

CrossNorm and SelfNorm for Generalization under Distribution Shifts

Zhiqiang Tang
Amazon Web Services
zqtang@amazon.com

Yunhe Gao
Rutgers University
yunhe.gao@rutgers.edu

Yi Zhu
Amazon Web Services
yzaws@amazon.com

Zhi Zhang
Amazon Web Services
zhiz@amazon.com

Mu Li
Amazon Web Services
mli@amazon.com

Dimitris Metaxas
Rutgers University
dnm@cs.rutgers.edu

Abstract

Traditional normalization techniques (e.g., Batch Normalization and Instance Normalization) generally and simplistically assume that training and test data follow the same distribution. As distribution shifts are inevitable in real-world applications, well-trained models with previous normalization methods can perform badly in new environments. Can we develop new normalization methods to improve generalization robustness under distribution shifts? In this paper, we answer the question by proposing CrossNorm and SelfNorm. CrossNorm exchanges channel-wise mean and variance between feature maps to enlarge training distribution, while SelfNorm uses attention to recalibrate the statistics to bridge gaps between training and test distributions. CrossNorm and SelfNorm can complement each other, though exploring different directions in statistics usage. Extensive experiments on different fields (vision and language), tasks (classification and segmentation), settings (supervised and semi-supervised), and distribution shift types (synthetic and natural) show the effectiveness. Code is available at <https://github.com/amazon-research/crossnorm-selfnorm>

1. Introduction

Normalization methods, e.g., Batch Normalization [22], Layer Normalization [1], and Instance Normalization [46], play a pivotal role in training deep neural networks by making training more stable and convergence faster, assuming that training and test data come from the same distribution. However, distribution shifts in various real-world scenarios [15, 38, 16] make traditional normalization techniques impractical. For instance, a driving scene segmentation model trained on one city usually does not generalize well to another city. In this paper, we aim to explore how normalization can improve generalization under distribution shifts. Specifically, we tackle the distribution shift problem from

two respects: *enlarging training distribution* and *reducing test distribution*.

First, enlarging the training distribution is not in line with the conventional purpose of normalization which is to stabilize and accelerate training. So, can we employ normalization for a different goal—augmenting training data? Our inspiration comes from a simple observation that exchanging the RGB mean and variance between two images can transfer style between them, as shown in Figure 1 (a). For many tasks such as CIFAR image classification [24], style, encoded by channel-wise mean and variance, is usually less critical in recognizing the object than other information, such as object shape. Therefore, augmenting style is safe enough that content labels remain unchanged. To augment style, we propose CrossNorm, which swaps channel-wise mean and variance of feature maps in training so that the model becomes more robust to changes in appearance.

Even with the augmented training data, a model will still encounter data with unforeseen appearances in deployment. Hence, another question comes: how to make normalization reduce test data distribution, i.e., bridging distribution gaps between training and test data? Similarly, our method is motivated by an observation illustrated in Figure 1 (b). Given one image in different styles, we can reduce the style discrepancy when adjusting the RGB means and variances properly. Intuitively, style recalibration can reduce appearance variance so that training and test data will share more consistent styles. To this end, we propose SelfNorm by using attention [19] to adjust channel-wise mean and variance.

It is interesting to analyze the distinction and connection between CrossNorm and SelfNorm. At first glance, they take opposite actions (style augmentation vs. style reduction). Even so, they use the same tool: channel-wise statistics and pursue the same goal: generalization robustness. Additionally, CrossNorm can increase the capacity of SelfNorm by letting SelfNorm learn from more diverse styles in training. Overall, the key contributions are three-fold:

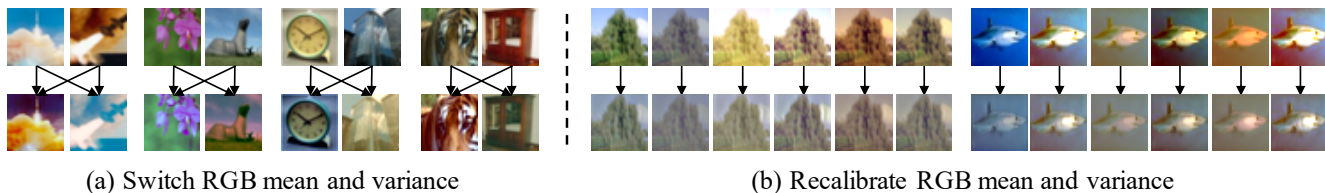


Figure 1: Examples of exchanging (**Left**) and adjusting (**Right**) RGB mean and variance. Swapping the statistics can enrich image styles and thus enlarge training distribution, while recalibrating the statistics properly can encourage style consistency, reducing the train-test distribution gap.

- Unlike previous efforts, we explore a new direction of using feature normalization for generalization under distribution shifts.
- We propose CrossNorm and SelfNorm, two simple yet effective normalization techniques that complement each other to improve generalization robustness.
- CrossNorm and SelfNorm can advance state-of-the-art robustness performance for different fields (vision or language), tasks (classification and segmentation), settings (fully or semi-supervised), and distribution shift types (synthetic and natural).

2. Related Work

Generalization under synthetic distribution shifts.

