A Unified Framework for Robustness on Diverse Sampling Errors

Myeongho Jeon Myungjoo Kang Joonseok Lee
Seoul National University
{andyjeon, mkang, joonseok}@snu.ac.kr

Abstract

Recent studies have substantiated that machine learning algorithms including convolutional neural networks often suffer from unreliable generalizations when there is a significant gap between the source and target data distributions. To mitigate this issue, a predetermined distribution shift has been addressed independently (e.g., single domain generalization, de-biasing). However, a distribution mismatch cannot be clearly estimated because the target distribution is unknown at training. Therefore, a conservative approach robust on unexpected diverse distributions is more desirable in practice. Our work starts from a motivation to allow adaptive inference once we know the target, since it is accessible only at testing. Instead of assuming and fixing the target distribution at training, our proposed approach allows adjusting the feature space the model refers to at every prediction, i.e., instance-wise adaptive inference. The extensive evaluation demonstrates our method is effective for generalization on diverse distributions.

1. Introduction

A fundamental assumption of machine learning is that the training dataset represents the true distribution and the test set also follows the same true distribution. For instance, a face recognition model is trained on a dataset of thousands or millions of face images, assuming that this dataset represents a set of all possible faces of human beings under all conditions like illumination or angle. In this sense, any dataset is nothing but a collection of samples from the real world, and it may not perfectly represent the true distribution, even if it is huge. No dataset is free from this sampling error; rather it is just a matter of degree. Obviously, a model trained on a particular set of samples is affected by the distribution. In some fortunate cases, we may have similar distribution over training samples and over inference set, while distribution shift between source and target data caused by extreme sampling error could give rise to unreliable generalization of the model.

Although we are not able to know the true distribution in most cases, there are many practical situations where we need to build a model based only on a limited set of training samples and apply it to another target data, which may neither necessarily follow the training distribution, nor the true distribution. For instance, a gender classifier trained only on images of the young may not perform well for the old. For this practical need, previous research has been dedicated to tackling this issue of domain shift. Particularly, several works suggest that the performance of the model trained on the source domain could often be degraded on the out-of-distribution (OOD) target domains. This is called *single domain generalization* (SDG) [25, 22, 32, 7, 20, 30, 26, 5], mainly tackling an unknown domain shift between the training and test sets.

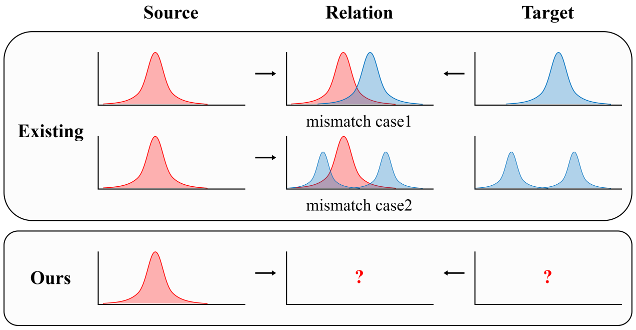


Figure 1. **Our motivation.** At training, we do not know to which distribution the model would be deployed. Thus, generalization for a predetermined distribution mismatch (*e.g.*, SDG or UBL) is often unfavorable. The image at the bottom illustrates our scenario, where an arbitrary distribution could be the target.

On the other hand, *unbiased learning* (UBL) is another related research area, aiming at removing bias that potentially a machine learning model may rely on. As an example, a gender classifier may consider 'age' as an important factor if trained on (young, female) and (old, male) samples and hence frequently fail to correctly predict (old, female) and (young, male) test samples. Biased attributes could make the model biased and several recent works propose various approaches to mitigate this issue [3, 21, 19, 11, 19, 13, 23].

However, the type of distribution shift can be defined only when we know both the source and target distributions. As illustrated in Fig. 1, we do not know at training on which

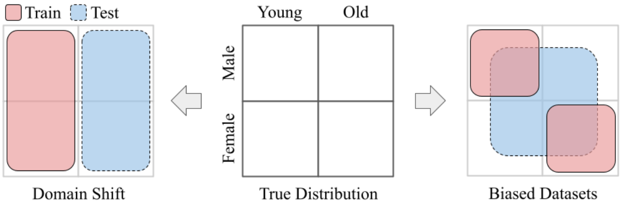


Figure 2. **Distribution mismatch.** Domain generalization and unbiased learning can be seen as two special cases of the distribution mismatch problem.

distribution the model will be used for inference. We hypothesize that conventional methods may perform well or not depending on the distribution shift because they are designed considering one predetermined scenario.

From this perspective, we explore the generalizability of the previous methods on two special cases of distribution shift problems (*i.e.*, biased and domain-limited distributions depicted in Fig. 2) by conducting a motivating experiment to apply state-of-the-art unbiased learning (UBL) models to single domain generalization (SDG) and vice versa. Not surprisingly, we observe they significantly underperform beyond the designated situations they were trained on (See Sec. 3 for more details). In other words, previous methods tackling either SDG or UBL tend to be over-optimized to each specific case, failing to generalize to the other problem, even if they are indeed the same problem, *i.e.*, domain mismatch. This quantitatively demonstrates that inaccurate estimation of data distribution and inappropriate model selection during training could significantly drop performance.

In this paper, we aim to design a framework robust on diverse data distribution mismatch problems, especially focusing on convolutional neural networks (CNNs) for image recognition. CNNs learn two manuals: learning to represent a feature space from input images and fitting a classifier from the features to the target labels. From this point of view, we hypothesize that the features helpful for image discrimination would be different depending on the data. This would be because image data distribution implies correlations of the attributes in the image space including the label, which can be used as a predictive logic to match the feature space to the label space. In this respect, the model optimized to represent features and select appropriate ones only for a specific source data may not be directly applied to OOD data.

