95	75.66	93.57	89.84	-	-
ACE [Singla *et al.*, 2023]	92.02	87.33	55.41	50.54	90.76	75.09	-	-
DISC [Wu *et al.*, 2023]	94.13	89.75	76.38	74.52	93.49	89.67	-	-
PDE [Deng *et al.*, 2024]	92.05	<u>91.07</u>	78.08	75.92	92.41	90.53	86.87	<u>72.69</u>
DKT (Ours)	<u>94.57</u>	**92.34**	**91.45**	**78.36**	<u>95.50</u>	**92.56**	**92.30**	**76.02**

Table 1: Comparison of average and worst-group accuracy (%) across four subpopulation shift datasets. (↑) indicates that higher values are better. The best and second-best results are highlighted in bold and underlined. DKT consistently achieves the highest worst-group accuracy, indicating its efficacy in breaking spurious correlations.

subproblem for the target network can be defined as

$$\min_{\boldsymbol{\theta}} \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \mathrm{E}_{\mathcal{A}\sim p(\mathcal{A}|\boldsymbol{\xi};\boldsymbol{\Omega})} \left[\mathcal{L}^{DKT}(\boldsymbol{\theta},\boldsymbol{\Omega})\right]. \tag{10}$$

Given the characteristics of a sample, the strategy network generates a strategy distribution $p(\mathcal{A}_i|\boldsymbol{\xi}_i;\boldsymbol{\Omega})$, from which an augmentation strategy is randomly sampled. Utilizing these sampled strategies, we update the parameters of the target model by minimizing $\mathcal{L}^{DKT}$ via gradient descent:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \eta_1 \frac{1}{n} \sum\nolimits_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \mathcal{L}_i^{DKT}, \tag{11}$$

where η_1 represents the learning rate of the target network and n denotes the mini-batch size.

Additionally, with $\boldsymbol{\theta}$ given, the optimization problem for the parameters of the strategy network can be formulated as

$$\min_{\boldsymbol{\Omega}} \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \mathrm{E}_{\mathcal{A}\sim p(\mathcal{A}|\boldsymbol{\xi};\boldsymbol{\Omega})} \left[\mathcal{L}^{DKT}(\boldsymbol{\theta},\boldsymbol{\Omega})\right]. \tag{12}$$

Following the REINFORCE algorithm [Williams, 1992], we compute the derivative of the objective function, $\mathcal{H}(\boldsymbol{\Omega}) := \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \mathrm{E}_{\mathcal{A}\sim p(\mathcal{A}|\boldsymbol{\xi};\boldsymbol{\Omega})} \mathcal{L}^{DKT}$, with respect to $\boldsymbol{\Omega}$ as:

$$\begin{aligned}
&\nabla_{\boldsymbol{\Omega}} \mathcal{H}(\boldsymbol{\Omega}) = \nabla_{\boldsymbol{\Omega}} \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \mathrm{E}_{\mathcal{A}\sim p(\mathcal{A}|\boldsymbol{\xi};\boldsymbol{\Omega})} \mathcal{L}^{DKT} \\
&= \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \sum_{\mathcal{A}} \mathcal{L}^{DKT} \cdot \nabla_{\boldsymbol{\Omega}} p(\mathcal{A} \mid \boldsymbol{\xi};\boldsymbol{\Omega}) d\mathcal{A} \\
&= \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \sum_{\mathcal{A}} \mathcal{L}^{DKT} \cdot p(\mathcal{A} \mid \boldsymbol{\xi};\boldsymbol{\Omega}) \nabla_{\boldsymbol{\Omega}} \log p(\mathcal{A} \mid \boldsymbol{\xi};\boldsymbol{\Omega}) d\mathcal{A} \\
&= \mathrm{E}_{(\boldsymbol{x},y)\sim\mathcal{D}^{tr}} \mathrm{E}_{\mathcal{A}\sim p(\mathcal{A}|\boldsymbol{\xi};\boldsymbol{\Omega})} \left[\mathcal{L}^{DKT} \cdot \nabla_{\boldsymbol{\Omega}} \log p(\mathcal{A} \mid \boldsymbol{\xi};\boldsymbol{\Omega})\right].
\end{aligned} \tag{13}$$