Following the categorization in [45], distribution shifts are synthetic if they modify existing images to get shifted test datasets. Adversarial examples [11, 33] are one class of synthetic distribution shifts that were widely studied. Recently, various image corruptions [15], as another synthetic type, have attracted increasing attention. To improve the robustness to corruptions, Stylized-ImageNet [10] conducts style augmentation to reduce the texture bias of CNNs. Recently, AugMix [17] trains robust models by mixing multiple augmented images based on random image primitives or image-to-image networks [14]. Adversarial noises training (ANT) [39] and unsupervised domain adaptation [40] can also improve the robustness against corruption.

CrossNorm has two advantages over Stylized-ImageNet, though they are related. First, CrossNorm is efficient as it transfer styles directly in the feature space of target CNNs. However, Stylized-ImageNet relies on external style datasets and pre-trained style transfer models. Second, CrossNorm can advance the performance on both clean and corrupted data, while Stylized-ImageNet hurts clean generalization because external styles can result in massive training distribution shifts. Also, CrossNorm is orthogonal to AugMix and ANT, making it possible for their joint usage.

Generalization under natural distribution shifts.

Compared to synthetic distribution shifts, natural shifts refers to distribution gaps between unmodified data. One type of natural shifts is from video data, where adjacent

frames are perceptually similar for humans, but they usually get inconsistent predictions from deep models [12, 41]. Another type is dataset gaps [37, 2, 3] arising from different factors, e.g., where and when, in collecting two separate datasets. For example, the semantic segmentation dataset GTA5 [38] comes from computer games, which naturally has distribution gaps with realistic segmentation datasets such as Cityscape [4]. To address the distribution gaps, IBN [35] mixes Instance and Batch Normalizations to narrow the distribution gaps. Domain randomization [50] uses style augmentation for domain generalization on segmentation datasets. It suffers from the same issues of Stylized-ImageNet as it also uses pre-trained style transfer models and additional style datasets.

Compared to IBN and domain randomization, SelfNorm can bridge the distribution gaps with style recalibration, and CrossNorm is more efficient and balances better between the source and target datasets’ performance. Beyond the vision field, natural language processing (NLP) applications also face the generalization challenges [16] posed by distribution shifts. Fortunately, SelfNorm and CrossNorm can also improve model robustness in NLP.

Normalization and attention. Batch Normalization [22] is a milestone technique that inspires many following normalization methods such as Instance Normalization [46], Layer Normalization [1], and Group Normalization [48]. Recently, some works integrate attention [19] into normalization. Mode normalization [7] and attentive normalization [26] use attention to weigh a mixture of Batch Normalizations. IEBN [29] uses attention to regulate the batch noises in Batch Normalization. Exemplar Normalization [54] learns to combine multi-type normalizations by attention. By contrast, SelfNorm uses attention with only Instance Normalization. With attention’s help, SelfNorm can emphasize important styles and suppress trivial ones, reducing the distribution gaps caused by appearance discrepancy.

Data augmentation. Data augmentation is an important tool in training deep models. Current popular data augmentation techniques are either label-preserving [5, 30, 18] or label-perturbing [53, 51]. The label-preserving methods usually rely on domain-specific image primitives, e.g., rotation and color, making them inflexible for tasks beyond the

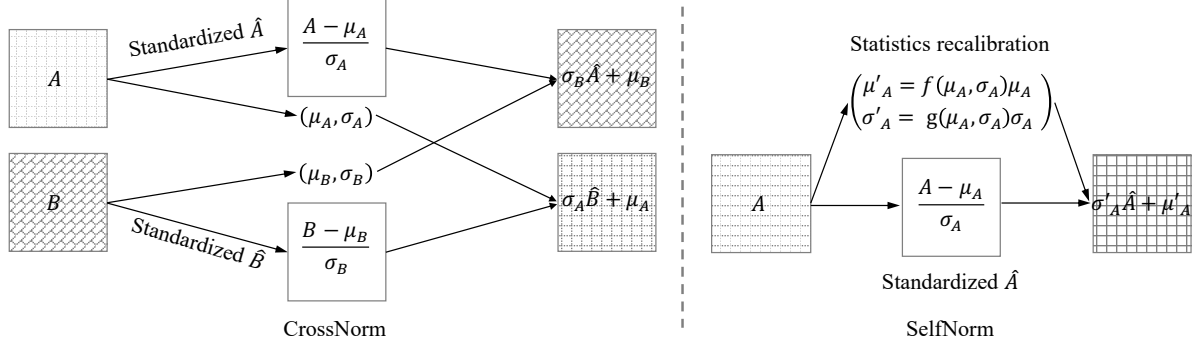


Figure 2: CrossNorm (**left**) and SelfNorm (**right**). CrossNorm swaps the mean and variance between a pair of feature maps, while SelfNorm uses attention to recalibrate a feature map’s statistics.

vision field. The label-perturbing techniques mainly work for classification and may have trouble in broader applications, e.g., segmentation. CrossNorm, as a data augmentation method, is readily applicable to diverse fields (vision and language) and tasks (classification and segmentation). The goal of CrossNorm is to boost generalization robustness under distribution shifts, which is also different from many former data augmentation methods.

3. CrossNorm and SelfNorm

This section elaborates CrossNorm, SelfNorm, their relation, and their application in deep neural networks. Before that, we introduce some preliminaries regarding Instance Normalization [46] and the style concept.