Therefore, we propose *instance-wise adaptive inference* (IAI) that adjusts the referenced feature space to be more suitable for each test instance at inference. Although we do not know the type of distribution mismatch at training, it can be indirectly estimated by comparing each target sample with the learned representations. Specifically, our operative idea is threefold: (*i*) to widen the feature space (since if the model has a limited feature space, there may be little opportunity to apply IAI), (*ii*) to disentangle the widened feature space, and (*iii*) to adaptively select features among them, considering each test instance at inference.

In consequence, we demonstrate that our proposed method exhibits remarkable generalizability on both SDG and UBL tasks compared to the state-of-the-art methods. Our experimental results quantitatively verify our hypothesis that target data needs different features from source data for better image discrimination, justifying the effectiveness of IAI for robustness.

We summarize our main contributions as follows:

- We present a novel framework for robust learning on arbitrary diverse distributions. This is desirable in scenarios where a model is deployed in a dynamic environment where queries often drift while the model cannot be frequently revised.
- We propose a novel method that exploits *adaptive inference* re-weighting instance-wise on disentangled representations in widened feature space. Extensive evaluations demonstrate our method is robust on diverse distributions.

2. Problem Statement

Consider an instance set $X = \{\mathbf{x}_i | \mathbf{x}_i \in \mathcal{X}, i = 1, \dots, N\}$ for a classification problem, where $\mathcal{X}$ denotes the input space. The instance set X is generated by collecting N samples $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ from the real world, and this process is not free from the sampling error, resulting in distribution shift, as mentioned at the beginning. In machine learning, a model may not suffer from generalization problem when distribution shift is subtle. However, if the model is trained on a dataset with severe sampling error, its performance could be significantly degraded at inference.

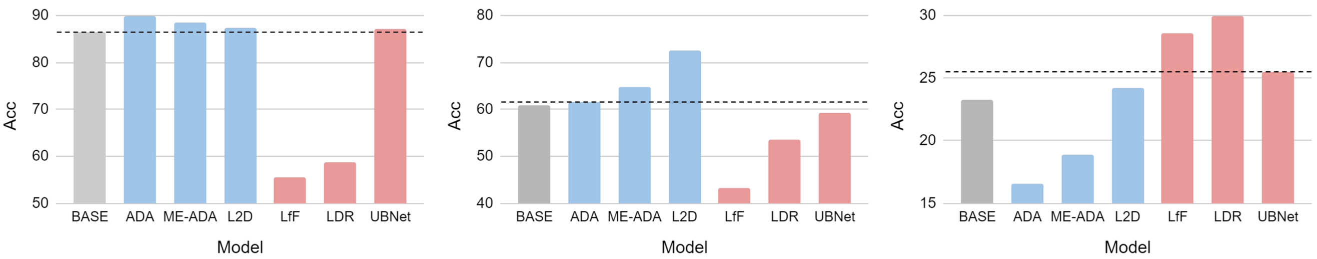
Our problem setting. Let us denote the true, training (source), and testing (target) distributions by p, p_S, and p_T, respectively. Also, we denote the source and target datasets by $\mathcal{D}_S \sim p_S$ and $\mathcal{D}_T \sim p_T$, respectively, composed of samples from the corresponding distributions. A machine learning model $\hat{f}$ is trained to minimize the empirical risk

$$\hat{f} = \underset{f}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y_i, f(\mathbf{x}_i)), \tag{1}$$

where $(\mathbf{x}_i, y_i) \in \mathcal{D}_S$ for $i = 1, ..., N$. The ideal goal of our problem is maximizing its performance on true distribution $p(\mathbf{x}, y)$; that is,

$$f^* = \underset{\hat{f}}{\text{argmin}} \, \mathbb{E}_{p(\mathbf{x},y)} \left[\mathcal{L}(y, \hat{f}(\mathbf{x})) \right]. \tag{2}$$

However, since constructing $\mathcal{D}_T$ with $p_T \simeq p$ is impractical, we aim to make the model f robust on diverse $\{\mathcal{D}_S, \mathcal{D}_T\}$ pairs where there is significant distribution mismatch between them, denoted by $\mathcal{D}_S \not\approx \mathcal{D}_T$.



(a) (b) (c)

Figure 3. **Exploration of previous methods on diverse distributions.** (a) no-shift distribution, (b) domain-limited distribution, (c) biased distribution. Blue bars in the graphs mean SDG models and red ones are UB models. Dotted lines are the lowest performance among the corresponding task models to compare the methods of other tasks, meaning the accuracy of `BASE`, `ADA`, and `UBNet` for (a), (b), and (c) respectively.

Domain-limited distribution. A domain $\mathcal{O}$ is defined as a tuple of the input space $\mathcal{X}$ and a marginal distribution p_X over samples $X \subset \mathcal{X}$, *i.e.*, $\mathcal{O} = (\mathcal{X}, p_X)$. The source distribution p_S is called *domain-limited* when the source domain $\mathcal{O}_S$ is significantly different from the target domain $\mathcal{O}_T$ with $\mathcal{X}_S \neq \mathcal{X}_T, p_S \neq p_T$. In this sense, the domain-limited distribution is a special case of $\mathcal{D}_S \not\approx \mathcal{D}_T$.

Biased distribution. Biased distribution is another special case of $\mathcal{D}_S \not\approx \mathcal{D}_T$. The label space $\mathcal{Y}$ can be defined as a set of all possible assignments to each instance $\mathbf{x}$. For $(\mathbf{x}_S, y_S) \in \mathcal{D}_S$, y_S is a particular one selected from $\mathcal{Y}$, *e.g.*, gender of the person in an image. Most $y' \neq y_S \in \mathcal{Y}$ may be independent of the y_S, but some might be significantly correlated with y_S [13]. For the latter y', if it also has a correlation with labels y_T of the target dataset $\mathcal{D}_T$, y' can be used as a meaningful factor for prediction. For instance, the mustache is a meaningful indicator that the person is likely a male. However, if y' is spuriously correlated with the label y_S, relying on y' would cause $\mathcal{D}_S \not\approx \mathcal{D}_T$, misleading the generalization of the model, considered as a bias. For example, hair length is a well-known bias for gender classification due to its high correlation with the label, although it is biologically independent.

