Better May Not Be Fairer: A Study on Subgroup Discrepancy in Image Classification

Ming-Chang Chiu
University of Southern California
Los Angeles, CA
mingchac@usc.edu

Pin-Yu Chen
IBM Research
Boston, MA
pin-yu.chen@ibm.com

Xuezhe Ma
University of Southern California
Los Angeles, CA
xuezhema@isi.edu

Abstract

*In this paper, we provide 20,000 non-trivial human annotations on popular datasets as a first step to bridge gap to studying how natural semantic spurious features affect image classification, as prior works often study datasets mixing low-level features due to limitations in accessing realistic datasets. We investigate how natural background colors play a role as spurious features by annotating the test sets of CIFAR10 and CIFAR100 into subgroups based on the background color of each image. We name our datasets **CIFAR10-B** and **CIFAR100-B**[1] and integrate them with CIFAR-Cs.*

*We find that overall human-level accuracy does not guarantee consistent subgroup performances, and the phenomenon remains even on models pre-trained on ImageNet or after data augmentation (DA). To alleviate this issue, we propose **FlowAug**, a* semantic *DA that leverages decoupled semantic representations captured by a pre-trained generative flow. Experimental results show that FlowAug achieves more consistent subgroup results than other types of DA methods on CIFAR10/100 and on CIFAR10/100-C. Additionally, it shows better generalization performance.*

Furthermore, we propose a generic metric, MacroStd*, for studying model robustness to spurious correlations, where we take a macro average on the weighted standard deviations across different classes. We show MacroStd being more predictive of better performances; per our metric, FlowAug demonstrates improvements on subgroup discrepancy. Although this metric is proposed to study our curated datasets, it applies to all datasets that have subgroups or subclasses. Lastly, we also show superior out-of-distribution results on CIFAR10.1.*

1. Introduction

Deep neural networks (DNNs, e.g., [25, 19]), properly trained via empirical risk minimization (ERM), have been demonstrated to significantly improve benchmark performances in a wide range of application domains. However, minimizing empirical risk over finite or biased datasets often results in models latching on to *spurious correlations* that do not show a robust relationship between the input data and output labels. Moreover, benchmark evaluations based solely on average accuracy may overlook these critical issues. For instance, Fig. 1 shows that on CIFAR10, even though a standard ERM model reaches human-level test accuracy (*red line*), if we dive deeper into each class and compute their respective worst test accuracy stratified by background colors, they are inconsistent across the ten classes and the degradation from total accuracy is huge (*black line*) for some. Such inconsistency and discrepancy have huge real-world implications, suggesting DNN models may make biased decisions against or in favor of specific spurious factors, such as certain background colors.

Researchers have been working in different directions to understand the effect of spurious correlations, including model over-parameterization [35], causality [1] and information theory [27, 53]. Various techniques have emerged over the years to address this challenge, among which DA [41] has stood out for its simplicity and effectiveness. DA shows better generalization results in various machine learning tasks than other approaches [52, 50, 47, 18, 46, 39]. These augmentation methods, however, are often based on heuristic and coarse image processing techniques such as flipping, rotating, blurring, or manipulating images by mixing attributes from other inputs [52, 50, 11, 21] (Fig. 2); therefore, they can only address limited aspects of spurious correlations, for which we will show an example in § 2. To address this limitation, instead of mixing low-level features, we seek to augment the training set by learning *semantic* deep representations and then using them to generate new images.

In this paper, as the very *first* step towards comprehensive evaluation of subgroup performance against *semantically meaningful and realistic spurious correlations* in image classification, we conduct a case study experiment to

[1] Dataset is released at https://github.com/charismaticchiu/CIFAR-B

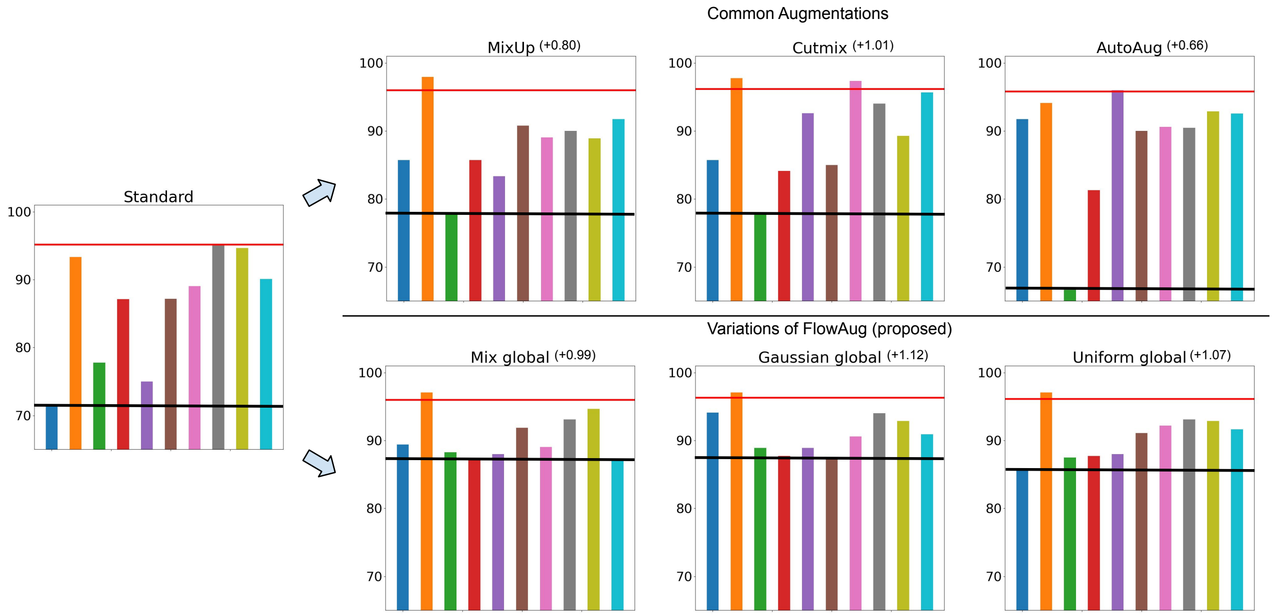


Figure 1: **FlowAug reduces subgroup discrepancy. CIFAR10-B** enables us to observe the worst test time subgroup accuracy in each class. Standard ERM shows *subgroup discrepancy*, uneven subgroup performances across all classes, and a huge gap between total accuracy (*red line*) and the worst subgroup accuracy (*black line*). This issue persists even after common DAs are used (*top*). Our proposed **FlowAug** mitigates this issue (*bottom*) and also reports improved overall performance.