Similar to solving Eq. (10), we sample augmentation strategies from the conditional distribution to calculate the sample losses. The gradient with respect to the parameters of the strategy network can be approximately computed as:

$$\nabla_{\boldsymbol{\Omega}} \mathcal{H}(\boldsymbol{\Omega}) \approx \frac{1}{n} \sum\nolimits_{i=1}^{n} \mathcal{L}_i^{DKT} \cdot \nabla_{\boldsymbol{\Omega}} \log p\left(\mathcal{A}_i \mid \boldsymbol{\xi}_i;\boldsymbol{\Omega}\right). \tag{14}$$

Consequently, the parameters of the strategy network can be updated using gradient descent as follows:

$$\boldsymbol{\Omega}^{t+1} = \boldsymbol{\Omega}^{t} - \eta_2 \nabla_{\boldsymbol{\Omega}} \mathcal{H}\left(\boldsymbol{\Omega}^{t}\right), \tag{15}$$

where η_2 denotes the learning rate for the strategy network.

Convergence Analysis

Based on Eqs. (11) and (15), we have the following convergence result for the effectiveness of the proposed optimization algorithm.

Theorem 1. *Suppose that the objective function for the strategy network $\mathcal{L}^{DKT}$ satisfies the gradient Lipschitz conditions w.r.t. $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$, and $\mathcal{L}^{DKT}$ is λ-strongly concave in $\hat{\boldsymbol{\Omega}}$, the feasible set of $\boldsymbol{\Omega}$. If $\mathcal{A}'$ is a δ-approximation of the optimal augmentation strategy $\mathcal{A}^*$, the variance of the stochastic gradient is bounded by a constant $\sigma^2 > 0$, and we set the learning rate of $\boldsymbol{\theta}$ as*

$$\eta_1 = \min\left(\frac{1}{L_{DKT}}, \sqrt{\frac{\Gamma}{\sigma^2 T L_{DKT}}}\right), \tag{16}$$

where $L_{DKT} = L_{\boldsymbol{\theta}\boldsymbol{\Omega}} L_{\boldsymbol{\Omega}\boldsymbol{\theta}}/\lambda + L_{\boldsymbol{\theta}\boldsymbol{\theta}}$ denotes the Lipschitz constant of $\mathcal{L}^{DKT}$ and $\Gamma = \mathcal{L}^{DKT}\left(\boldsymbol{\theta}^0\right) - \min_{\boldsymbol{\theta}} \mathcal{L}^{DKT}(\boldsymbol{\theta})$, it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathrm{E}\left[\left\|\nabla \mathcal{L}^{DKT}\left(\boldsymbol{\theta}^t\right)\right\|_2^2\right] \leq 4\sigma\sqrt{\frac{\Gamma L_{DKT}}{T}} + \frac{5\delta L_{\boldsymbol{\theta}\boldsymbol{\Omega}}^2}{\lambda}, \tag{17}$$

where T represents the maximum training epoch.

According to Theorem 1, if the inner minimization process yields a δ-approximation of the optimal augmentation strategy $\mathcal{A}^*$, then the training algorithm can converge to a stationary point at a sub-linear rate with an accuracy of $5\delta L_{\boldsymbol{\theta}\boldsymbol{\Omega}}^2/\lambda$. Moreover, if $5\delta L_{\boldsymbol{\theta}\boldsymbol{\Omega}}^2/\lambda$ is sufficiently small, our method effectively identifies the desired robust model $\boldsymbol{\theta}^T$ by generating a good augmentation strategy capable of approximating $\mathcal{A}^*$ well.

4 Experimental Investigation

We conduct experiments on three typical learning scenarios susceptible to spurious correlations: subpopulation shift learning, GLT learning, and domain shift learning. The experiments are repeated three times using different random