3. Motivating Experiments

Experimental setup. We explore the performance of the state-of-the-art SDG and UBL models on three distributions: no-shift ($p_S \simeq p_T$), domain-limited, and biased one. We set `ADA` [25], `ME-ADA` [32], and `L2D` [30] for SDG as competing models and `LfF` [21], `LDR` [19] and `UBNet` [13] for UBL. The experiments are conducted on CIFAR-10 [17]. We assume the training and test sets in CIFAR-10 have no-shift distribution, *i.e.*, $p_S \simeq p_T$. The domain-limited and biased CIFAR-10 are made by widening the target domain and by intentionally planting bias, respectively, following the setup in `LfF` [21]. Specifically, one of the 10 types of corruption, {*fog, snow, frost, brightness, contrast, spatter, elastic, jpeg compression, pixelate, saturate*}, is applied to the training images per each label; *e.g.*, (airplane, snow), (automobile, frost), and so on. The unbiased validation set is corrupted uniformly randomly for all the labels. For SDG, 12 validation sets are created with {*fog, snow, frost, zoom blur, defocus blur, glass blur, speckle noise, shot noise, impulse noise, jpeg compression, pixelate, spatter*}. That is, all the validation image samples are corrupted by ‘snow’, meaning the ‘snow’ target domain. We use ResNet18 [10] as the baseline model.

Results and Analysis. Figure 3 shows that all the approaches for SDG and `UBNet` outperform the base model on the no-shift distribution. Yet, `LfF` and `LDR` degrade accuracy significantly. All the UBL models generalize worse than SDG models for SDG task, and vice versa. Consequently, we observe that conventional methods are limited to addressing various sampling errors. This result implies that most previous works perform well only on a special case of distribution mismatch problems, where `L2D` is the only one that performs reasonably on most distributions according to our experiments.

4. Methodology

For generalization on diverse distributions, we propose three ideas: (*i*) widening the feature space to be referred, (*ii*) disentangling multiple independent representations from one another in the widened feature space, and (*iii*) performing inference adaptively to each test example, weighting differently on disentangled features based on distribution mismatch between the source and target data. Towards this, the overall architecture of our method consists of the style generation module G and the classifier $[F = \{F_a, F_b\}; H]$ (See Fig. 4). Aiming to train $[F; H]$ to be robust, G aids it to be generalized on diverse distributions by creating diversely stylized inputs. Our detailed model designs are depicted in the following subsections.

4.1. Background to Widen Feature Space: L2D

Widening the Feature Space. In our framework, the disentangled features are re-weighted for every instance at test time. This approach works well when disentangled features

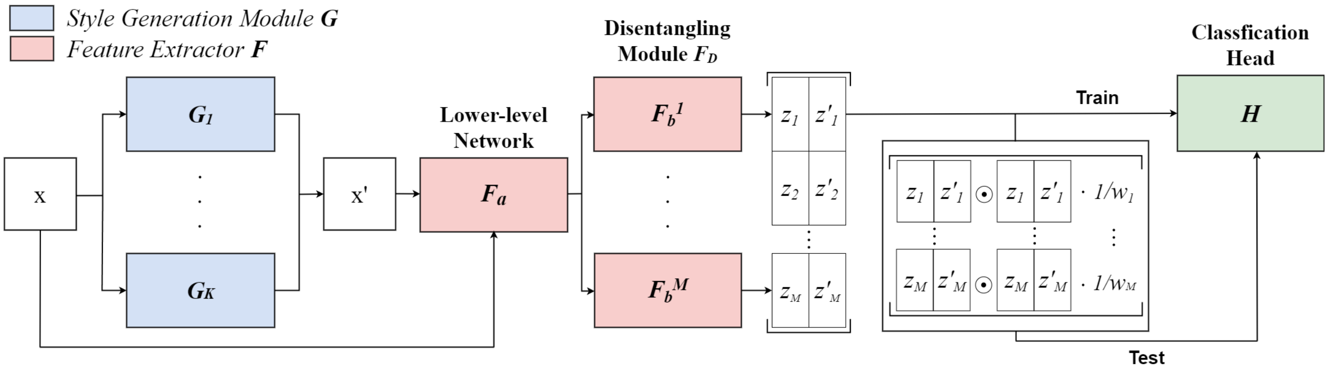


Figure 4. **Our model architecture.** An input image $\mathbf{x}$ and its stylized image $\mathbf{x}'$ by $\{G_i\}_{i=1}^{M}$ are fed into the lower-level network F_a. Then, the higher-level networks $\{F_b^i\}_{i=1}^{M}$ disentangle the features into $\mathbf{z}_1, \cdots, \mathbf{z}_M$. At training, the simply concatenated features $[\mathbf{z}_1, \cdots, \mathbf{z}_M]$ are passed to H, the classification head. At testing, the re-weighted (by $\mathbf{w} = [w_1, ..., w_M]$) and activated (by $\odot$) concatenated features are passed.

have rich enough information, because there may be little opportunity to select more meaningful features for a test instance if the model represents limited feature space. We widen the feature space by exploiting techniques, `L2D` [30].

Learning to Diversify. Let us denote the source dataset as $\mathcal{D}_S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. To make the model represent a wider feature space, the style generation module $G : \mathcal{X} \rightarrow \mathcal{X}'$, composed of K sets of {convolution (`Conv`), style-transfer (`Trans`), transposed convolution ($\texttt{Conv}^\top$)}, synthesizes various stylized images, where $\mathcal{X}$ and $\mathcal{X}'$ denote the original and augmented input space, respectively. All scaled inputs, $\mathbf{x}_i'^k = \texttt{Conv}^\top(\texttt{Trans}(\texttt{Conv}(\mathbf{x}_i)))$ for $k = 1, ..., K$, are aggregated to $\mathbf{x}_i'$ by a linear combination with Gaussian random weights. Then, the stylized image $\mathbf{x}_i'$ is fed to the feature extractor F in conjunction with the original image $\mathbf{x}_i$.