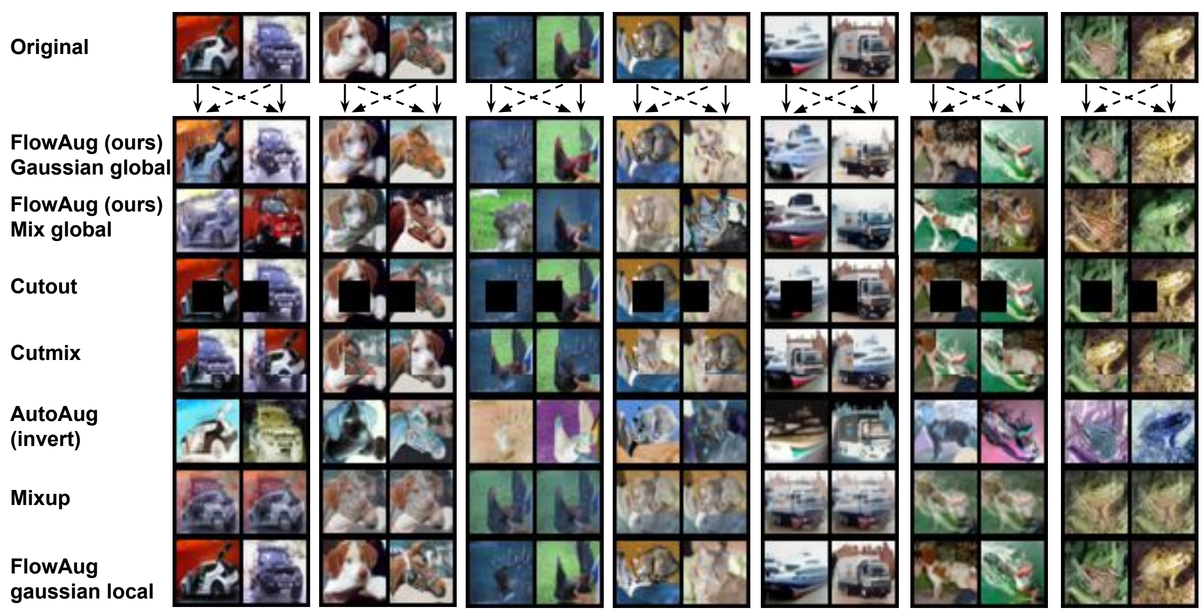
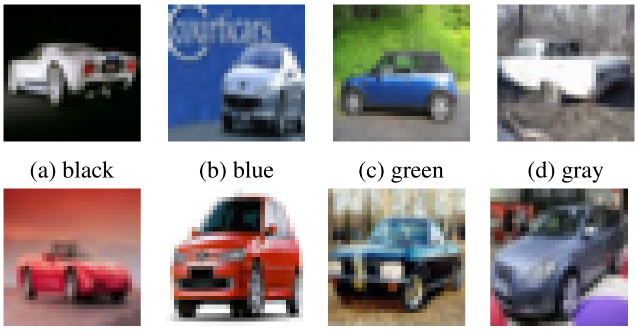


Figure 2: **Examples of different augmentation methods.** Row 2 & 3 are generated by our methods.

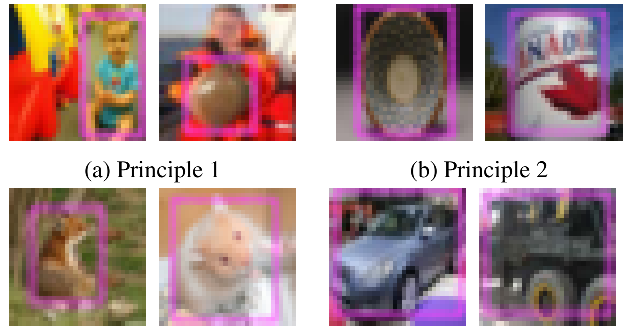
investigate background colors as spurious features (§2), for their commonality in image classification and immediate implications for trustworthiness [34]. To directly quantify the results, we annotated the test data of CIFAR10 and CIFAR100 into subgroups based on natural image background colors (see Fig. 3), yielding **CIFAR10-B**ackground and **CIFAR100-B**ackground. To the best of our knowledge, our datasets are *two of the only human-annotated* benchmark datasets with a *natural semantic bias*. We argue that the background color bias should be a *necessary spurious correlation* for future studies on robustness to benchmark on and so our work can facilitate future works to benchmark their capabilities on reducing learning spurious factors. Equipped with our datasets, we can investigate the



(a) black (b) blue (c) green (d) gray

(e) red (f) white (g) brown (h) others

Figure 3: **Examples of CIFAR10-B (Car).** We label seven common background colors for CIFAR10 and CIFAR100. Difficult examples are categorized as "others".



(a) Principle 1 (b) Principle 2

(c) Principle 3 (d) Principle 4

Figure 4: **Examples of the four main principles of our labeling philosophy.** See § 3.2 for philosophy descriptions.

reliance on background color of deep neural models in a multi-class multi-subgroup setup.

