On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness

Anonymous Author(s)
Affiliation
Address
email

Abstract

Invariance to a broad array of image corruptions, such as warping, noise, or color shifts, is an important aspect of building robust models in computer vision. Recently, several new data augmentations have been proposed that significantly improve performance on ImageNet-C, a benchmark of such corruptions. However, there is still a lack of basic understanding on the relationship between data augmentations and test-time corruptions. To this end, we develop a feature space for image transforms, and then use a new measure in this space between augmentations and corruptions called the Minimal Sample Distance to demonstrate a strong correlation between similarity and performance. We then investigate recent data augmentations and observe a significant degradation in corruption robustness when the test-time corruptions are sampled to be perceptually dissimilar from ImageNet-C in this feature space. Our results suggest that test error can be improved by training on perceptually similar augmentations, and data augmentations may not generalize well beyond the existing benchmark. We hope our results and tools will allow for more robust progress towards improving robustness to image corruptions.

1 Introduction

Robustness to distribution shift, *i.e.* when the train and test distributions differ, is an important feature of practical machine learning models. Among many forms of distribution shift, one particularly relevant category for computer vision are image corruptions. For example, test data may come from sources that differ from the training set in terms of lighting, camera quality, or other features. Post-processing transforms, such as photo touch-up, image filters, or compression effects are commonplace in real-world data. Models developed using clean, undistorted inputs typically perform dramatically worse when confronted with these sorts of image corruptions [8, 13]. The subject of corruption robustness has a long history in computer vision [1, 6, 27] and recently has been studied actively with the release of benchmark datasets such as ImageNet-C [13].

One particular property of image corruptions is that they are low-level distortions in nature. Corruptions are transformations of an image that affect structural information such as colors, textures, or geometry [5] and are typically free of high-level semantics. Therefore, it is natural to expect that *data augmentation* techniques, which expand the training set with random low-level transformations, can help learn robust models. Indeed, data augmentation has become a central technique in several recent methods [14, 19, 24] that achieve large improvements on ImageNet-C and related benchmarks.

One caveat for data augmentation based approaches is the test corruptions are expected to be *unknown* at training time. If the corruptions are known, they may simply be applied to the training set as data augmentations to trivially adapt to the test distribution. Instead, an ideal robust model needs to be robust to *any* valid corruption, including ones unseen in any previous benchmark. Of course, in practice the robustness of a model can only be evaluated approximately by measuring its corruption

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

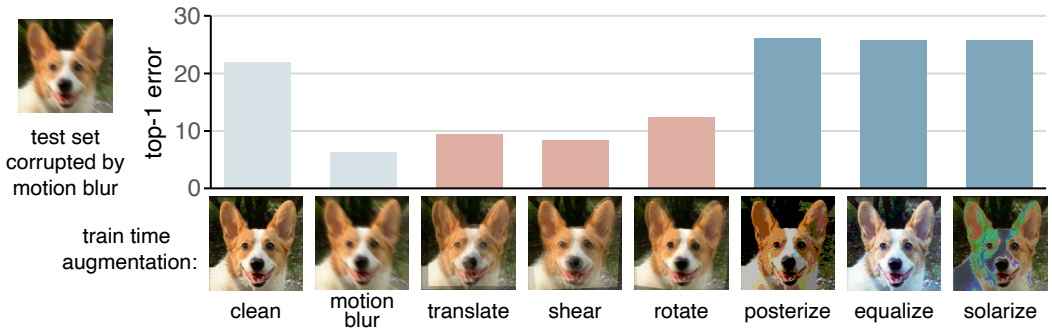

Figure 1: **A toy experiment.** We train multiple models on CIFAR-10 [17] using different augmentation schemes. Each scheme is based on a single basic image transformation type and enhanced by overlaying random instantiations of the transformation for each input image following Hendrycks et al. [14]. We compare these models on the CIFAR-10 test set corrupted by the motion blur, a corruption used in the ImageNet-C corruption benchmark [13]. None of the augmentation schemes contains motion blur; however, the models trained with geometric-based augmentations significantly outperform the baseline model trained on the clean images while color-based augmentations show no gains. We note the geometric augmentations can produce a result visually similar to a blur by overlaying copies of shifted images[2].

error on a representative corruption benchmark. To avoid trivial adaptation to the benchmark, recent works manually exclude test corruptions from the training augmentations. However, with a toy experiment presented in Figure 1, we argue that this strategy alone might not be enough and that visually similar augmentation outputs and test corruptions can lead to significant benchmark improvements even if the exact corruption transformations are excluded.