To train G to generate diverse stylized $\mathbf{x}'$, mutual information (MI) between $\mathbf{x}$ and $\mathbf{x}'$ is minimized in the high-level feature space $\mathcal{Z}$. For this, feature extractor $F : \mathcal{X} \cup \mathcal{X}' \rightarrow \mathcal{Z}$ encodes $\mathbf{x}$ and $\mathbf{x}'$ to $\mathbf{z}$ and $\mathbf{z}'$, respectively. Formally, the MI is defined as:

$$I(\mathbf{z}, \mathbf{z}') = \mathbb{E}_{p(\mathbf{z}, \mathbf{z}')}\left[\log \frac{p(\mathbf{z}'|\mathbf{z})}{p(\mathbf{z}')}\right]. \quad (3)$$

If $p(\mathbf{z}'|\mathbf{z})$ and $q(\mathbf{z}'|\mathbf{z})$ have similar distribution, $I(\mathbf{z}, \mathbf{z}')$ can be approximated to a tractable upper bound as:

$$\hat{I}(\mathbf{z}, \mathbf{z}') = \frac{1}{N}\sum_{i=1}^{N}[\log q(\mathbf{z}_i'|\mathbf{z}_i) - \frac{1}{N}\sum_{j=1}^{N}\log q(\mathbf{z}_j'|\mathbf{z}_i)], \quad (4)$$

where the variational distribution $q(\mathbf{z}'|\mathbf{z})$ is estimated by a neural network. The difference between $p(\mathbf{z}'|\mathbf{z})$ and $q(\mathbf{z}'|\mathbf{z})$ can be minimized by Kullback-Leibler divergence (KLD). For an implementation, KLD can be reduced by minimizing the negative log-likelihood $\mathcal{L}_{\text{NLL}}$ between $\mathbf{z}$ and $\mathbf{z}'$:

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{N}\sum_{i=1}^{N}\log q(\mathbf{z}_i'|\mathbf{z}_i). \quad (5)$$

Although G generates various stylized $\mathbf{x}'$, it is meaningless if their semantic information is not maintained. To prevent this problem, another loss term $\mathcal{L}_{\text{MMD}}$ based on class-conditional Maximum Mean Discrepancy (MMD) is applied:

$$\mathcal{L}_{\text{MMD}} = \frac{1}{C}\sum_{j=1}^{C}\left\|\frac{1}{n_s^j}\sum_{i=1}^{n_s^j}\phi(\mathbf{z}_i^j) - \frac{1}{n_t^j}\sum_{i=1}^{n_t^j}\phi(\mathbf{z'}_i^j)\right\|^2, \quad (6)$$

where $\mathbf{z}$ and $\mathbf{z}'$ denote the feature vector of $\mathbf{x}$ and $\mathbf{x}'$, respectively. n_s^j and n_t^j are the number of original and augmented images, respectively. C denotes the number of classes, and $\phi(;)$ denotes a Gaussian kernel that represents the distribution in the kernel Hilbert space to compute the difference.

With the aforementioned process, various styles generated by G cover a wider input space and hence help the model to make wider feature space. Yet, the classifier $[F; H]$ needs to maximize the MI between the same semantic labels for better classification. Towards this, a supervised contrastive loss $\mathcal{L}_{\text{CL}}$ [14] is exploited instead of directly maximizing MI:

$$\mathcal{L}_{\text{CL}} = -\sum_{i=0}^{N}\frac{1}{|P(i)|}\sum_{p \in P(i)}\log\frac{e^{z_i \cdot z_p/\tau}}{\sum_{a \in A(i)} e^{z_i \cdot z_a/\tau}}, \quad (7)$$

where $P(i) = \{\mathbf{z}_p, \mathbf{z}_p' \in A(i) : y_p = y_i\}$, $A(i)$ is the set of the source and generated latent representations $\mathbf{z}, \mathbf{z}'$ of all images in the same class, and $1 \leq i, p \leq N$ denotes the index of the instance of $\mathbf{z}$. τ denotes the temperature coefficient. Consequently, the min-max adversarial training is performed for MI between G and $[F; H]$.

4.2. Instance-wise Adaptive Inference

The widened feature space is disentangled via the *disentangling module* and re-weighted for each test instance. The operative idea for *instance-wise adaptive inference* (IAI) is that the feature attributes needed for accurate prediction of each test example may be different from those of the training set due to distribution shift. With similar intuition, test time adaptation techniques for domain generalization are recently presented, where the classification head is re-trained at test time with pseudo-labeled test samples with high confidence [12] or by minimizing the entropy [27]. Dubey *et al.* [6] encodes a few unlabeled samples belonging to a domain via kernel mean embedding and then feeds it in conjunction with input data. Our IAI is different from these previous adaptive approaches in that additional training at test time or pre-processed domain information in the input space is not needed. Besides, they are designed assuming multiple domains for training (*i.e.*, domain generalization), and hence exhibit limited adaptiveness in our scenario receiving only one source distribution (*i.e.*, SDG). See Sec. 5 for empirical comparison.

4.2.1 Disentangling Module

To define a *disentangling module*, let the feature extractor F have L layers. For an arbitrary $1 \leq l \leq L$, F can be divided into two sub-modules, *i.e.*, $F = \{F_{[1,\dots,l]}, F_{[l+1,\dots,L]}\}$, where $[\cdot]$ denotes the indices of the layers included in the sub-modules. Then, let $\mathcal{M} \subset \mathcal{F}$ be the sub-feature space obtained right after F_l and $\mathcal{Z} \subset \mathcal{F}$ be another sub-space after F_L. For simplicity, we call $F_{[1,\dots,l]} : \mathcal{X} \cup \mathcal{X}' \rightarrow \mathcal{M}$, $F_{[l+1,\dots,L]} : \mathcal{M} \rightarrow \mathcal{Z}$ by the lower-level network F_a, the higher-level network F_b, respectively. We employ M number of F_b denoting each of them by $F_b^1, \cdots F_b^M$, which are expected to encode different semantic information one another, as depicted in Fig. 4. We call $\{F_b^m\}_{m=1}^M$ as *disentangling module* F_D.