3.1. Preliminary

Instance Normalization. Technically, SelfNorm and CrossNorm share the same origin: Instance Normalization [46]. In 2D CNNs, each instance has C feature maps of size $H \times W$. Given a feature map $\mathcal{A} \in \mathbb{R}^{H \times W}$, Instance Normalization first normalizes the feature map and then conducts affine transformation:

$$\gamma \frac{\mathcal{A} - \mu_{\mathcal{A}}}{\sigma_{\mathcal{A}}} + \beta, \quad (1)$$

where $\mu_{\mathcal{A}}$ and $\sigma_{\mathcal{A}}$ are the mean and standard deviation; γ and β denotes learnable affine parameters. As shown in Figure 1 and also pointed out by the style transfer practices [9, 47, 21], $\mu_{\mathcal{A}}$ and $\sigma_{\mathcal{A}}$ can encode some style information.

Style Concept. In this paper, the style concept refers to a family of weak cues associated with the semantic content of interest. For instance, the image style in object recognition can include many appearance-related factors such as color, contrast, and brightness. Style sometimes may help in decision-making, but the model should rely more on vital content cues to become robust. To reduce its bias rather than discarding it, we use CrossNorm with probability in training. The insight beneath CrossNorm is that each instance,

or feature map, has its unique style. Further, style cues are not equally important. For example, the yellow color seems more useful than other style cues in recognizing an orange. In light of this, the intuition behind SelfNorm is that attention may help emphasize essential styles and suppress trivial ones. Although we use the channel-wise mean and variance to modify styles, we do not assume that they are sufficient to represent all style cues. Better style representations are available with more complex statistics [27] or even style transfer models [47, 21]. We choose the first and second-order statistics mainly because they are simple, efficient to compute, and can naturally connect normalization to generalization robustness.

3.2. CrossNorm

To enlarge training distribution, CrossNorm exchanges $\mu_{\mathcal{A}}$ and $\sigma_{\mathcal{A}}$ of channel \mathcal{A} with $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ of channel \mathcal{B} , i.e., changing β and γ to each other’s μ and σ , seen in Figure 2:

$$\sigma_{\mathcal{B}} \frac{\mathcal{A} - \mu_{\mathcal{A}}}{\sigma_{\mathcal{A}}} + \mu_{\mathcal{B}} \quad \sigma_{\mathcal{A}} \frac{\mathcal{B} - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}} + \mu_{\mathcal{A}}, \quad (2)$$

where \mathcal{A} and \mathcal{B} seem to normalize each other, hence CrossNorm. CrossNorm is motivated by the key observation that a target dataset, such as a classification dataset, has rich, though subtle, styles. Specifically, each instance, or even every channel, has its unique style. CrossNorm, turned on with some probability in training, can perform efficient style augmentation and thus enlarge training distribution. To further diversify styles, we investigate different feature map choices, resulting in different CrossNorm variants.

1-instance mode. For 2D CNNs, given one instance $\mathcal{X} \in R^{C \times H \times W}$, CrossNorm can exchange statistics between its C channels:

$$\{(A, B) \in (\mathcal{X}_{i,:}, \mathcal{X}_{j,:}) \mid i \neq j, 0 < i, j < C\}, \quad (3)$$

where \mathcal{A} and \mathcal{B} refer to the channel pair in Equation 2.

2-instance mode. If two instances $\mathcal{X}, \mathcal{Y} \in R^{C \times H \times W}$ given, CrossNorm can swap statistics between their corre-

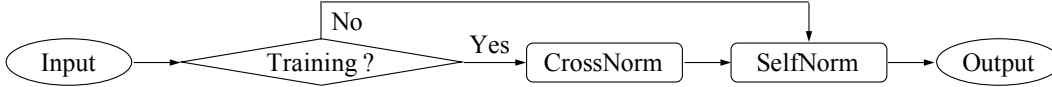


Figure 3: Flowchart for CrossNorm and SelfNorm. CrossNorm works only in training, while SelfNorm learns in training and functions in testing.

sponding channels, i.e., \mathcal{A} and \mathcal{B} become:

$$\{(\mathcal{A}, \mathcal{B}) \in (\mathcal{X}_{i,:}, \mathcal{Y}_{i,:}) \mid 0 < i < C\}. \quad (4)$$

Compared to 1-instance CrossNorm, 2-instance CrossNorm considers instance-level style instead of channel-level.

Crop. Moreover, distinct spatial regions probably have different mean and variance statistics. To promote the style diversity, we propose to crop regions for CrossNorm:

$$\{(\mathcal{A}, \mathcal{B}) \in (\text{crop}(\mathcal{A}), \text{crop}(\mathcal{B})) \mid r_{\text{crop}} \geq t\} \quad (5)$$

where the crop function returns a square with area ratio r no less than a threshold t ($0 < t \leq 1$). The whole channel is a special case in cropping. There are three cropping choices: content only, style only, and both. For content cropping, we crop A only when we use its standardized feature map. In other words, no cropping applies to A when it provides its statistics to B. Cropping both means cropping A and B no matter we employ their standardized feature map or statistics. The cropping strategy can produce diverse styles for both the 1-instance and 2-instance CrossNorms.