This observation raises two important questions. One, *how exactly does the similarity between train time augmentations and corruptions of the test set affect the error?* And two, if the gains are due to the similarity, they may not translate into better robustness to other possible corruptions, so *how well will data augmentations generalize beyond a given benchmark?* In this work, we take a step towards answering these questions, with the goal of better understanding the relationship between data augmentation and test-time corruptions. Using a feature space on image transforms and a new measure called Minimal Sample Distance (MSD) on this space, we are able to quantify the distance between augmentation schemes and classes of corruption transformation. With our approach, we empirically show an intuitive yet surprisingly overlooked finding:

Augmentation-corruption perceptual similarity is a strong predictor of corruption error.

Based on this finding, we perform additional experiments to show that data augmentation aids corruption robustness by increasing perceptual similarity between a (possibly small) fraction of the training data and the test set. To further support our claims, we introduce a set of new corruptions, called CIFAR/ImageNet-$\overline{\text{C}}$, to test the degree to which common data augmentation methods generalize from the original the CIFAR/ImageNet-C. To choose these corruptions, we expand the set of natural corruptions and sample new corruptions that are far away from CIFAR/ImageNet-C in our feature space for measuring perceptual similarity. We then demonstrate that augmentation schemes designed specifically to improve robustness show significantly degraded performance on CIFAR/ImageNet-$\overline{\text{C}}$. Some augmentation schemes still show some improvement over baseline, which suggests meaningful progress towards general corruption robustness is being made, but different augmentation schemes exhibit different degrees of generalization capability. As an implication, caution is needed for fair robustness evaluations when additional data augmentation is introduced.

These results suggest a major challenge that is often overlooked in the study of corruption robustness: *generalization is often poor*. Since perceptual similarity can predict performance, for any fixed finite set of test corruptions, improvements on that set may generalize poorly to dissimilar corruptions. We hope that these results, tools, and benchmarks will help researchers better understand *why* a given augmentation scheme has good corruption error and whether it should be expected to generalize to

[2]Example transforms are for illustrative purpose only and are exaggerated. Base image © Sehee Park.

dissimilar corruptions. On the positive side, our experiments show that *generalization does emerge* among perceptually similar transforms, and that only a *small fraction* of sampled augmentations need to be similar to a given corruption. Section 6 discusses these points in more depth.

2 Related Work

Corruption robustness benchmarks and analysis. ImageNet-C [13] is a corruption dataset often used as a benchmark in robustness studies. Other corruption datasets [15, 26] collect corrupted images from real world sources and thus have a mixture of semantic distribution shifts and perceptual transforms. Corruption robustness differs from adversarial robustness [29], which seeks invariance to small, worst case distortions. One notable difference is that improving corruption robustness often slightly improves regular test error, instead of harming it. Yin et al. [35] analyzes corruption robustness in the context of transforms' frequency spectra; this can also influence corruption error independently from perceptual similarity. Here we study the relationship between augmentations and corruptions more generically, and explore the relationship between perceptual similarity and generalization to new corruptions. Dao et al. [3] and Wu et al. [33] study the theory of data augmentation for regular test error. Hendrycks et al. [15] and Taori et al. [31] study how the performance on synthetic corruption transforms generalizes to performance on corruption datasets collected from the real world. Here we do not address this issue directly but touch upon it in the discussion.