To disentangle $\{\mathbf{z}_i\}_{i=1}^M$, we regularize them by cosine similarity loss $\mathcal{L}_{\text{COS}}$, defined as:

$$\mathcal{L}_{\text{COS}} = \frac{1}{\binom{M}{2}} \sum_{m \neq n} \frac{\mathbf{z}_m \cdot \mathbf{z}_n}{\|\mathbf{z}_m\| \|\mathbf{z}_n\|}. \tag{8}$$

Exactly the same regularization is applied to $\mathbf{z}'$. By this regularization, we intend to make $F_b^1, \cdots F_b^M$ encode different semantic representations one another, *e.g.*, F_b^1 clearly activates 'texture', while F_b^2 detects 'object shape', and so on. The disentangled features $\{\mathbf{z}_i\}_{i=1}^M$ are concatenated to $\mathbf{z}_{[1,\dots,M]} = [\mathbf{z}_1, \cdots, \mathbf{z}_M]$ and fed into the classification head H. Because an ordinary CNN optimized for image classification frequently represents semantically entangled feature space [4], IAI is limited to be applied directly with a vanilla CNN. However, if we disentangle multiple independent representations from one another by F_D, useful features for the prediction can be exploited, without using other unnecessary (sometimes detrimental) features for generalization.

In overall, our method is trained by iterative two phases, optimizing (*a*) the style-complement module G, and (*b*) the discriminative model including F_a, $\{F_b^i\}_{i=1}^M$, q, and H. The comprehensive loss $\mathcal{L}_{\text{G}}$ for (*a*) and $\mathcal{L}_{\text{D}}$ for (*b*) are

$$\mathcal{L}_{\text{G}} = \hat{I}(\mathbf{z}_m, \mathbf{z}'_m) + \alpha \mathcal{L}_{\text{MMD}}, \tag{9}$$

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{CE}} + \beta_1 \mathcal{L}_{\text{CL}} + \beta_2 \mathcal{L}_{\text{NLL}} + \beta_3 \mathcal{L}_{\text{COS}}, \tag{10}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss for the classifier, and α, β_1, β_2, and β_3 are hyper-parameters for balancing.

4.2.2 Re-weighting

By minimizing $\mathcal{L}_{\text{D}}$, the classification head H learns to select meaningful information for discrimination of source data from semantically diverse representations (*e.g.*, concentrating more on $\mathbf{z}_1$ (texture) than $\mathbf{z}_2$ (shape) with divergent weights). However, if there is a significant distribution shift from the source to the target dataset, the features helpful for prediction would also be shifted. Thereby, we adjust referred feature space differently for each test instance. Formally, IAI makes activated features $\mathbf{z}^A \in \mathbb{R}^{N \times M \times C}$ from $\mathbf{z}_{1,\dots,M} \in \mathbb{R}^{N \times C}$ by

$$\mathbf{z}_i^A = \left[\frac{1}{w_1^s} \mathbf{z}_1 \odot \mathbf{z}_1; \cdots; \frac{1}{w_M^s} \mathbf{z}_M \odot \mathbf{z}_M \right], \tag{11}$$

$$w_m^s = \frac{1}{NC} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathbf{z}_{mij}^{(s)}, \tag{12}$$

where $[;]$ denotes concatenation and $\odot$ denotes element-wise multiplication. $\mathbf{z}_{mij}^{(s)} \in \mathbb{R}$ denotes the j-th element in the latent vector obtained from instance i in the source data $\mathbf{z}_{mi}^{(s)} \in \mathbb{R}^C$. Note that this operation is only applied at inference, not at training. The activated features $\mathbf{z}^A$ are fed into the classification head H in exactly the same way as $\mathbf{z}_{[1,\cdots,M]}$ to H at training.

Adjustment of the referred feature space. Our IAI is grounded in the notion that *stronger activations typically correspond to the detection of highly discriminative features* [24, 31]. From this perspective, considering that $\{F_b^m\}_{m=1}^M$ represents distinct features from one another (by $\mathcal{L}_{COS}$), the activation level from each module F_b^m induced as $\mathbf{z}_m$ offers the importance of each module in detecting discriminative factors for either the source or target. Hence, if the m-th module is well activated for the target and less activated for the source ($w_m^t > w_m^s$), it is desirable to use that module more at testing, and vice versa. Implementing this intuition involves adjusting the weight of each module at test time by multiplying the ratio $\mathbf{z}_m / w_m^s \simeq w_m^t / w_m^s$ when inducing $\mathbf{z}_m$, resulting in $1/w_m^s \times \mathbf{z}_m \odot \mathbf{z}_m$.

Overall, we claim that IAI generalizes the model on diverse distribution mismatches. From a semantic point of

Dataset		DG-CIFAR					PACS			
		Weather	Blur	Noise	Digits	*avg.*	A	C	S	*avg.*
	Base*	64.43	65.31	45.95	68.20	60.97	54.66	19.37	26.11	33.38
SDG	ADA	72.67	67.04	39.97	66.62	61.58	58.72	45.58	48.26	50.85
	ME-ADA	72.67	67.04	54.21	65.10	64.65	58.96	44.09	49.96	51.00
	L2D	75.98	69.16	73.29	72.02	72.61	56.26	**51.04**	58.42	55.24
UBL	LfF*	49.77	48.53	22.29	51.30	43.22	56.35	22.65	26.70	35.23
	LDR*	49.89	52.97	55.06	55.79	53.43	27.98	21.63	19.14	22.92
	UBNet*	63.41	64.44	42.18	67.04	59.27	59.42	26.80	26.20	37.47
TTA	Tent-C*	46.35	47.33	35.78	43.75	43.30	58.01	21.59	24.79	34.80
	Tent-B*	52.86	55.25	52.26	57.76	54.53	57.18	43.39	39.25	46.61
	T3A*	45.21	48.12	43.40	58.11	48.71	56.98	23.25	32.32	37.52
	IAI	**76.37**	**71.36**	**74.14**	**79.57**	**75.36**	**60.81**	45.62	**65.78**	**57.40**

Table 1. **Single domain generalization.** We display the best performance (accuracy) by **bold** and second performance by underline. We experiment on different seeds three times and report the average of them. * denotes the result by our own experiment and the others are from the papers.