Benchmark	ImageNet						MSCOCO	
Protocol	CLT		GLT		ALT		GLT	
Metric	Acc. (↑)	Prec. (↑)	Acc. (↑)	Prec. (↑)	Acc. (↑)	Prec. (↑)	Acc. (↑)	Prec. (↑)
CE loss†	42.52	47.92	34.75	40.65	41.73	41.74	63.79	70.52
Re-balancing techniques								
LDAM† [Cao *et al.*, 2019]	46.74	46.86	38.54	39.08	42.66	41.80	67.26	70.70
De-confound-TDE† [Tang *et al.*, 2020]	45.70	44.48	37.56	37.00	41.40	42.36	66.07	68.20
BLSoftmax† [Ren *et al.*, 2020]	45.79	46.27	37.09	38.08	41.32	41.37	64.07	68.59
BBN† [Zhou *et al.*, 2020]	46.46	49.86	37.91	41.77	43.26	43.86	64.48	70.20
LA [Menon *et al.*, 2021]	46.53	45.56	37.80	37.56	41.73	41.74	66.17	68.35
IFL† [Tang *et al.*, 2022]	45.97	52.06	37.96	44.47	45.89	46.42	65.31	72.24
BKD [Zhang *et al.*, 2023]	46.51	50.15	37.93	41.50	42.17	41.83	65.48	70.59
Augmentation methods								
MixUp† [Zhang *et al.*, 2018]	38.81	45.41	31.55	37.44	42.11	42.42	64.45	71.13
ISDA [Wang *et al.*, 2019]	42.66	44.98	36.44	37.26	43.34	43.56	66.57	71.09
RandAug† [Cubuk *et al.*, 2020]	46.40	52.13	38.24	44.74	46.29	46.32	67.71	72.73
MetaSAug [Li *et al.*, 2021]	48.47	54.09	38.31	43.24	43.15	43.50	65.89	71.91
Meta-IADA [Zhou *et al.*, 2024]	53.45	58.05	44.36	50.07	52.54	53.23	70.06	74.55
Ensemble learning approaches								
RIDE† [Wang *et al.*, 2020]	52.08	51.65	43.00	43.32	47.24	46.67	68.59	72.20
TADE† [Zhang *et al.*, 2021]	50.47	51.85	41.75	44.15	47.10	47.32	66.98	71.22
DKT (Ours)	**54.87**	**59.49**	**45.64**	**52.10**	**54.32**	**54.71**	**72.36**	**75.69**

Table 2: Accuracy and precision (%) of the CLT, GLT, and ALT protocols on the ImageNet-GLT and MOCOCO-GLT benchmarks. † indicates the results reported in [Tang *et al.*, 2022]. DKT surpasses all compared baselines in terms of both accuracy and precision.

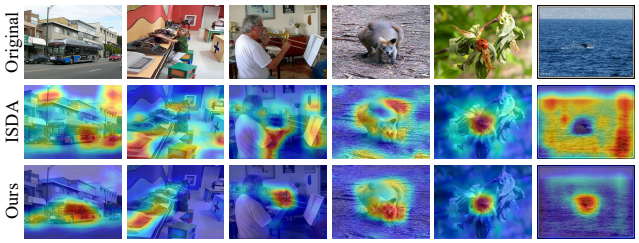

Figure 3: Visualization of the input regions utilized by the model for predictions. Blue and red indicate regions that are non-discriminative and highly discriminative, respectively.

seeds. For all experiments, the parameter sets for α_i, β_i, and γ_i range from 0.1 to 1, with intervals of 0.1.

4.1 Subpopulation Shift Learning

Settings. We evaluate the performance of DKT under four subpopulation shift datasets: Colored MNIST (CMNIST) [Yao *et al.*, 2022], Waterbirds [Sagawa *et al.*, 2020], CelebA [Liu *et al.*, 2016], and CivilComments [Borkan *et al.*, 2019]. In these datasets, certain attributes are highly spuriously correlated with the labels. Following Yao et al. [2022], we adopt pre-trained ResNet-50 [He *et al.*, 2016] and DistilBERT [Sanh *et al.*, 2019] as the model for image (i.e., CMNIST, Waterbirds, CelebA) and text data (i.e., CivilComments), respectively. We compare DKT with a variety of robust methods designed to break spurious correlations and learn invariant predictors. To ensure a comprehensive assessment, we report both the average and worst-group accuracy.

Results. The comparative results are reported in Table 1. DKT achieves a higher average accuracy compared to other robust learning methods across diverse datasets, highlighting its strong generalization capability. Moreover, it consistently surpasses other methods in terms of worst-group accuracy, demonstrating its efficacy in enhancing model robustness for underrepresented groups, such as samples belonging to the landbird class with a water background. These findings manifest that models trained using the DKT framework rely more on causal attributes for predictions, rather than non-causal factors like colors and backgrounds, thereby bolstering the model's resilience to spurious correlations. Furthermore, data modification methods, such as PDE, and our approach, generally outperform previous regularization techniques (e.g., IRM and LISA), underscoring the effectiveness of data expansion in mitigating spurious correlations.