Improving corruption robustness. Data augmentations designed to improve robustness include AugMix [14], which composites common image transforms, Patch Gaussian [19], which applies Gaussian noise in square patches, and ANT [24], which augments with an adversarially learned noise distribution. AutoAugment [2] learns augmentation policies that optimize clean error but has since been shown to improve corruption error [35]. Mixup [37] can improve robustness [18], but its label augmentation complicates the dependence on image augmentation. Stylized-ImageNet [9], which applies style transfer to input images, can also improve robustness. Noisy Student [34] and Assemble-ResNet [18] combine data augmentation with new models and training procedures and greatly enhance corruption robustness.

3 Perceptual similarity for augmentations and corruptions

First, we study the importance of similarity between augmentations and corruptions for improving performance on those corruptions. To do so, we need a means to compare augmentations and corruptions. Both types of transforms are perceptual in nature, meaning they affect low-level image structure while leaving high-level semantic information intact, so we expect a good distance to be a measure of *perceptual similarity*. Then, we need to find the appropriate measure of distance between the augmentation and corruption *distributions*. We will argue below that distributional equivalence is not appropriate in the context of corruption robustness, and instead introduce the *minimal sample distance*, a simple measure that does capture a relevant sense of distribution distance.

Measuring similarity between perceptual transforms. We define a perceptual transform as a transform that acts on low-level image structure but not high-level semantic information. As such, we expect two transforms should be similar if their actions on this low-level structure are similar, independent of algorithmic or per-pixel differences between them. A closely related, well-studied problem is the perceptual similarity between *images*. A common approach is to train a neural network on a classification task and use intermediate layers as a feature space for measuring distances [39]. We adapt this idea to obtain a feature space for measuring distances between perceptual transforms.

We start with a feature extractor for images, which we call $\hat{f}(x)$. To train the model from which we will extract features, we assume access to a dataset $\mathbb{D}$ of image label pairs (x, y) associated with a classification task. The model should be trained using only default data augmentation for the task in question so that the feature extractor is independent of the transforms we will use it to study. In order to obtain a very simple measure, we use just the last hidden layer of the network as a feature space.

A perceptual transform $t(x)$ may be encoded by applying it to all images in $\mathbb{D}$, encoding the transformed images, and averaging the features over these images. For efficiency, we find it sufficient to average over only a randomly sampled subset of images $\mathbb{D}_S$ in $\mathbb{D}$. We show in the supplementary

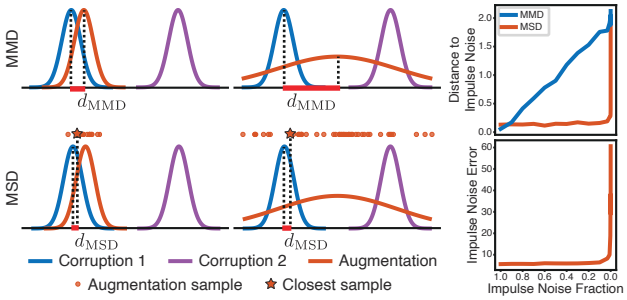

Figure 2: (a) Schematic comparison of MMD to MSD. MMD measures the distance between distribution centers and is only small if the augmentation overlaps with a corruption. MSD measures to the nearest sampled point in the set of samples (marked by a star) and is small even for broad distributions that overlap with multiple corruptions. (b) We test on images corrupted with *impulse noise*, and train on images augmented with a mixture of *impulse noise* and *motion blur*. As the mixing fraction of *impulse noise* decreases, MMD between the augmentation and corruption grows linearly while MSD and error stay low until nearly 0% mixing fraction.

material this produces stable estimates for reasonable numbers of images. The random choice of images is a property of the feature extractor, and so remains fixed when encoding multiple transforms. This reduces variance when computing distances between two transforms. The transform feature extractor is given by $f(t) = \mathbb{E}_{x \in \mathbb{D}_S}[\hat{f}(t(x)) - \hat{f}(x)]$. The *perceptual similarity* between an augmentation and a corruption can be taken as the L_2 distance on this feature space f.