3.3. SelfNorm

To bridge the train-test distribution gap, SelfNorm replaces β and γ in Equation 1 with recalibrated mean $\mu'_A = f(\mu_A, \sigma_A)\mu_A$ and standard deviation $\sigma'_A = g(\mu_A, \sigma_A)\sigma_A$, as illustrated in Figure 2, where f and g are the attention functions. The adjusted feature map becomes:

$$\sigma'_A \frac{\mathcal{A} - \mu_A}{\sigma_A} + \mu'_A. \quad (6)$$

As f and g learn to scale μ_A and σ_A based on themselves, \mathcal{A} normalizes itself by self-gating, hence *SelfNorm*. SelfNorm is inspired by the fact that attention can help the model emphasize informative features and suppress less useful ones. In terms of recalibrating μ_A and σ_A , SelfNorm expects to highlight the discriminative styles shared by training and test distributions and understate trivial one-sided styles. In practice, we use two fully connected (FC) networks to wrap attention functions f and g , respectively. Each network is efficient as the input and output are two and one scalars.

Note that SelfNorm is different from SE [19], though they use similar attention. First, SE models the interdependency between channels, while SelfNorm deals with each channel independently. Second, SelfNorm learns to recalibrate channel-wise mean and variance instead of channel

features in SE. Also, a SelfNorm unit, with complexity $O(C)$, is more lightweight than a SE one, of $O(C^2)$, where C denotes the channel number.

3.4. Relation and Application

Unity of opposites. CrossNorm and SelfNorm both start from Instance Normalization but head in opposite directions. CrossNorm transfers statistics between channels, enriching the combinations of standardized features (zero-mean and unit-variance) and statistics. In contrast, SelfNorm recalibrates statistics to focus on only necessary styles, reducing standardized features and statistics mixtures' diversity. They perform opposite operations mainly because they target different stages. CrossNorm functions only in training, whereas SelfNorm dedicates to style recalibration during testing. Note that SelfNorm is a learnable module, requiring training to work. Figure 3 shows the flowchart of CrossNorm and SelfNorm. Despite these differences, they both can facilitate generalization under distribution shifts. Further, CrossNorm can boost SelfNorm's performance because its style augmentation can prevent SelfNorm from overfitting to specific styles. Overall, the two seemingly opposed methods form a unity of using normalization statistics to advance generalization robustness.

Modular design. CrossNorm and SelfNorm can naturally work in the feature space, making it flexible to plug them into many network locations. Two questions arise: how many units are necessary and where to place them? To simplify the questions, we turn to the modular design by embedding them into a network cell. For example, in ResNet [13], we put them into a residual module. The search space significantly shrinks for the limited positions in a residual module. We will investigate the position choices in experimental ablation study. The modular design allows using multiple CrossNorms and SelfNorms in a network. We will show in the ablation study that accumulated style recalibrations are helpful for model robustness.

4. Experiment

We evaluate CrossNorm (CN) and SelfNorm (SN) in various distribution shifts settings.

Image classification datasets. We use benchmark datasets: CIFAR-10 [24], CIFAR-100, and ImageNet[8]. To evaluate the model robustness against corruption, we use the datasets: CIFAR-10-C, CIFAR-100-C, and ImageNet-C [15]. These datasets are the original test data poisoned by

Table 1: mCE (%) on CIFAR-10-C and CIFAR-100-C. CNSN obtains lower errors than most previous methods with different backbones. Albeit some higher errors than AugMix, it is more general without relying on the image primitives, e.g., rotation, in AugMix. As CNSN and AugMix are orthogonal, their joint usage brings new state-of-the-art results.

CIFAR-10-C	Basic	Cutout	Mixup	CutMix	AutoAug	AdvTr.	AugMix	CN	SN	CNSN	CNSN+AugMix
AllConvNet	30.8	32.9	24.6	31.3	29.2	28.1	15.0	26.0	24.0	17.2	11.8
DenseNet	30.7	32.1	24.6	33.5	26.6	27.6	12.7	24.7	22.0	18.5	10.4
WideResNet	26.9	26.8	22.3	27.1	23.9	26.2	11.2	21.6	20.8	16.9	9.9
ResNeXt	27.5	28.9	22.6	29.5	24.2	27.0	10.9	22.4	21.5	15.7	9.1
Mean	29.0	30.2	23.5	30.3	26.0	27.2	12.5	23.7	22.1	17.0	10.3
CIFAR-100-C	Basic	Cutout	Mixup	CutMix	AutoAug	AdvTr.	AugMix	CN	SN	CNSN	CNSN+AugMix
AllConvNet	56.4	56.8	53.4	56.0	55.1	56.0	42.7	52.2	50.3	42.8	36.8
DenseNet	59.3	59.6	55.4	59.2	53.9	55.2	39.6	55.4	53.9	48.5	37.0
WideResNet	53.3	53.5	50.4	52.9	49.6	55.1	35.9	48.8	47.4	43.7	33.4
ResNeXt	53.4	54.6	51.4	54.1	51.3	54.4	34.9	47.0	47.6	40.8	30.8
Mean	55.6	56.1	52.6	55.5	52.5	55.2	38.3	50.9	49.8	43.5	34.7

Table 2: Clean error and mCE (%) of ResNet50 trained 90 epochs on ImageNet. CNSN, using simple general statistics, achieves comparable performance as domain-specific AugMix. Jointly applying CNSN with AugMix and IBN can produce the lowest clean and corruption errors.