4.2 Generalized Long-Tail Learning

Settings. GLT learning accounts for both long-tailed class and attribute distributions within the training data, as both types of imbalances contribute to spurious correlations. We employ two GLT benchmarks [Tang *et al.*, 2022]: ImageNet-GLT and MSCOCO-GLT. Each benchmark consists of three protocols: Class-wise Long Tail (CLT), Attribute-wise Long Tail (ALT), and GLT, showcasing variations in class distribution, attribute distribution, and combinations of both between training and testing datasets. We report the mean accuracy and precision for all approaches. Following Tang et al. [2022], the ResNeXt-50 [Xie *et al.*, 2017] model serves as the backbone network. We compare DKT with three cate-

Dataset Metric	Camelyon17 Avg. (↑)	FMoW Worst (↑)	RxRx1 Avg. (↑)
ERM	70.32	32.29	29.91
CORAL [Li *et al.*, 2018]	59.53	31.74	28.45
IRM [Arjovsky *et al.*, 2019]	64.20	30.04	8.23
GroupDRO [Sagawa *et al.*, 2020]	68.45	30.82	23.07
DomainMix [Xu *et al.*, 2020]	69.71	34.23	30.86
IB-IRM [Ahuja *et al.*, 2021]	68.90	28.41	6.44
V-REx [Krueger *et al.*, 2021]	71.52	27.26	7.55
LISA [Yao *et al.*, 2022]	77.15	35.52	31.90
C-GAN [Reddy *et al.*, 2023]	68.74	32.41	27.86
ACE [Singla *et al.*, 2023]	67.53	30.88	26.95
PDE [Deng *et al.*, 2024]	75.72	35.91	31.92
DKT (Ours)	**78.94**	**37.35**	**32.43**

Table 3: Comparison results (%) on three domain shift datasets. The performance of DKT surpasses that of all other compared methods across the various domain shift datasets.

gories of approaches: re-balancing techniques, augmentation methods, and ensemble learning approaches.

Results. The comparative results are reported in Table 2, where DKT demonstrates significant performance improvements, with its accuracy and precision surpassing the best comparative results by 1.70% and 1.52%, respectively. These findings underscore its effectiveness in enhancing model generalization and robustness against imbalanced class and attribute distributions. Moreover, the results presented in Fig. 3 further illustrate its efficacy in mitigating spurious associations between non-causal attributes and class labels, thereby guiding the model to focus more on causal attributes. Additionally, methods tailored for long-tailed learning generally exhibit inferior performance on the GLT and ALT protocols, primarily due to their class-level characteristics. Nevertheless, DKT dynamically generates sample-specific augmentation strategies, making it more effective in breaking sample-wise spurious correlations.

4.3 Domain Shift Learning

Settings. We examine three domain shift benchmarks featuring out-of-distribution test data. These benchmarks (i.e., Camelyon17 [Bandi *et al.*, 2018], FMoW [Christie *et al.*, 2018], and RxRx1 [Taylor *et al.*, 2019]) are sourced from WILDS [Koh *et al.*, 2021], covering domains including healthcare and vision. Following previous studies [Yao *et al.*, 2022], the evaluation metrics are average accuracy for Camelyon17 and RxRx1, and worst-group accuracy for FMoW. The compared baselines are consistent with those for subpopulation shift learning.

Results. The comparison results for the three datasets are presented in Table 3. DKT consistently outperforms other compared methods across diverse datasets, achieving an average accuracy improvement of 1.25%. Moreover, we have validated the significance of the performance improvement achieved by DKT using the Wilcoxon signed-rank test, which produced a p-value of 0.04, smaller than the threshold of 0.05. These findings demonstrate the efficacy of DKT in enhancing model robustness against spurious correlations.

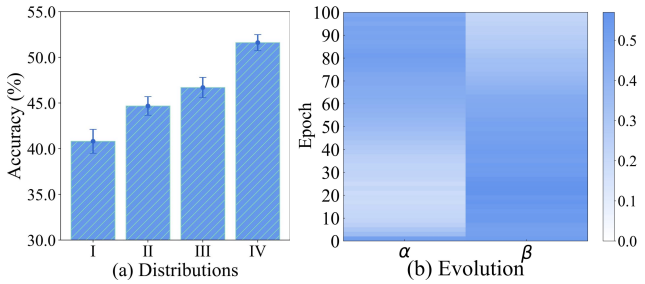

Figure 4: (a) Ablation studies of augmentation distributions. (b) Evolution of the average α and β values during the training process. The ImageNet-GLT benchmark is utilized.