Minimal sample distance. We now seek to compare the distribution of an augmentation scheme p_a to a distribution of a corruption benchmark p_c. If the goal was to optimize error on a *known* corruption distribution, exact equivalence of distributions is the correct measure to minimize. But since the goal is robustness to general, *unknown* corruption distributions, a good augmentation scheme should be equivalent to no single corruption distribution.

To illustrate this behavior, consider a toy problem where we have access to the corruption transforms at training time. A very rough, necessary-but-insufficient measure of distributional similarity is $d_{\mathrm{MMD}}(p_a, p_c) = ||\mathbb{E}_{a \sim p_a}[f(a)] - \mathbb{E}_{c \sim p_c}[f(c)]||$. This is the maximal mean discrepancy on a fixed, finite feature space, so for brevity we will refer to it as MMD. We still employ the featurization $f(t)$, since we are comparing transforms and not images, unlike in typical domain adaptation. Consider two corruption distributions, here *impulse noise* and *motion blur*, and an augmentation scheme that is a mixture of the two corruption distributions. Figure 2b shows MMD between the augmentation and *impulse noise* corruption scales linearly with mixing fraction, but error on *impulse noise* remains low until the mixing fraction is almost 0% impulse noise. This implies distributional similarity is a poor predictor corruption error. Indeed, low d_{MMD} with any one corruption distribution suggests the augmentation overlaps significantly it, so the augmentation is unlikely to aid dissimilar corruptions.

Our expectation for the behavior of the error in Figure 2b is that networks can often successfully memorize rare examples seen during training, so that only a very small fraction of sampled images need *impulse noise* augmentations to perform well on *impulse noise* corruptions. An appropriate distance should then measure how close augmentation samples can come to the corruption distribution, even if the density of those samples is low. We thus propose a very simple measure called *minimal sample distance (MSD)*, which is just the perceptual similarity between an average corruption and the closest augmentation from a finite set of samples $\mathbb{A} \sim p_a$:

$$d_{\mathrm{MSD}}(p_a, p_c) = \min_{a \in \mathbb{A} \sim p_a} ||f(a) - \mathbb{E}_{c \sim p_c}[f(c)]|| \,. \tag{1}$$

A schematic comparison of MMD and MSD is shown in Figure 2a. While both MMD and MSD are small for an augmentation scheme that is distributionally similar to a corruption distribution, only MSD remains small for a broad distribution that occasionally produces samples near multiple corruption distributions. Figure 2b shows MSD, like test error, is small for most mixing fractions in the toy problem described above. Note the measure's need to accommodate robustness to general, unknown corruption distributions has led it to be asymmetric, so it differs from more formal distance metrics that may be used to predict generalization error, such as the Wasserstein distance [40].

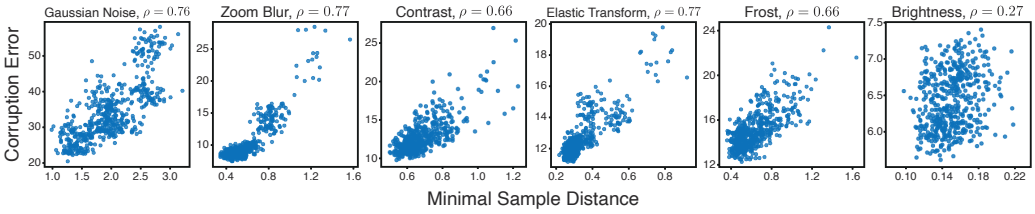

Figure 3: Example relationships between MSD and corruption error. ρ is the Spearman rank correlation. MSD correlates well with error across all four categories of corruption in CIFAR-10-C. For completeness, we also show *brightness*, a negative example where correlation is poor.

4 Perceptual similarity is predictive of corruption error

We are now equipped to measure how important this augmentation-corruption similarity is for corruption error. For a large number of augmentation schemes, we will measure both the MSD to a corruption distribution and the corruption error of a model trained with that scheme. We will find a correlation between MSD and corruption error, which provides evidence that networks generalize across perceptually similar transforms. Then, we will calculate MSD for augmentation schemes in the literature that have been shown to improve error on corruption benchmarks. We will find a correlation between MSD and error here as well, suggesting their success is in part explained by their perceptual similarity to the benchmark. This implies there may be a risk of poor generalization to different benchmarks, since we would not expect this improvement to transfer to a dissimilar corruption.