Aug.	Clean	Noise			Blur				Weather				Digital				mCE
		Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	
Standard	23.9	79	80	82	82	90	84	80	86	81	75	65	79	91	77	80	80.6
Patch Uniform	24.5	67	68	70	74	83	81	77	80	74	75	62	77	84	71	71	74.3
Random AA*	23.6	70	71	72	80	86	82	81	81	77	72	61	75	88	73	72	76.1
MaxBlur pool	23.0	73	74	76	74	86	78	77	77	72	63	56	68	86	71	71	73.4
SIN	27.2	69	70	70	77	84	76	82	74	75	69	65	69	80	64	77	73.3
AugMix*	23.4	66	66	66	69	80	65	68	72	72	66	60	63	78	66	71	68.4
CN	23.3	73	75	75	78	89	79	82	79	75	66	61	69	97	69	74	75.1
SN	23.7	69	71	69	77	87	77	80	75	77	70	61	73	83	61	71	73.8
CNSN	23.3	66	67	65	77	89	76	80	72	72	67	59	47	83	62	72	69.7
CNSN+AugMix	22.3	61	62	60	70	77	62	68	62	65	63	55	43	73	55	66	62.8

15 everyday image corruptions from 4 general types: noise, blur, weather, and digital. Each noise has 5 intensity levels when injected into images. In addition, we conduct domain adaptation experiments with Office-31 including 4652 images and 31 categories from 3 domains: Amazon (A) (amazon.com images), Webcam (W) (web camera images) and DSLR (D) (digital SLR camera images).

Image segmentation datasets. We further validate our method using a domain generalization setting, where the models are trained in a source domain and tested on a unforeseen target domain. We use the synthetic dataset Grand Theft Auto V (GTA5) [38] as the source domain and generalize to the real-world dataset Cityscapes [4]. GTA5 has the training, validation, and test divisions of 12,403, 6,382, and 6,181, more than those of 2,975, 500, and 1,525 from Cityscapes. Despite the differences, their pixel categories are compatible with each other, allowing to evaluate models' generalization capability from one to another. **Sentiment classification datasets.** Besides vision tasks, we demonstrate that our method can work well on NLP

tasks. In particular, we use the cross-dataset binary sentiment classification setting, where a model is trained on the IMDB dataset [32] and then tested on the SST-2 dataset [42]. The IMDB dataset collects highly polarized full-length lay movie reviews with 25,000 positive and 25,000 negative reviews. The SST-2, with 9613 and 1821 reviews for training and testing, is also a binary sentiment dataset but instead contains pithy expert movie reviews.

Metric. For image classification, we use test errors to measure robustness. Given corruption type c and severity s , let E_s^c denote the test error. For CIFAR datasets, we use the average over 15 corruptions and 5 severities: $1/75 \sum_{c=1}^{15} \sum_{s=1}^5 E_{c,s}$. In contrast, for ImageNet, we normalize the corruption errors by those of AlexNet [25]: $1/15 \sum_{c=1}^{15} (\sum_{s=1}^5 E_s^c / \sum_{s=1}^5 E_{c,s}^{AlexNet})$. The above two metrics follow the convention [17] and are denoted as mean corruption errors (mCE) whether they are normalized or not. Different from classification, segmentation uses the mean Intersection over Union (mIoU) of all categories. For sentiment classification, we report accuracy as its metric.

Table 3: Semi-supervised results on CIFAR-10 with WideResNet-28-2. We use FixMatch with weak augmentation (WA) or strong RandAugment (RA). In either case, CN substantially reduces both clean and corruption errors.

		FixMatch+WA		FixMatch+RA	
		Baseline	+CN	Baseline	+CN
250 labels	Clean err(%)	57.7	51.4	10.9	7.4
	mCE(%)	62.9	56.5	23.4	16.8
1000 labels	Clean err(%)	16.7	13.1	6.7	5.8
	mCE(%)	37.7	29.6	19.8	15.5

Hyper-parameters. In the experiments, an attention function in SN uses one fully connected layer, followed by Batch Norm and a sigmoid layer. We put CN ahead of SN, and plug them into every cell in a network, e.g., each residual module in a ResNet. During training, we turn on only some CNs with probability to avoid excessive data augmentation. Unless specified, 2-instance CN is used with cropping. We sample the cropping bounding box ratio uniformly and set the threshold $t = 0.1$. Refer to the appendix for details regarding CN’s active number and probability.

4.1. Robustness against Unseen Corruptions

Supervised training on CIFAR. Following AugMix [17], we use four different backbones: an All Convolutional Network [44], a DenseNet-BC ($k = 12, d = 100$) [20], a 40-2 Wide ResNet [52], and a ResNeXt-29 (32×4) [49]. CN and SN are plugged into their cell modules. We use the same hyper-parameters in the AugMix Github repository¹.