4.4 Analytical Experiments

We compare the performance of DKT under four augmentation distributions: (I) excluding knowledge transfer among categories, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{y_i})$, (II) excluding the augmentation of samples within their own classes, $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{\Sigma}_c)$, (III) adopting the augmentation distribution with a zero mean, $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_c + \mathbf{\Sigma}_{y_i})$, and (IV) our adopted augmentation distribution, $\mathcal{N}(\boldsymbol{\mu}_c, \mathbf{\Sigma}_c + \mathbf{\Sigma}_{y_i})$. The accuracy across all protocols is presented in Fig. 4(a). Our augmentation distribution proves to be the most effective, due to its ability to enhance intra-class diversity and reduce inter-class spurious associations. Additionally, the evolution of the average α and β values during training is illustrated in Fig. 4(b). At the start of training, both the average values of α and β are approximately equal to 0.5, as the selected values encompass the entire range of possible options. During the early stages of training, the average value of β surpasses that of α, prioritizing feature representation. In contrast, later in training, the average value of β drops below that of α, enhancing the model's robustness to spurious correlations.

5 Conclusion

This paper proposes enhancing model robustness against spurious correlations by transferring knowledge across categories. Specifically, the deep features of the samples are transformed along semantic vectors derived from their respective category distributions, as well as those of spuriously correlated categories. We then demonstrate that our knowledge transfer strategy can be achieved solely by optimizing a surrogate robust loss and introduce a reinforcement learning-based framework to train classifiers using this loss, where the direction and intensity of knowledge transfer are dynamically determined based on the unique training dynamics of samples. Extensive experiments across various learning scenarios prone to spurious correlations validate the effectiveness of our approach in mitigating spurious correlations and enhancing model robustness.

Acknowledgments

This work was partially supported by the Baidu Scholarship.

References

[Ahuja *et al.*, 2021] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, et al. Invariance principle meets information bottleneck for out-of-distribution generalization. In *NeurIPS*, pages 3438–3450, 2021.

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

[Bandi *et al.*, 2018] Peter Bandi, Oscar Geessink, Quirine Manson, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018.

[Borkan *et al.*, 2019] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, pages 491–500, 2019.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578, 2019.

[Chew *et al.*, 2024] Oscar Chew, Hsuan-Tien Lin, Kai-Wei Chang, et al. Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In *ACL Findings*, pages 1013–1025, 2024.

[Christie *et al.*, 2018] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, pages 6172–6180, 2018.

[Cubuk *et al.*, 2020] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 3008–3017, 2020.

[Deng *et al.*, 2023] Xun Deng, Wenjie Wang, Fuli Feng, Hanwang Zhang, Xiangnan He, and Yong Liao. Counterfactual active learning for out-of-distribution generalization. In *ACL*, pages 11362–11377, 2023.

[Deng *et al.*, 2024] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. In *NeurIPS*, pages 1390–1402, 2024.

[Gupta *et al.*, 2022] Umang Gupta, Jwala Dhamala, Varun Kumar, et al. Mitigating gender bias in distilled language models via counterfactual role reversal. In *ACL*, page 658–678, 2022.

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

[Koh *et al.*, 2021] 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.

[Krueger *et al.*, 2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, pages 5815–5826, 2021.

[Li *et al.*, 2018] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[Li *et al.*, 2021] Shuang Li, Kaixiong Gong, Chi-Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, pages 5208–5217, 2021.

[Lin *et al.*, 2024] Zhihao Lin, Qi Zhang, Zhen Tian, Peizhuo Yu, and Jianglin Lan. Dpl-slam: enhancing dynamic point-line slam through dense semantic methods. *IEEE Sensors Journal*, 2024.

[Liu *et al.*, 2016] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2016.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[Moayeri *et al.*, 2022] Mazda Moayeri, Phillip Pope, Yogesh Balaji, et al. A comprehensive study of image classification model sensitivity to foregrounds, backgrounds, and visual attributes. In *CVPR*, pages 19087–19097, 2022.

[Reddy *et al.*, 2023] Abbavaram Gowtham Reddy, Saketh Bachu, Saloni Dash, et al. On counterfactual data augmentation under confounding. *arXiv preprint arXiv:2305.18183*, 2023.

[Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, pages 4175–4186, 2020.

[Sagawa *et al.*, 2020] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Selvaraju *et al.*, 2017] 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.

[Singla *et al.*, 2023] Sumedha Singla, Nihal Murali, Forough Arabshahi, et al. Augmentation by counterfactual explanation-fixing an overconfident classifier. In *WACV*, pages 4720–4730, 2023.