4.1 Experimental setup

Corruptions. We use CIFAR-10-C [13], which is a common benchmark used for studying corruption robustness. It consists of 15 corruptions, each further split into five different severities of transformation, applied to the CIFAR-10 test set. The 15 corruptions fall into four categories: per-pixel noise, blurring, synthetic weather effects, and digital transforms. We treat each corruption at each severity as a separate distribution for the sake of calculating MSD and error; however, for simplicity we average errors and distances over severity to present a single result per corruption. Examples of each corruption are shown in the supplementary material.

Space of augmentation schemes. To build each sampled augmentation transform, we will composite a set of base augmentations. For base augmentations, we consider the nine common image transforms used in Hendrycks et al. [14], shown in the supplementary material. There are five geometric transforms and four color transforms. By taking all subsets of these base augmentations, we obtain $2^9 = 512$ unique augmentation schemes, collectively called the *augmentation powerset*. Also following Hendrycks et al. [14], we composite transforms in two ways: by applying one after another, or by applying them to copies of the image and then linearly superimposing the results.

Computing similarity and corruption error. A WideResNet-40-2 [36] model is pre-trained on CIFAR-10 using default augmentation and training parameters from Hendrycks et al. [14]. WideResNet is a common baseline model used when studying data augmentation on CIFAR-10 [2, 14, 37]. Its last hidden layer is used as the feature space. For MSD, we average over 100 images, 100 corruptions, and minimize over 100k augmentations. We argue in the supplementary material that these are reasonable choices and show that using VGG [28] instead of WideResNet for the feature extractor gives similar results. Images are from the training set and do not have default training augmentation. A WideResNet-40-2 with the same training parameters is used for corruption error evaluation.

4.2 Analysis

MSD correlates with corruption error. First, we establish the correlation between MSD and corruption error on the augmentation powerset. MSD shows strong correlation with corruption error across corruptions types in all four categories of CIFAR-10-C, and for a large majority of CIFAR-10-C corruptions in general: 12 of 15 have Spearman rank correlation greater than 0.6. Figure 3 shows the relationship between distance and corruption error on six example corruptions, including one negative example for which correlation is low. This corruption, *brightness*, may give poor results because it

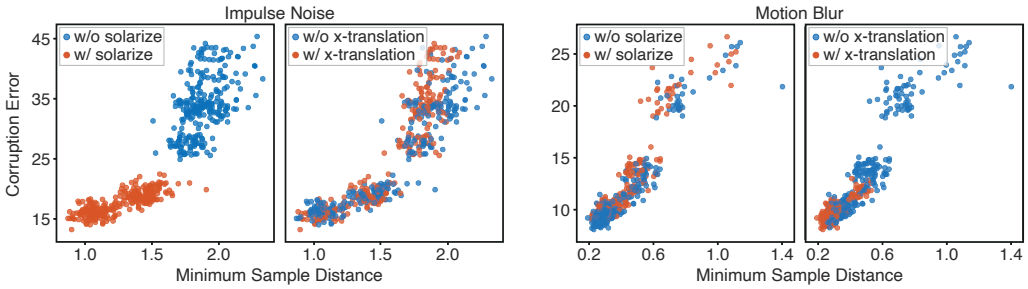

Figure 4: Example relationships between base augmentations and corruptions. Including *solarize* reduces MSD on the perceptually similar *impulse noise* corruption. Including *x translation* reduces MSD on the perceptually similar *motion blur* corruption. MSD is not decreased for dissimilar augmentation-corruption pairs.

is a single low-level image statistic that can vary significantly from image to image, and thus may not be well represented by our feature extractor. In the supplementary material, we confirm MMD correlates poorly with corruption error, as expected. In particular, we expect broad augmentation schemes produce samples similar to a larger set of corruptions, leading to both lower MSD and lower corruption error but higher MMD. Additionally, the correlation between MSD and corruption error suggests that perceptual similarity is a predictor of corruption error. A complete set of correlation plots is shown in the supplementary material. Also in the supplementary material, we consider compatibility with Schneider et al. [25], which shows adapting a model's batch norm statistics to a corruption can help robustness. We repeat our experiment but do not train on the augmentations, instead only adapting the batch norm statistics of a pre-trained model to them. We still find a strong correlation, suggesting our methods are indeed compatible with Schneider et al. [25].