We reveal that even though standard DNNs have achieved human-level accuracy in image classification tasks, the performances fluctuate across different subgroups. This phenomenon demonstrates the reliance on background colors as spurious features. Moreover, applying some popular DA methods or pre-training on larger dataset such as ImageNet do not prevent the models from producing uneven accuracies across subgroups, as shown in Fig. 1 & 5, which further shows that low-level feature manipulations or brute-force pre-training are not sufficient to address spurious correlations and better methods are needed. To quantify our observations, we propose *MacroStd*, a metric to quantify subgroup performance discrepancy and imply the reliance on spurious correlations (§ 3.4).

To enable semantic data augmentations and address the issue of uneven accuracies, we propose **FlowAug**, a novel DA method which is capable of manipulating images semantically via decoupled representations learned from invertible generative flows [28] (§3). Concretely, our deep generative augmentation approach incorporates a novel flow-based generative model that encourages disentanglement of local and global representations from images, which arguably correspond to the image "style" and "content" [15, 54], respectively. By operating on the global representation that is isolated with the image class label, FlowAug semantically creates new images for DA.

More consistent performance across subgroups demonstrates the effectiveness of FlowAug. Also, we integrate our CIFAR-Bs with CIFAR-Cs [20] for broader out-of-distribution (OOD) evaluations and observe similar consistent subgroup performances. Furthermore, though not our main foci, we also find that superior experimental results on various in-distribution (ID) and OOD benchmarks, including CIFAR10, CIFAR100 [23], CIFAR10.1[33, 42] bolster our belief that low-level manipulations or brute-force pre-training are not sufficient.

To summarize, our contributions are four-fold,

- We curate 20,000 human-annotated labels and two **CIFAR-B** datasets that reveal the *subgroup discrepancy* phenomenon and allow us to (1) study semantically meaningful and realistic spurious correlations in a multi-class multi-subgroup setup, and (2) integrate with CIFAR-C for OOD evaluations.
- We propose **FlowAug**, a novel augmentation method that leverages "expert knowledge" of the deep generative model to change semantic attributes of images and empirically shown to reduce subgroup discrepancy.
- We propose a generic metric that captures the subgroup discrepancy phenomenon of ERM and common DA methods and measures the sensitivity of model performances to spurious correlations, and demonstrate FlowAug's effectiveness in this regard.
- As an additional benefit, we conduct experiments on CIFAR10/100 and CIFAR10.1 and show FlowAug can further provide superior performances on ID and OOD datasets.

2. A Motivating Example of Subgroup Discrepancy

We investigate background color as the spurious correlation with our CIFAR10-B (§ 3.2) by first training a standard Resnet18 for 250 epochs with weight decay 5×10^{-4}, initial learning rate 0.1 and learning rate decay at [100, 150] epochs by a factor of 0.1. We observe significant performance degradation in the subgroups of some classes, for example class "airplane," "bird" and "deer" (Fig. 5 (left)). Moreover, after applying DAs such as AutoAug [9], the same phenomenon remains, for instance observe the "bird" class in the mid-left plot in Fig. 5. More surprisingly, even after we fine-tune Resnet18 pre-trained on ImageNet (pre-Resnet) with similar protocol to [22], the degradation continues to exist (Fig. 5 (right)).

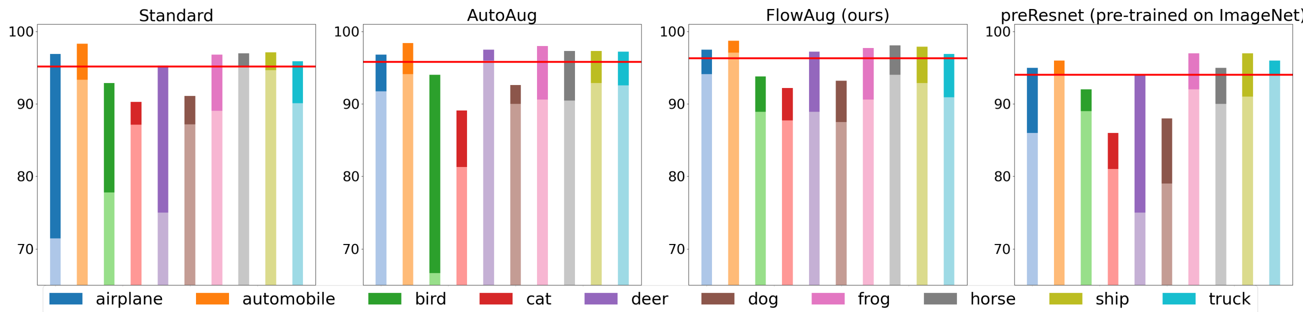


Figure 5: **Gaps between class accuracies (dark bars) and their worst subgroup accuracies (light bars).** Although a standard CNN model can reach human-level accuracy (*red line*), we find that the subgroup performances can be surprisingly low. Even after data augmentation (*mid-left*) or fine-tuned from ImageNet (*right*), the same phenomenon remains. FlowAug *(mid-right) shows more consistent results and mitigates the performance gaps (dark bars) the most.*

In summary, though a Resnet model with or without popular DAs achieve more than 90% class accuracies, their respective background subgroup performances can be surprisingly low. This phenomenon, which we call "subgroup discrepancy" or "in-class variability," shows that background colors play a role in the performance of a standard DNN model and constitute spurious correlations; otherwise, the performances should be relatively consistent. This triggers our interest in mitigating the performance variability in subgroups, i.e. the reliance on background attributes. And we show in Fig. 1 & Fig. 5 that using FlowAug achieves more consistent subgroup results.

Furthermore, fine-tuning model pre-trained on larger benchmark such as ImageNet does not reduce *subgroup discrepancy* even on dataset like CIFAR, so we reasonably conjecture that this phenomenon will exist in other datasets.

3. Methods

In principle, DA takes the form of a particular set of transformation functions T where each $t \sim T$ transforms an input x in a particular fashion. Moreover, an expert may have the knowledge to design label-preserving transformations T in a way that $t(x)$'s leave the label unchanged.