[Sun *et al.*, 2021] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recover-

ing latent causal factor for generalization to distributional shifts. In *NeurIPS*, pages 16846–16859, 2021.

[Tang *et al.*, 2020] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, pages 1513–1524, 2020.

[Tang *et al.*, 2022] Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. Invariant feature learning for generalized long-tailed classification. In *ECCV*, pages 709–726, 2022.

[Taylor *et al.*, 2019] James Taylor, Berton Earnshaw, Ben Mabey, Mason Victors, and Jason Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *ICLR Workshops*, 2019.

[Tian *et al.*, 2025] Zhen Tian, Zhihao Lin, Dezong Zhao, Wenjing Zhao, David Flynn, Shuja Ansari, and Chongfeng Wei. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886*, 2025.

[Veitch *et al.*, 2021] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *NeurIPS*, pages 16196–16208, 2021.

[Wang *et al.*, 2019] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. In *NeurIPS*, pages 12635–12644, 2019.

[Wang *et al.*, 2020] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2020.

[Wang *et al.*, 2025] Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. Abkd: Pursuing a proper allocation of the probability mass in knowledge distillation via alpha-beta-divergence. *arXiv preprint arXiv:2505.04560*, 2025.

[Williams, 1992] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[Wu *et al.*, 2023] Shirley Wu, Mert Yuksekgonul, et al. Discover and cure: Concept-aware mitigation of spurious correlation. In *ICML*, pages 37765–37786, 2023.

[Wu *et al.*, 2024] Dongming Wu, Lulu Wen, Chao Chen, and Zhaoshu Shi. A novel counterfactual data augmentation method for aspect-based sentiment analysis. In *ACML*, pages 1479–1493, 2024.

[Xiao *et al.*, 2023] Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, et al. Masked images are counterfactual samples for robust fine-tuning. In *CVPR*, pages 20301–20310, 2023.

[Xie *et al.*, 2017] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017.

[Xu *et al.*, 2020] Minghao Xu, Jian Zhang, Bingbing Ni, et al. Adversarial domain adaptation with domain mixup. In *AAAI*, pages 6502–6509, 2020.

[Yao *et al.*, 2022] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, pages 25407–25437, 2022.

[Young *et al.*, 2019] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, et al. Deep neural network or dermatologist? In *IMIMIC*, pages 48–55, 2019.

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

[Zhang *et al.*, 2021] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021.

[Zhang *et al.*, 2022] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *ICML*, pages 26484–26516, 2022.

[Zhang *et al.*, 2023] Shaoyu Zhang, Chen Chen, Xiyuan Hu, and Silong Peng. Balanced knowledge distillation for long-tailed learning. *Neurocomputing*, 527:36–46, 2023.

[Zhao *et al.*, 2023] Shiman Zhao, Wei Chen, and Tengjiao Wang. Learning few-shot sample-set operations for noisy multi-label aspect category detection. In *IJCAI*, pages 5306–5313, 2023.

[Zhao *et al.*, 2024] Shiman Zhao, Yutao Xie, Wei Chen, Tengjiao Wang, Jiahui Yao, et al. Metric-free learning network with dual relations propagation for few-shot aspect category sentiment analysis. *TACL*, 12:100–119, 2024.

[Zhao *et al.*, 2025] Shiman Zhao, Wei Chen, Tengjiao Wang, Jiahui Yao, Dawei Lu, and Jiabin Zheng. Less is enough: Relation graph guided few-shot learning for multi-label aspect category detection. In *ICASSP*, pages 1–5, 2025.

[Zhou and Wu, 2023] Xiaoling Zhou and Ou Wu. Implicit counterfactual data augmentation for deep neural networks. *arXiv preprint arXiv:2304.13431*, 2023.

[Zhou *et al.*, 2020] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020.

[Zhou *et al.*, 2023a] Xiaoling Zhou, Ou Wu, and Mengyang Li. Investigating the sample weighting mechanism using an interpretable weighting framework. *IEEE TKDE*, 36(5):2041–2055, 2023.

[Zhou *et al.*, 2023b] Xiaoling Zhou, Nan Yang, and Ou Wu. Combining adversaries with anti-adversaries in training. In *AAAI*, pages 11435–11442, 2023.

[Zhou *et al.*, 2024] Xiaoling Zhou, Wei Ye, Zhemg Lee, Rui Xie, and Shikun Zhang. Boosting model resilience via implicit adversarial data augmentation. In *IJCAI*, pages 5653–5661, 2024.