According to Table 1, individual CN and SN can outperform most previous approaches on robustness against unseen corruptions and combining them can decrease the mean error by $\sim 12\%$ on both CIFAR-10-C and CIFAR-100-C. As the corruptions mainly change image textures, one possible explanation is that CN and SN, through style augmentation and recalibration, may help reduce the texture sensitivity and bias, making the classifiers more robust to unseen corruptions. Also, CN and SN are orthogonal to AugMix, which relies on domain-specific image operations. Their joint application can continue to lower the mCEs by 2.2% and 3.6% on top of AugMix.

Supervised training on ImageNet. Following the AugMix Github repository, we train a ResNet-50 for 90 epochs with weight decay $1e-4$. The learning rate starts from 0.1, divided by 10 at epochs 30 and 60. Note that AugMix reports the results of 180 epochs in their paper. For a fair comparison, we also train it 90 epochs in our experiments. Different from CIFAR experiments, we apply CN only to the image space. Besides, we also add Instance-batch normalization (IBN) [35] in the final combination with AugMix. It was initially designed for domain generalization but can also boost model robustness against corruption.

¹<https://github.com/google-research/augmix>

Table 4: Image segmentation results (mIoU) on GTA5-Cityscapes domain generalization using a FCN with ResNet50. CN and SN are comparable to domain randomization (DR) and IBN on the target domain (Cityscapes). Combining CN and SN can achieve state-of-the-art results.

Methods	Baseline	IBN	DR	CN	SN	CNSN
Source	63.7	64.2	49.0	61.2	64.6	63.5
Target	21.4	29.6	32.7	32.0	29.9	36.5

Table 2 gives the results on ImageNet. We can observe that both clean and corrupted errors decrease when applying CN and SN separately. Their joint usage can make the clean and corruption errors drop by 0.6% and 10.9% simultaneously, closing the gap with AugMix. Moreover, applying CN and SN on top of AugMix can significantly lower its clean and corruption errors by 1.1% and 5.6%, respectively, achieving state-of-the-art performance. IBN also makes some contributions here since it is complementary to other components.

Semi-supervised training on CIFAR. Apart from supervised training, we also evaluate CN in semi-supervised learning. Following state-of-the-art FixMatch [43] setting, we train a 28-2 Wide ResNet for 1024 epochs on CIFAR-10. The SGD optimizer applies with Nesterov momentum 0.9, learning rate 0.03, and weight decay $5e-4$. The probability threshold to generate pseudo-labels is 0.95, and the weight for unlabeled data loss is 1. We sample 250 and 4,000 labeled data with random seed 1, leaving the rest as unlabeled data. In each experiment, we apply CrossNorm to either all data or only unlabeled data and choose the better one. Our experiments use the Pytorch FixMatch implementation².

Table 3 shows the semi-supervised results. Whether FixMatch uses only weak random flip and crop augmentations or strong RandAugment [6], CN can always decrease both the clean and corruption errors, demonstrating its effectiveness in semi-supervised training. Especially with the help of CN, training with 250 labels even has 3% lower corruption error than with 1000 labels, according to the columns 5 and 6. Additionally, two points are noteworthy here. First, we try FixMatch with only weak augmentations to simulate more general situations. For new areas other than natural images, humans may have the limited expertise to design advanced augmentation operations. Fortunately, CN is area-agnostic and easily applicable to such situations. Moreover, previous semi-supervised methods mainly focus on clean generalization. Here we introduce corruption robustness as another metric for comprehensive evaluation.

4.2. Generalization from Synthetic to Realistic Data

Setup. We perform cross-dataset generalization from GTA5 (synthetic) to Cityscapes (realistic), following the

²<https://github.com/kekmodel/FixMatch-pytorch>

Table 5: Sentiment classification accuracy with cross-dataset generalization (IMDb→SST-2) using GloVe embedding and ConvNets. CN and SN work in the NLP field.

Methods	Baseline	CN	SN	CNSN
Source	85.7	85.1	86.3	85.9
Target	71.9	73.0	73.9	74.9

setting of IBN [35]. It uses 1/4 training data in GTA5 to match the data scale of Cityscapes. We train the FCN [31] with ResNet50 backbone in source domain GTA5 for 80 epochs with batch size 16. The network is initialized with ImageNet pre-trained weights. We test the trained model on both the source and target domains. The training uses random scaling, flip, rotation, and cropping (713×713) for data augmentation. We use the 2-instance CN with style cropping in this setting. Besides, we re-implement the domain randomization [50] and make the training iterations the same as ours. It transfers the synthetic images to 15 auxiliary domains with ImageNet image styles.

Results. Table 4 shows that CN and SN both can substantially increase the segmentation accuracy on the target domain by 10.6% and 8.5%. CN performs style augmentation to make the model focus more on domain-invariant features. SN learns to highlight the discriminative styles that are likely to share across domains. CN and SN get comparable generalization performance as state-of-the-art domain randomization [50] and IBN [35]. However, CN significantly outperforms the domain randomization method by 12.2% on the source accuracy because the domain randomization transfers external styles to the source training data, causing dramatic distribution shifts. Moreover, combining CN and SN gives the best generalization performance while still maintaining high source accuracy.