After the transformations, the dataset $\mathcal{D}$ will be augmented to $\{(x_i^{1:K}, y_i)\}_{i=1}^n$, where K is the number of times x_i is transformed. From a frequentist point of view, we can apply any MLE algorithm to the augmented dataset, and the hope is that the learned model can better estimate the true model since we have more data.

In this section, we discuss our generative flow model, present our datasets CIFAR10-B and CIFAR100-B for studying spurious correlation, and detail our augmentation algorithms. Lastly, we introduce two metrics to quantify the effect of spurious correlation.

3.1. Decoupling representations with Flow-based Generative Models

Prior work has shown that embedding a invertible normalizing flow model as a decoder in a variational autoencoder (VAE) can decouple global (z) and local (ν) representations of images in an unsupervised fashion [28], and can *switch* the decoupled representations of different images to alter their semantic attributes (see Appendix). We presume the global information corresponds to the *style* of the image and local leans toward the *content* in the neural style transfer literature [15, 54]. In this work, we apply the flow model $\mathcal{F}$ to encode images into global and local representations and also decode them back to image space like VAEs,

$$z, \nu \leftarrow \mathcal{F}_{enc}(x); \quad x' \leftarrow \mathcal{F}_{dec}(z, \nu) \tag{1}$$

where $z \sim \mathcal{N}(\mu(x)), \sigma(x))$, $\mu(x)$ and $\sigma(x)$ are neural networks learned from the data, and $\nu \sim \mathcal{N}(0, I)$. z is a d_z-dimensional vector where d_z is the dimension of the latent space and the size of ν is the same as the input image x.

We further hypothesize that z includes information on colors or more, which are spurious to the ground truth, and ν bears information about the shape, object, etc., which are more indicative of the labels. In § 5.1, we will do an ablation study to attest this hypothesis.

3.2. Datasets quantifying spurious background correlations

We curate CIFAR-10-B & CIFAR-100-B to identify and study spurious information in images, and we choose to label the major background colors of CIFAR10 and CIFAR100 validation sets. By learning the subgroup performances, we can measure the sensitivity of a model to different spuriously correlated colors. As shown in Fig. 3, we manually label the background colors of CIFAR10 and CIFAR100, and split them into eight separate groups. We understand people have different criteria toward determining the background color; therefore, we provide our four main

Algorithm 1: FlowAug-Gaussian Global z

```
Input: Flow: F, Dataset: X, L, μ, σ, b
for l = 1, ..., L do
    x ~ X;                      // Sample image
    z, ν ← F_enc(x) ;   // Encode the image
    ε ~ N_trunc(μ, σ^2; b);           // Sample
     perturbations
    z ← z + ε;            // explore space
    x_aug ← F_dec(z, ν) ;   // Decode global
     and local back to image space
end
Output: X_aug
```

Algorithm 2: FlowAug-Mix Gloabl z

```
Input: Flow: F, Dataset: X, Threshold: tr, L, α
for l = 1, ..., L do
    x_1, x_2 ~ X;           // Sample images
    z_1, ν_1 ← F_enc(x_1) ;          // Encode x_1
    z_2, ν_2 ← F_enc(x_2) ;          // Encode x_2
    m ~ Beta(α, α);                   // Sample
     interpolation parameter
    if m < tr then
        m ← 1 − m ;      // Avoid drastic
         change in style
    end
    z_1 ← m z_1 + (1 − m) z_2;        // explore
     space
    x_aug ← F_dec(z_1, ν_1) ;  // Decode global
     and local
end
Output: X_aug
```

labeling principles as follows,

1. We label the color that has the most coverage around the object. In Fig. 4 (a), one may argue the red patch or blue ocean has taken up most of the image in the background, but the "baby" is surrounded completely by the green area, and the "flatfish" is in the red area.
2. When two colors take almost the same coverage other than the object, we choose the color that appears further away. In Fig. 4 (b), black is farther away from the "bowl", and so is the blue sky for the "can".
3. When two colors take almost the same coverage and appear to be at a similar distance, we make a judgment call on the color that has more coverage (Fig. 4 (c)).
4. When multiple colors appear in the background and none is significantly larger than the rest (Fig. 4 (d)), or when the object takes up almost all the space in the picture so that we cannot judge the color in the background, or when the perceived color does not belong to our categories, we put it in the "others" category.

3.3. Algorithms

Knowing properties of ν and z discussed in §3.1, we design two families of transformations to operate on global z: (1) T_1: we add perturbations to z, and (2) T_2: we interpolate global information extracted from different images. The over-arching rationale behind is: by equipping models with label-preserving images under diverse environments (i.e., backgrounds), the model should learn more robust correlations[1]. The second and third row of Fig. 2 demonstrate our method ability in this regard.

More specifically, in T_1 we add truncated Gaussian perturbation ϵ to z,

$$T_1 := \{t(x) = \mathcal{F}_{dec}(z+\epsilon, \nu) | (z, \nu) = \mathcal{F}_{enc}(x), \quad \epsilon \sim \mathcal{N}_{trunc}(\mu, \sigma^2; b), \ \forall x\}.$$

instead of a Gaussian noise, since a Gaussian noise may sample large numbers that potentially destroy the decoding of $\mathcal{F}_{dec}(z, \nu)$. For T_2, we decode two random images x_1, x_2 to retrieve z_1, z_2 and then interpolate z_1 and z_2 stochastically with a parameter m drawn from a Beta distribution,

$$T_2 := \{t(x_i) = \mathcal{F}_{dec}(z_{new}, \nu) | z_{new} = m z_i + (1-m) z_j, \quad m \sim Beta(\alpha, \alpha), (z_i, \nu_i) = \mathcal{F}_{enc}(x_i), \ \forall i \neq j\}.$$

Detailed transformations are elaborated in Algorithm 1 & 2.