Dataset		B-CIFAR	IMDB				
			train on EB1		train on EB2		
			EB2	TEST	EB1	TEST	*avg.*
	Base*	23.26	57.84	69.75	59.86	84.42	67.97
SDG	ADA*	16.56	66.90	77.71	63.11	86.54	73.57
	ME-ADA*	18.91	65.93	76.91	63.10	85.56	73.13
	L2D*	24.18	72.68	81.80	65.87	88.01	77.09
UBL	LfF*	28.61	75.72	77.88	**72.98**	81.79	77.09
	LDR*	**29.95**	74.71	79.19	71.81	83.20	77.23
	UBNet*	25.52	75.00	83.56	71.17	88.90	79.66
TTA	Tent-C*	23.41	61.82	72.81	70.82	88.26	73.43
	Tent-B*	26.19	72.00	81.90	61.11	86.34	75.34
	T3A*	25.53	70.12	79.80	66.92	86.18	75.76
	IAI	27.37	**77.63**	**84.84**	72.14	**89.52**	**81.03**

Table 2. **Unbiased learning.** We follow the exactly same way as single domain generalization to report table. We set the backbone of L2D as ResNet18 because it performs better than original setup using AlexNet [18].

view, the distribution mismatch between the source and target data means that the attributes constituting the image are distributed differently. Thus, the features that the model relies on for better prediction would be also divergent. This is not limited only to one specific problem (*e.g.*, biased or domain-limited) but corresponds to an arbitrary distribution shift. Therefore, disentangling and re-weighting framework can be applied for diverse scenarios in a general manner.

5. Experiments

5.1. Experimental Setup

Evaluation protocol. Since it is practically hard to know the true distribution p, as pointed out in Sec. 2, we evaluate the models on both SDG and UBL tasks on the target distribution p_T to estimate the generalizability of the model. We intentionally make the domain-limited and biased training distributions as follows. For SDG (DG-CIFAR, PACS), the model is trained on a single domain and evaluated on multiple OOD domains. For UBL, we follow the two scenarios, each of which is followed by B-CIFAR [3] and IMDB [15]. As a first scenario (B-CIFAR), we make an extremely biased set towards a certain attribute (bias), *i.e.*, 'extreme bias (EB)'. We use EB as training data and evaluate the model on a uniformly distributed set for bias attribute, *i.e.*, an unbiased set. For the second scenario (IMDB), mutually exclusive sets, EB1 and EB2, and an unbiased test set are utilized. We use one of $\{$EB1, EB2$\}$ as training data and evaluate it on the other one. The test set is used for both.

Datasets. We evaluate our method on two modified versions of CIFAR10, domain generalization (DG)-CIFAR and biased (B)-CIFAR, following the exact same way as in Sec. 3. We further add real world-like datasets PACS [33] and IMDB [28] for SDG and UBL, respectively. PACS consists of 4 domains, P (photo), A (art painting), C (cartoon), and S (sketch). We set P as source data and $\{$A,C,S$\}$ as target data. We divide IMDB into EB1 and EB2 considering 'gender' as the target label and 'age' as a bias. Consequently, EB1 $\{$(female, young), (male, old)$\}$ and EB2 $\{$(female, old), (male, young)$\}$ are sampled, and we set an unbiased test set with all four possible combinations of (gender, age). Detailed data distribution is illustrated in the appendix.

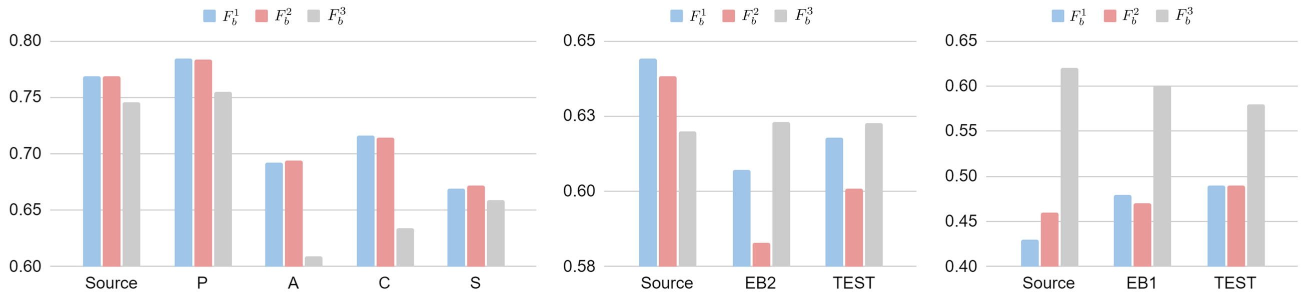
Competing methods. We use ADA [25], ME-ADA [32], and L2D [30] for SDG and LfF [21], LDR [19] and UBNet [13] for UBL as the competing methods. To investigate the generalization of TTA toward diverse distribution, we add Tent-C, Tent-B [27] and T3A [12] as our baselines. For hyperparameters in the competing models, we follow the same setting as presented in the original papers, and we report our reproducedresults if they are better than the original ones.

We use ResNet18 [10] as our base model. For the experiment, we use Adam optimizer [16] and grid search for learning rate (initial value and decay schedule), stopping criterion, and batch size. More detailed implementation details are provided in the appendix.