4.3. Cross-dataset Generalization in NLP

Setup. To show that CN and SN are independent of application fields, we also evaluate their generalization robustness on a binary sentiment classification setup in the NLP area. The model is trained on the IMDb dataset and tested on the SST-2 dataset. Follow the setting of [16], we use the GloVe [36] word embedding and Convolutional Neural Networks (ConvNets) [23] as the classification model. We use the implementation of ConvNets in this repository³. The convolutional layers with three kernel sizes (3,4,5) are used to extract $n - gram$ features within the review texts. CN and SN units are placed between the embedding layer and the convolutional layers. We use the Adam optimizer and train the model for 20 epochs.

Results. From Table 5, we can find that SN improves the performance in both the source and target domains by 0.6% and 2.0%. CN can also increase target accuracy with-

³<https://github.com/bentrevett/pytorch-sentiment-analysis>

Table 6: Domain adaptation accuracy on Office-31. CN outperforms AdaBN in 4 settings and mean accuracy.

	A→W	D→W	W→D	A→D	D→A	W→A	Avg.
Source Only	68.4	96.7	99.3	68.9	62.5	60.7	76.1
AdaBN	74.1	97.1	99.5	72.3	61.9	61.2	77.7
CN	77.6	98.0	100.0	77.9	60.9	61.0	79.2

out much degradation in the source domain. Combining them gives a 3.0% boost of target accuracy. This experiment indicates that CN and SN can also work in the NLP area, not limited to the vision tasks. Despite the lack of intuitive explanations as for the image data, the mean and variance statistics in NLP data are also useful in facilitating generalization under distribution shifts.

4.4. Domain Adaptation

Setup. In addition to generalization, we evaluate CN on domain adaptation. In particular, we compare CN with closely related AdaBN [28] in 6 adaptation settings of the Office-31 dataset. CN is applied to both image and feature space without cropping. We follow a Github repo⁴ to use ResNet50, 100 epochs, batch size 32, and Adam optimizer with constant learning rate $1e-5$ and weight decay $2.5e-5$.

Results. According to Table 6, CN improves 3.1% average accuracy over the baseline, which almost doubles the improvements (1.6%) of AdaBN. We notice that CN and AdaBN make accuracy increase slightly or even decrease in the D→A and W→A settings. This may be due to that D (498 labeled images) and W (795 labeled images) domains have much fewer images than A domain (2817 images).

4.5. Visualization

Apart from the quantitative comparisons, we also provide some visualization results of CN and SN to better understand their effects. To this end, we map the feature changes made by CN and SN back to image space by inverting the feature representations [34]. For detailed experimental settings, refer to the appendix.

In visualizing CN, we pair one content image with multiple style images for better illustration. We first forward them to get their feature representations at a chosen position. Then we compute standardized features from the content image representation and means and variances of the style image representations. The optimization starts from the content image and tries to fit its representation to the target one mixing the standardized features with different means and variances. Figure 4 shows diverse style changes made by CN. The style changes become more local and subtle as CN moves deeper in the network.

To visualize SN at a network location, we first forward an image to obtain the target representation immediately af-

⁴<https://github.com/fazilaltinel/ADDA.PyTorch-resnet>

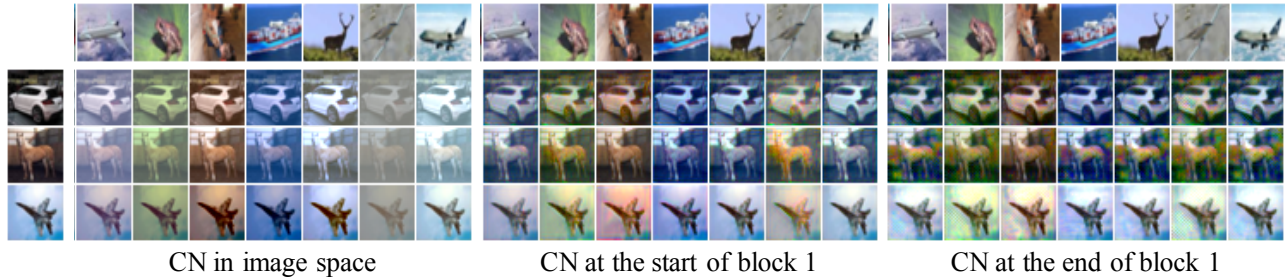


Figure 4: CN visualization at image level (**Left**), the head (**Middle**) and tail (**Right**) of block 1 in WideResNet-40-2. Both the content (**Row**) and style (**Column**) images are from CIFAR-10. The style rendering changes from global to local as CN gets deeper in the network.

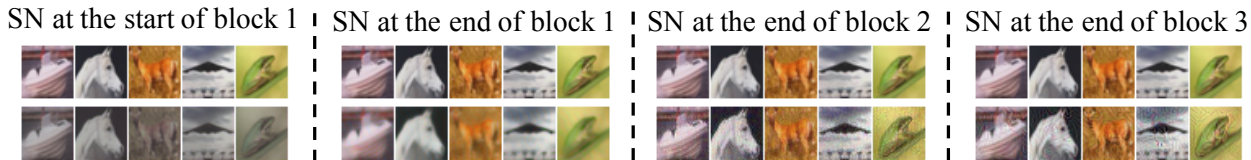


Figure 5: Visualizing 4 single SNs in WideResNet-40-2 by comparing images before (**Top**) and after (**Bottom**) SN. The left two, lying in shallow locations, can adjust styles by suppressing color and adding blur. As SN goes deeper, the recalibration effect becomes subtle because the statistics of high-level features do not directly connect to low-level visual cues.