We train our models with the following learning objectives: (1) training only with transformed images from T_1 or T_2 instead of the original examples, (2) in addition to transformed images, adding the original dataset, and (3) combining the two algorithms and the original dataset,

$$\mathcal{L}_{FlowAug} = \mathcal{L}(f(t(x)), y; \theta), \ \ t \sim T_1 \ \ or \ \ t \sim T_2, \quad (2)$$

$$\mathcal{L}_{FlowAug+std} = \mathcal{L}(f(t(x)), y; \theta) + \lambda \mathcal{L}(f(x), y; \theta), \quad t \sim T_1 \ \ or \ \ t \sim T_2, \quad (3)$$

$$\mathcal{L}_{combine} = \mathcal{L}(f(t_1(x)), y; \theta) + \lambda_1 \mathcal{L}(f(t_2(x)), y; \theta) + \lambda_2 \mathcal{L}(f(x), y; \theta), t_1 \sim T_1 \ \ and \ \ t_2 \sim T_2, \quad (4)$$

3.4. Quantifying subgroup discrepancy

To quantify the reliance on background attributes, we first propose using the weighted standard deviation,

$$\sigma_w = \sqrt{\frac{\sum_{i=1}^{G} w_i (s_i - \bar{s}^*)^2}{\sum_{i=1}^{G} w_i - 1}}, \quad (5)$$

where s_i's are the subgroup accuracies, $\bar{s}^*$ the weighted mean, w_i's the weights determined by the number of examples in the subgroup, G the number of groups. Weighted Std can be applied to subgroups performances within a class (as in Fig. 6), and across all accuracies from different classes and subgroups.

The second metric we propose is macro standard deviation (*MacroStd*),

$$\sigma_{Macro} = \sqrt{\frac{1}{C}\sum_{i=1}^{C} {\sigma_w^{(i)}}^2}, \tag{6}$$

where $\sigma_w^{(i)}$ is the weighted standard deviation for each class, and C is the number of classes.

MacroStd treats each class equally and measures the sensitivity of a model performance across classes. If *MacroStd* is high, this suggests the model has imbalanced performances across classes and also could be affected by background colors. We conduct a correlation analysis to show our metric is a better indicator for both sensitivity and accuracy (see Appendix).

4. Experiments

In this section, we discuss our empirical results on the study of spurious correlation with our CIFAR10-B & CIFAR100-B and their integration with OOD datasets such as CIFAR10-C and CIFAR100-C. Secondly, although not our primary foci, we present ID and OOD image classification experiments on three datasets — CIFAR10, CIFAR100, CIFAR10.1 — to test the generalization capabilities of applying FlowAug. Lastly, we analyze and provide intuitions on how our approach is superior. Furthermore, the comparing baselines and implementation details are provided.

Due to human resource limit, we are not able to scale labeling efforts to larger benchmark such as ImageNet, but our work has pinpointed critical issues in the subgroup discrepancy in image classification. And we reasonably believe the phenomenon will persist in other datasets given the result from preResnet (Fig. 5 (*right*)). In addition, a recent benchmark work [13] shows that CIFARs are *not necessarily easier than ImageNet*, which also validates our efforts.

4.1. Datasets

Other than our CIFAR10-B and CIFAR100-B that are based on CIFAR10 and CIFAR100 [23], we integrate them with CIFAR10-C & CIFAR100-C [20], which are benchmark datasets to model generalization abilities in the presence of 18 shallow corruptions including blurring, contrast, shift, etc. Finally, CIFAR10.1 [33] is a test set consists of 2000 images collected from TinyImages [42] and contains the same class labels as CIFAR10. Additionally, we include ImageNet-10 based on our labeling method and discuss the results in the Appendix.

4.2. Baselines

We compare our proposed method with four types of low-level DA methods (1) mixing by interpolations, (2) fill-in-with-blank, (3) mixing by fill-in-the-blank, (4) combinations of image manipulations. In our experiments, we compare with the best setups reported in their papers. We include more discussion on rationale behind the selecting the chosen baselines and additional comparisons with composite data augmentations such as AugMix and AugMax in the Appendix.

Mixup [52] does linear interpolation on two random images x_1, x_2 and mix them as $x_{new} = \lambda x_1 + (1-\lambda)x_2$, where $\lambda \sim Beta(\alpha, \alpha)$, and the same applies to the label, $y_{new} = \lambda y_1 + (1-\lambda)y_2$.

Cutout [11] randomly crops out a portion of an image and fills it with a specific color, and the label remains unchanged.

Cutmix [50] crops out an area of image, but fills the area with a portion of the same size from another image. The label of the augmented image is adjusted according to the proportion of the area of two engaging examples.

Autoaug [9] uses reinforcement learning to optimize a pre-defined set of policies, combinations of low-level image manipulation, and then learns the best policy for DA.

Standard refers to the models trained on the original datasets, without using any DA methods.

4.3. Implementation Details

Generative models We pre-train the normalizing flow models as in [28], and they achieve the negative log-likelihood scores in bits/dim (BPD) 3.27 and 3.31 on CIFAR10 and CIFAR100, respectively.

Hyperparameters In Algorithm 1, we simply set $\mu = 0$ and $\sigma = 0.1$ for the truncated Gaussian distribution. As for truncation b, we empirically find that z has an average maximum value around 4 and so we set $b = 4$. In Algorithm 2, we simply set $\alpha = 1$ and $tr = 0.5$. For all models reported in Table 2, we train Resnet18 for 250 epochs with weight decay 0.0005. Also, the learning rate starts at 0.1 and is divided by 10 at [100, 150] epochs. For our learning objectives, we lightly fine-tune λ in Eq. (3) with values of $\{0.01, 0.05, 0.1\}$, and λ_1, λ_2 in Eq. (4) with $\lambda_1 = 1$ and $\lambda_2 \in \{0.01, 0.05, 0.1\}$. The generative flow models are trained on two NVIDIA A40 GPUs, while the Resnet18 are trained on one NVIDIA A40 GPU.