5.2. Experimental Result and Discussion

First of all, we observe that previous methods are over-optimized to each specific mismatch problem. Similarly to the experimental results in Sec. 3, all the SDG methods generalize worse than UBL for biased distribution, and vice versa. Although L2D boosts the generalizability of the base model for B-CIFAR and IMDB, it is limited compared to UBL models. We conjecture this is because simply stylized images cannot guarantee a less-biased distribution for bias attributes.

However, our *instance-wise adaptive inference* contributes to generalization. According to Tab. 1 and Tab. 2, applying *instance-wise adaptive inference* on L2D, our method significantly improves the generalizability on both



(a) (b) (c)

Figure 5. **Weight shift by *instance-wise adaptive inference*.** (a) PACS, (b) IMDB trained with EB1, and (C) IMDB trained with EB2. Source and P in (a) are split from the same domain. The y-axis means w_i corresponded to each data indicated on the x-axis. We average the three results of the same experiments in Tab. 1 and Tab. 2.

SDG and UBL, implying that our framework is robust on arbitrary sampling errors. Our method exhibits the best performance for DG-CIFAR, especially in the Digit domain. Intuitively, the art painting (A) domain is most similar to the photo (P) domain and this is quantitatively explained by multiple methods (`Base`, `ADA`, `ME-ADA`, `LfF`, `UBNet`, `Tent`, `T3A`), predicting much better on A than on C and S. However, our method exhibits the best performance on the sketch (S), showing remarkable generalization. Although `LDR` performs the best for B-CIFAR, our model shows competitive generalizability with a significantly improved accuracy from `L2D`. The test accuracy on EB1 and EB2 means the generalization on conflicting samples to train data and TEST signifies more widespread; that is, more similar to the true distribution. Our method shows the best generalizability for the latter ones.

Lastly, test time adaptation methods have some limitations. Although TTA methods show competitive performance on B-CIFAR, their generalization for other datasets is limited because these frameworks are designed for domain generalization that requires multiple source domains for training. The performance improvement is insignificant because the UBL scenario learns with one biased source dataset as it is for SDG.

5.3. Ablation Study

Ablation study on sub-components. We investigate the contribution of sub-components presented in Sec. 4 for the generalizability of the model. Tab. 3 exhibits that *disentanglement* and *re-weighting* improve performance for both SDG and UBL. One might argue that this performance improvement is due to the increased model size. However, the accuracy with the *multi-features extractors* in Tab. 3 with a comparable number of parameters to our final method demonstrates that it is not the case.

We additionally estimate weight shift by *instance-wise adaptive inference* and report it in Fig. 5. The activation scores $w_{1,2,3}$, corresponding to $F_b^{1,2,3}$, are significantly different between the source and other target data for both SDG and UBL. Especially, it is notable that ‘Source’ and ‘P’ (disparate data but same domain) in (a) have highly similar activation distribution, while others do not. Consequently, Tab. 3 and Fig. 5 conclude that the target data need different features from the source data for prediction.

Multi-features extractors		✓	✓	✓
Disentanglement ($\mathcal{L}_{\text{COS}}$)			✓	✓
Re-weighting				✓
Acc (SDG)	52.73	51.00	56.99	**57.40**
Acc (UBL)	77.09	77.30	78.03	**81.03**

Table 3. **Ablation Study on sub-components.** The sub-components are applied step by step. The model with none of them applied is `L2D` with ResNet18 as the backbone. Acc (SDG) and Acc (UB) denote accuracy on PACS and IMDB, respectively.

Comparison on disentangling regularization. Several regularizations including distance-based and similarity-based techniques can be applied to disentangle feature space. To find effective disentanglement for *intance-wise adaptive inference*, we compare them. Tab. 4 shows cosine similarity contributes the best.

	L1	L2	KLD	COS
Acc (SDG)	51.10	47.45	52.22	**57.40**
Acc (UBL)	74.40	74.89	78.98	**81.03**

Table 4. **Comparison on regularization techniques for disentanglement.** L1, L2, KLD, and COS denote L1-norm, L2-norm, Kullback-Leibler divergence, and cosine similarity, respectively.

Generalization when $p_S \simeq p_T$. Despite the generalizability for the distribution mismatch problem, it is useless if the model degrades performance when $p_S \simeq p_T$. Thus, assuming training and testing data in the dataset sampled from the same distribution, we compare the accuracy of our model to the baseline model (ResNet18). Since the baseline model is designed only for $p_S \simeq p_T$, the performance of `Base` implies the guidance for this scenario. Figure 6 shows that our model even outperforms the baseline model.

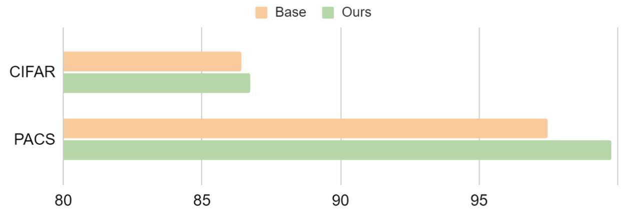


Figure 6. **performance when $p_S \simeq p_T$.** we use a train and validation set in CIFAR and P(train) and P(test) for PACS.

Model size. One might argue that *instance-wise adaptive inference* generalizes well due to the increased number of model parameters. To address this concern, we compare the performance of the proposed method to several baselines. The results in Tab. 5 show that simply increasing the number of model parameters is not sufficient for generalization.

Model	ResNet18	ResNet50	ResNet101	`L2D` (base)	`L2D` (large)	`IAI` (Ours)
Params.	11,702K	25,557K	42,521K	12,037K	26,709K	28,827K
Acc (SDG)	33.38	40.47	43.11	55.24	56.78	**57.40**
Acc (UBL)	67.97	62.42	65.37	77.09	77.77	**81.03**

Table 5. **Investigation for the contribution of model size.** `L2D` (base) and (large) use ResNet18 and ResNet50 as backbone network, respectively.