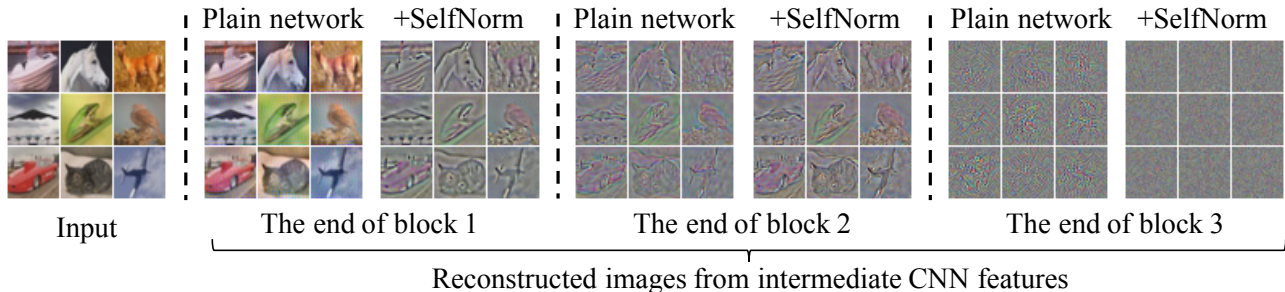


Figure 6: Visualizing accumulated SNs in WideResNet-40-2 by comparing reconstructed images from intermediate features. SNs in block 1 can wash away much style information preserved in the vanilla network. Similarly, the plain network’s final representation retains some high-frequency signals which are suppressed by SNs.

ter the SN. Then we turn off the chosen SN and optimize the original image to make its representation fit the target one. In this way, we can examine a SN’s effect by observing the changes in image space. As shown in Figure 5, SN can primarily reduce the contrast and color at the first network block. The effect becomes more subtle as SN goes deeper into the network. One possible explanation is that the high-level representations lose too many low-level details, making it difficult to visualize the changes.

In addition to visualizing individual SNs, it is also interesting to see their compound effect. To this end, we reconstruct an image from random noises by matching its representation with a given one. The reconstructed image can show what information is preserved by the feature representation. By comparing two reconstructed images from a network with or without SN, we can observe the joined recalibration effects of SNs before a selected location. From Figure 6, we can find SNs in the first two network blocks

can suppress much style information and preserve object shapes. The reconstructions from block 3 do not look visually informative due to the high-level abstraction. Even so, SNs can restrain the high-frequency signals kept in the vanilla network.

5. Conclusion

This paper has explored how normalization can enhance generalization under distribution shifts and presented CN and SN, two simple, effective, and complementary normalization techniques. Their extensive applications can shed light on developing general methods applicable to multiple fields, such as vision and language, and broad synthetic and natural distribution shift circumstances. Given the simplicity of CN and SN, we believe there is substantial room for improvement. One possible direction is to explore better style representations since the current channel-wise mean and variance are not optimal to encode diverse styles.

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Danny Gutfreund, Joshua Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. 2019.
- [3] Ali Borji. Objectnet dataset: Reanalysis and correction. *arXiv preprint arXiv:2004.02042*, 2020.
- [4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pages 113–123, 2019.
- [6] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. *arXiv preprint arXiv:1909.13719*, 2019.
- [7] Lucas Deecke, Iain Murray, and Hakan Bilen. Mode normalization. *arXiv preprint arXiv:1810.05466*, 2018.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [9] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.
- [10] 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. In *Proceedings of the International Conference on Learning Representations*, 2019.
- [11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [12] Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathon Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.
- [15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
- [16] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedziec, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.
- [17] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the International Conference on Learning Representations*, 2020.
- [18] Daniel Ho, Eric Liang, Ion Stoica, Pieter Abbeel, and Xi Chen. Population based augmentation: Efficient learning of augmentation policy schedules. *arXiv preprint arXiv:1905.05393*, 2019.
- [19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.
- [22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [23] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
- [24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.
- [26] Xilai Li, Wei Sun, and Tianfu Wu. Attentive normalization. *arXiv preprint arXiv:1908.01259*, 2019.
- [27] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, pages 386–396, 2017.
- [28] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.
- [29] Senwei Liang, Zhongzhan Huang, Mingfu Liang, and Haizhao Yang. Instance enhancement batch normalization: An adaptive regulator of batch noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4819–4827, 2020.
- [30] Sungbin Lim, Ildoo Kim, Taesup Kim, Chihyeon Kim, and Sungwoong Kim. Fast autoaugment. In *NeurIPS*, pages 6662–6672, 2019.
- [31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [32] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors

- for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.
- [33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [34] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.
- [35] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018.
- [36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.
- [38] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.
- [39] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. *arXiv preprint arXiv:2001.06057*, 2020.
- [40] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33, 2020.
- [41] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019.
- [42] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
- [43] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.
- [44] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- [45] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.
- [46] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.
- [48] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [49] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [50] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2100–2110, 2019.
- [51] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision (ICCV)*, 2019.
- [52] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [53] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- [54] Ruimao Zhang, Zhanglin Peng, Lingyun Wu, Zhen Li, and Ping Luo. Exemplar normalization for learning deep representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2020.