4.4. Empirical Results

MacroStd and WeightedStd Table 1 reports the *MacroStd* and the weighted standard deviation of subgroup performances from the whole dataset. Our approach consistently has both lower *MacroStd* and lower WeightedStd over the baselines. Moreover, in Fig. 6, our approach also achieves lower WeightedStd at the class level. These results show evidence that our approach is less affected by the background colors and hence is more robust.

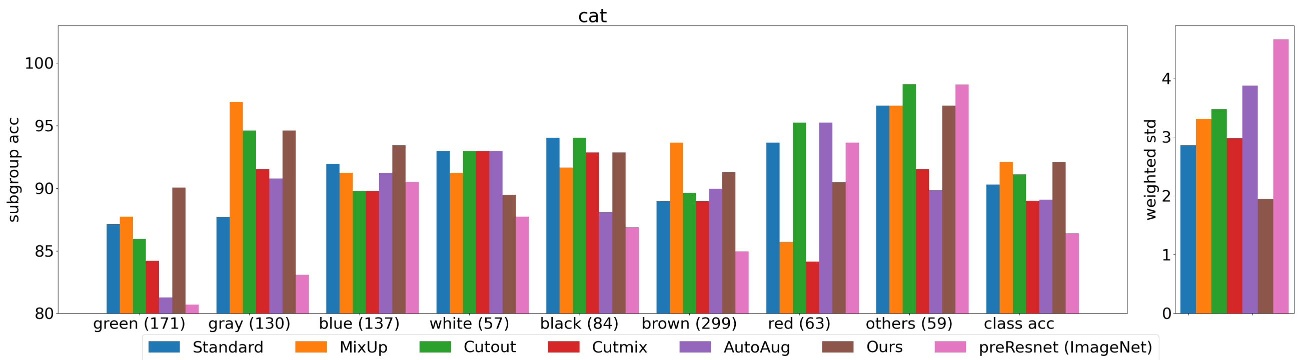


Figure 6: **Subgroup performances (CIFAR10-Cat).** FlowAug has more balanced results across subgroups and lower WeightedStd, suggesting our method is more resistant to spurious correlations such as background color. (.) indicates the number of instances in the subgroup.

	MacroStd		Weighted Std	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100
Standard	2.24	12.24	3.45	16.44
Mixup	2.17	11.94	3.02	16.75
Cutout	1.91	12.45	2.94	16.52
Cutmix	1.91	12.62	3.34	16.87
AutoAug	2.11	11.74	3.54	16.30
Ours (Mix z)	1.99	11.73	2.92	15.96
+ Std	1.81	12.49	2.76	16.76
Ours (Uniform on z)	1.83	12.17	2.98	16.57
+ Std	1.81	11.59	3.26	15.95
+ Mix z	1.85	11.72	2.89	16.31
+ Std + Mix z	1.86	**11.23**	3.12	16.01
Ours (Trunc Gaussian on z)	1.91	12.00	2.82	16.23
+ Std	**1.65**	11.55	3.08	16.02
+ Mix z	1.89	11.95	3.09	16.35
+ Std + Mix z	1.66	11.78	**2.71**	**15.92**
Trunc Gaussian on ν	1.94	12.68	2.81	16.70
Mix ν	2.50	13.27	4.21	18.46

Table 1: ***MacroStd* and Weighted Std.** Lower numbers represent lower reliance on spurious background color correlations and our algorithms are consistently better than the baselines.

CIFAR10-C and CIFAR100-C Another benefit of our datasets is the compatibility with CIFAR-Cs. Together with CIFAR-Cs we are able to evaluate the *subgroup discrepancy* phenomenon in an OOD setting. Fig. 7 shows that FlowAug has reduced *subgroup discrepancy* than other DAs. We exclude AutoAug in Fig. 7 because it contains policies resembling some corruption types of CIFAR-C so we deem it not a fair comparison.

CIFAR10 and CIFAR100 Although ID and OOD generalization performances are not our main foci, our FlowAug demonstrates significant gains on CIFAR10 and CIFAR100 and we report our experimental results in Table 2. Algorithm 1 itself achieves results better than all the baselines. Algorithm 2 also performs better than the *Standard* baseline and is competitive with other methods. When Algorithm 1 and 2 are combined or also add the *Standard* loss (Eq. (4)), they can further enhance the performances.

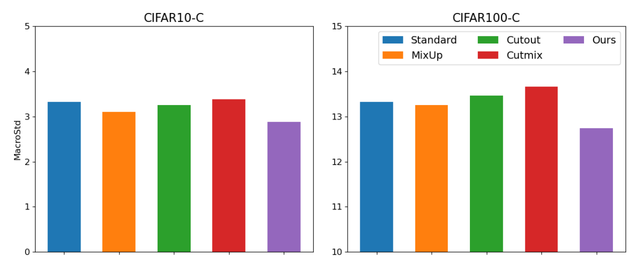


Figure 7: **CIFAR10-C and CIFAR100-C results (severity=1).** We integrate our datasets with CIFAR10/100-C. FlowAug demonstrates more consistent subgroup performances on OOD datasets.

In Algorithm 1, we do an ablation study with uniform distribution in Sec. 5.2. The best improvements on CIFAR10 and CIFAR100 are at 1.42% and 1.47% respectively. The superior results of our deep generative augmentation approach with decoupled representations have shown greater generalization potential.

CIFAR10.1 On another OOD dataset CIFAR10.1, we also observe significant improvements in performances from FlowAug over the baselines (up to +1% better than the best of all five baselines). These results again demonstrate that FlowAug is more robust and has better generalizability.

Our CIFAR10.1, CIFAR10-C, and CIFAR100-C experiments demonstrate FlowAug's generalizability to OOD data and validate our approach of using deep decoupled representations for DA.