6. Related Work

Single Domain Generalization. For single domain generalization, adversarial data augmentation (ADA) iteratively generate additional training examples with adversarial augmentation from a worst-case fictitious target domain. M-ADA [25] applied wasserstein auto-encoders (WAE) to generate adversarial samples. Domain shift to target distribution is encouraged by the maximum mean discrepancy [22]. Along with ADA and M-ADA, Zhao *et al.* [32] additionally applied the maximum-entropy-based regularization loss term to address limitations caused by the heuristic approach when finding hard cases. Fan *et al.* [7] presented adaptive standardization and rescaling normalization to attune gaps coming from different domains. Li *et al.* [20] suggested that expanding coverage of the training domain is limited due to the lack of appropriate safety and effectiveness constraints. Under this observation, they progressively expand the domain. Wang *et al.* [30] synthesized images minimizing MI to cover a wider range of domains and maximized MI between the samples from the same semantic category to learn discriminative logic. Based on the motivation that convolution features can be decomposed into universal and elemental visual features, Wan *et al.* [26] eliminated unfavorable features for generalization ability during the decomposition and composition of features. Cugu *et al.* [5] applied data augmentations to simulate multiple domains. Then, they enforced the output of the model to be consistent across original and simulated domains via class activation map (CAM) loss.

Unbiased Modeling. The first approaches to debias a model was in a supervised manner, assuming the bias attribute is known and annotated [2, 15, 9, 1]. After that, several studies mitigated it to an unlabeled but pre-defined bias [8, 29, 3].

To address the issue that known bias is not practical, the following works for unknown bias have been proposed. LfF [21] estimated bias-conflicting samples by the intentionally biased classifier and employed generalized cross-entropy loss with large weights to de-bias the target model. Lee *et al.* [19] proposed feature-level augmentation to generate various bias-conflicting samples. They disentangled representation into intrinsic and bias attributes to find bias-conflicting features. Hong *et al.* [11] suggested a contrastive bias learning approach for known biases. They proposed a soft bias-contrastive loss which weights bias-contrastive loss to tackle the unknown bias case. UBNet [13] addressed unknown bias by a conservative approach that widens feature space to be referred to via hierarchical features and orthogonal regularization. BPA [23] proposed a cluster-wise reweighting scheme to make the model consider minority groups to minimize the overall loss enough.

7. Summary and Discussion

In this paper, we consider a practical scenario that the type of distribution mismatch cannot be estimated during training due to limited access to the target data. Based on this motivation, we present a novel framework robust to diverse distributions. Specifically, a widened and disentangled feature space is referred flexibly according to the queried instance, *i.e.*, *instance-wise adaptive inference*. On the two special cases of distribution mismatch problems, our proposed method exhibits the best generalizability performance.

We present the effectiveness of our framework for generalization and the contribution of each component (widening, disentangling, re-weighting) through several experiments. Nonetheless, we observe in the experiments that the *re-weighting* rarely helps in some cases. We carefully suggest that an *adaptive inference* would have potential if more robustness is studied, leaving it as a promising future work. Further, to be more widespread, generalization on the distribution combination (*e.g.*, unbiased domain generalization) could be another practical scenario.

Acknowledgement. Corresponding authors are Joonseok Lee, Myoungjoo Kang. Joonseok Lee was supported by the New Faculty Startup Fund from SNU and by NRF grant (No. 2021H1D3A2A03038607, 2022R1C1C1010627, RS-2023-00222663) and IITP grant (No. 2022-0-00264) funded by the government of Korea. Myungjoo Kang was supported by the NRF grant [2012R1A2C3010887] and the MSIT/IITP ([1711117093], [2021-0-00077], [No. 2021-0-01343, Artificial Intelligence Graduate School Program(SNU)]).

References

[1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. Representation learning with statistical independence to mitigate bias. In *Pro. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 8

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proc. of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 8

[3] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020. 1, 6, 8

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 5

[5] Ilke Cugu, Massimiliano Mancini, Yanbei Chen, and Zeynep Akata. Attention consistency on visual corruptions for single-source domain generalization. In *CVPR*, 2022. 1, 8

[6] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *CVPR*, 2021. 5

[7] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *CVPR*, 2021. 1, 8

[8] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231*, 2018. 8

[9] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly debiasing face recognition and demographic attribute estimation. In *ECCV*, 2020. 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[11] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. In *NIPS*, 2021. 1, 8

[12] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *NIPS*, 2021. 5, 6

[13] Myeongho Jeon, Daekyung Kim, Woochul Lee, Myungjoo Kang, and Joonseok Lee. A conservative approach for unbiased learning on unknown biases. In *CVPR*, 2022. 1, 3, 6, 8

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NIPS*, 2020. 4

[15] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, 2019. 6, 8

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Univ. of Toronto, 2009. 3

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 6

[19] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. In *NIPS*, 2021. 1, 3, 6, 8

[20] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. *CVPR*, 2021. 1, 8

[21] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *NIPS*, 2020. 1, 3, 6, 8

[22] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *CVPR*, 2020. 1, 8

[23] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *CVPR*, 2022. 1, 8

[24] Ignacio Serna, Alejandro Pena, Aythami Morales, and Julian Fierrez. InsideBias: Measuring bias in deep networks and application to face gender biometrics. In *ICPR*, 2021. 5

[25] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NIPS*, 2018. 1, 3, 6, 8

[26] Chaoqun Wan, Xu Shen, Yonggang Zhang, Zhiheng Yin, Xinmei Tian, Feng Gao, Jianqiang Huang, and Xian-Sheng Hua. Meta convolutional neural networks for single domain generalization. In *CVPR*, 2022. 1, 8

[27] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv:2006.10726*, 2020. 5, 6

[28] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *ECCV*, 2018. 6

[29] Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. *arXiv:1903.06256*, 2019. 8

[30] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *ICCV*, 2021. 1, 3, 4, 6, 8

[31] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 5

[32] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *NIPS*, 2020. 1, 3, 6, 8

[33] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, 2020. 6