4.5. Analysis

Our two FlowAug algorithms 1 & 2 both improve over the baselines, and combining the two shows even superior results. Conceptually, we know that the flow model $\mathcal{F}$ can map X to a Gaussian prior distribution (cf. Eq. 1), but not necessarily all the points in the Gaussian distribution would follow the reverse g to a realistic image. Then intuitively, given that z_1, z_2 come from real images, Algorithm 2's in-

	CIFAR-10 best / last	CIFAR10.1 best / last	CIFAR-100 best / last
Standard	95.16 / 95.00	88.70 / 88.25	78.52 / 78.52
Mixup	95.96 / 95.82	89.75 / 88.85	77.91 / 76.99
Cutout	95.94 / 95.60	90.40 / 89.70	78.21 / 78.00
Cutmix	96.17 / 96.04	90.05 / 90.20	78.87 / 78.42
AutoAug	95.82 / 95.47	89.85 / 89.80	78.57 / 78.12
Ours (Mix z)	95.98 / 95.59	90.35 / 89.30	78.58 / 78.07
+ Std	96.15 / 95.73	90.25 / 90.20	78.96 / 78.33
Ours (Uniform on z)	96.12 / 96.11	90.35 / 90.45	79.45 / 79.24
+ Std	96.28 / 96.14	90.95 / 91.15	79.68 / **79.67**
+ Mix z	**96.58 / 96.37**	**91.40 / 91.65**	79.68 / 79.52
+ Std + Mix z	96.44 / 96.31	91.05 / 91.00	79.68 / 79.51
Ours (Trunc Gaussian on z)	96.23 / 96.23	90.15 / 90.25	79.30 / 79.04
+ Std	96.22 / 96.04	90.55 / 89.80	79.62 / 79.62
+ Mix z	96.49 / 96.42	90.05 / 90.25	79.51 / 79.18
+ Std + Mix z	96.53 / 96.29	90.75 / 91.20	**79.99** / 79.54
Trunc Gaussian on ν	95.70 / 95.45	88.70 / 89.25	78.50 / 78.41
Mix ν	94.22 / 93.43	87.55 / 85.50	74.05 / 71.42

Table 2: **Test results in % (best/last epoch).** Although ID and OOD generalization are not our foci, FlowAug consistently outperforms the baseline and we only highlight the top-2 results. Note: we simply apply the models trained from CIFAR-10 to obtain CIFAR10.1 results without fine-tuning.

terpolating of z_1, z_2 can be interpreted as finding an optimal point between two proven optimal points in the space, i.e., Algorithm 2 explores the Gaussian space in an efficient way.

On the other hand, adding a sampled perturbation to z as in Algorithm 1 can stretch the search space to outside of the Gaussian, which brings good performances. It also explains why a combined approach such as Eq. 4 can generally achieve superior performances over the rest since Algorithm 1 and 2 can be complementary.

5. Ablation Studies

To further study global and local representations, we can make some design choices applied to z and ν. In § 3, we assume z and ν carry information about the background and ground truth respectively, and we want to test the assumptions and the generality of Algorithm 1.

5.1. Perturbing Local (ν) or Global (z)?

§ 4 has shown that perturbing z improves generalization and the robustness of models. On the other hand, we can also decode realistic images by perturbing ν (Fig. 2(h)), which we assume affects the prediction. We apply Algorithm 1 & 2 with the same parameters on ν, and the results deteriorate on all datasets by at least 0.5% and up to 7%, suggesting that our assumption about ν's correspondence to the ground truth label is reasonable.

5.2. Does perturbation type matter? A case study of Gaussian vs Uniform distributions

Algorithm 1 uses a truncated Gaussian perturbation, but in fact, we can also add noise sampled from other distributions, such as a Uniform perturbation. To have about the same amount of probability density in the same range as $\mathcal{N}(\mu = 0, \sigma = 0.1)$, we choose $\mathcal{U}(-0.2, 0.2)$ for our study. Table 2 shows that adding uniform noise is comparable to adding truncated Gaussian, and when combined with algorithm 2 and(or) *Standard*, the improvements are top-2, achieving over a 1% gain on CIFAR10 and CIFAR100, and more than a 2% gain on CIFAR10.1. This study suggests the generalization capability of FlowAug on symmetric noise distributions.

	CIFAR10	MacroStd	W-Std	CIFAR100	MacroStd	W-Std
Standard	95.16	2.24	3.45	78.52	12.24	16.44
Cutmix	96.17	1.91	3.34	78.87	12.62	16.87
Ours-alg1	96.23	1.83	2.98	79.45	12.17	16.57
Ours-alg1+Cutmix	**96.34**	**1.65**	**2.91**	**81.57**	**11.20**	**15.34**

Table 3: **Chaining FlowAug with Cutmix on CIFAR10/100.** Our FlowAug has the flexibility of being combined with other works to further enhance performances.

5.3. Can FlowAug be composited with another method?

Composite DA sometimes offer additional benefit to generalization [45, 21, 44]. Thus, we investigate if FlowAug possesses the flexibility of being composited and run an additional experiment combining our simplest Algorithm 1 variation and Cutmix. Tab.3 shows combining FlowAug and Cutmix further mitigates subgroup degradation and also enhances generalization. In addition, on CIFAR100, the test accuracy is better than any methods using Resnet18 to our knowledge. This experiment showcases the possibility of chaining FlowAug with other methods for attaining "SOTA" performances in both mitigating bias or enhancing generalization. We compare FlowAug with other recent composite DA such as AugMix, AugMax in the Appendix.

6. Correlation Analysis on *MacroStd* and Performances

WeightedStd is the most common measure to quantify sensitivity in statistics. However, we want to justify our MacroStd to be a more suitable metric in quantifying subgroup degradation. We perform correlation analyses between accuracy and MacroStd(ours)/WeightedStd. The statistics are summarized in Table 4, and on both CIFAR10 and CIFAR10.1, our metric is a better indicator for sensitivity and accuracy (coefficients the lower the better), which validates the novelty of our metric.

7. Related Works

Representation Learning. Deep learning models' success is generally attributed to their ability to learn complex and meaningful representations [3], and most attempts to

	CIFAR-10 best / last	CIFAR10.1 best / last
MacroStd (ours)	-0.89 / -0.85	-0.83 / -0.88
WeightedStd	-0.78 / -0.77	-0.62 / -0.70

Table 4: **Correlation (↓) between accuracy and sensitivity metrics (best/last epoch).**

learning quality representations require certain inductive biases, for instance, space invariance of CNNs [26]. Of particular interest to our work, generative models such as VAEs [5, 31, 7, 40] enforce constraints such as independent multivariate Gaussain in the latent layers to learn disentangled representations. Our work leverages a model that learns two decoupled representations instead of the factorial ones.

Data Augmentation. DA often helps achieve improved generalization. One line of approach performs low-level basic image operations such as mixing examples [52], or random erasing [50, 11], etc. Another approach uses reinforcement learning to learn the best policy of basic image operations [9, 10]. [30] use causal inference to guide their method and add interventions during the generation process. Our work uses decoupled global representations to isolate spurious correlations and then learn robust correlations to the objects. We refer readers to [14] for recent surveys.

Robustness. Robustness in DNNs has drawn the attention of the community largely since [17, 24]. Multiple lines of research were proposed to study robustness, including using Distributionally Robust Optimization [12, 2], Adversarial Training [29, 6, 36, 49], and certifiable bounds [51, 8]. [1] proposed a scenario where representations learned should be robust in different environments, and [37, 38] suggests that learning Causal Representations can be an ultimate approach to robustness in deep learning. [16] studied the effect of shape and texture to CNNs. Our work is in line with the idea of [1] and studies color as a spurious factor orthogonal to concept of texture [4].

8. Conclusion and Future Work

In this work, we contributed 20,000 non-trivial human annotations in two datasets to reveal the phenomenon of subgroup discrepancy in various (pre)training techniques, and proposed a semantic DA method, **FlowAug** which trains more robust models evaluated on CIFARs and CIFAR-Cs. Additionally, we showed the potential of using disentangled representations for DA by achieving superior generalization performances on both ID and OOD datasets.

We believe our work serves as a leap forward in studies of fairness, robustness, and even causality in DNNs, as we can use CIFAR-Bs to quantify the effect of a hidden bias and we learn that high-level DA is suited to achieve consistent predictions. Last but not least, due to human and computing resource limits, we are not able to scale the labeling effort nor the experiments to larger datasets such as ImageNet, but our results on ImageNet-10 and CIFARs have shown it is an impactful direction and should further enhance the performances.

References

[1] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019.

[2] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[3] 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.

[4] Francesco Bianconi, Antonio Fernández, Fabrizio Smeraldi, and Giulia Pascoletti. Colour and texture descriptors for visual recognition: A historical overview. *Journal of Imaging*, 7(11):245, 2021.

[5] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nicholas Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv: Machine Learning*, 2018.

[6] Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. De-pois: An attack-agnostic defense against data poisoning attacks. In *IEEE Transactions on Information Forensics and Security*, 2021.

[7] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Kristjanson Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[9] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *ArXiv*, abs/1805.09501, 2018.

[10] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2020.

[11] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552, 2017.

[12] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *ArXiv*, 2020.

[13] Yang et. al. OpenOOD: Benchmarking generalized out-of-distribution detection. In *NeurIPS*, 2022.

[14] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy.

A survey of data augmentation approaches for nlp. *ArXiv*, abs/2105.03075, 2021.

[15] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *ArXiv*, abs/1508.06576, 2015.

[16] 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 preprint arXiv:1811.12231*, 2018.

[17] I. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, volume abs/1412.6572, 2015.

[18] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *ArXiv*, abs/1905.08941, 2019.

[19] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[21] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *ArXiv*, abs/1912.02781, 2020.

[22] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2019.

[23] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[24] Alexey Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *ArXiv*, abs/1611.01236, 2017.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting inductive biases of pre-trained models. In *International Conference on Learning Representations*, 2020.

[28] Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard Hovy. Decoupling global and local representations via invertible generative flows. In *Proceedings of the 9th International Conference on Learning Representations (ICLR-2021)*, May 2021.

[29] A. Madry, Aleksandar Makelov, Ludwig Schmidt, D. Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

[30] Chengzhi Mao, Amogh Gupta, Augustine Cha, Hongya Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3946–3955, 2021.

[31] Emile Mathieu, Tom Rainforth, Siddharth Narayanaswamy, and Yee Whye Teh. Disentangling disentanglement. *ArXiv*, abs/1812.02833, 2018.

[32] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

[33] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? 2018. `https://arxiv.org/abs/1806.00451`.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[35] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[36] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Black-box smoothing: A provable defense for pretrained classifiers. In *NeurIPS*, volume abs/2003.01908, 2020.

[37] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634, 2021.

[38] Bernhard Scholkopf and Julius von Kügelgen. From statistical to causal learning. *ArXiv*, abs/2204.00607, 2022.

[39] Dinghan Shen, Ming Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *ArXiv*, abs/2009.13818, 2020.

[40] Zheng sheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7917–7926, 2020.

[41] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[42] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[44] Vasilis Vryniotis. How to train state-of-the-art models using torchvision’s latest primitives, 2021.

[45] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. *Advances in neural information processing systems*, 34:237–250, 2021.

[46] Luyu Wang and Aäron van den Oord. Multi-format contrastive learning of audio representations. *ArXiv*, abs/2103.06508, 2021.

[47] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[48] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm, 2021.

[49] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, 2018.

[50] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Young Joon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019.

[51] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane S. Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *ArXiv*, abs/1906.06316, 2019.

[52] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018.

[53] Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pages 12857–12867. PMLR, 2021.

[54] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.