An example of perceptual similarity. Here we illustrate the perceptual nature of the similarity measure, using an example with two base augmentations and two corruptions. The augmentation *solarize* and the corruption *impulse noise* both insert bright pixels into the image, though in different ways. Linear superpositions of the augmentation *x translation* are visually similar to a blur, such as the corruption *motion blur*. Examples of these transforms are shown in the supplementary material. Figure 4 shows MSD vs error where augmentation schemes that include *solarize* and *x translation* are colored. It is clear that including an augmentation greatly decreases MSD to its perceptually similar corruption, while having little effect on MSD to its perceptually dissimilar corruption.

MSD and corruption error in real augmentation methods. The augmentation powerset may be used as a baseline for comparing real data augmentation schemes. Figure 5 shows example MSD-error correlations for Patch Gaussian [19], AutoAugment [2], and Augmix [14], along with the cloud of augmentation powerset points. The real augmentation schemes follow the same general trend that lower error predicts lower MSD. A few intuitive correlations are also captured in Figure 5. Patch Gaussian has low MSD to noise corruptions. AutoAugment, which contains contrast and Gaussian blurring augmentations in its sub-policies, has low MSD with *contrast* and *defocus blur*. We also show *fog*, an example where for AutoAugment MSD is not predictive of corruption error.

This correlation suggests generalization may be poor beyond an existing benchmark, since an augmentation scheme may be perceptually similar to one benchmark but not another. For augmentations and corruptions that are explicitly the same, such as *contrast* in AutoAugment and ImageNet-C, this is typically accounted for by removing such transforms from the augmentation scheme when testing corruption robustness[3]. But in addition to these explicit similarities, Figure 5 shows quantitatively that perceptual similarity between non-identical augmentations and corruptions is also strongly predictive of corruption error. This includes possibly unexpected similarities, such as between Patch Gaussian and *glass blur*, which introduces random pixel-level permutations as noise. This suggests that perceptually similar augmentations and corruptions should be treated with the same care as identical transforms. In particular, tools such as MSD help us determine *why* an augmentation scheme

[3]For this analysis, we wish to treat explicit transform similarity and perceptual transform similarity on the same footing, so we choose not to remove these overlapping augmentations.

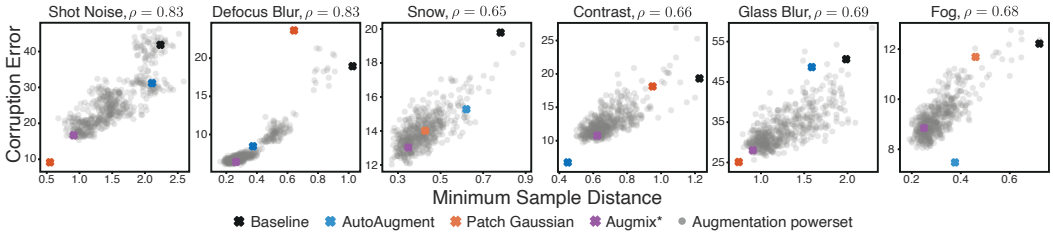

Figure 5: Example correlations for augmentation schemes from the literature. Patch Gaussian is similar to noise, while AutoAugment is similar to contrast and blur, as expected from their formulation. Glass blur acts more like a noise corruption than a blur for these augmentation schemes, likely because it randomly permutes pixels. As a negative example, MSD does not correlate well with error for AutoAugment on *fog*. *AugMix here refers to just the augmentation distribution in Hendrycks et al. [14], not the proposed Jensen-Shannon divergence loss.

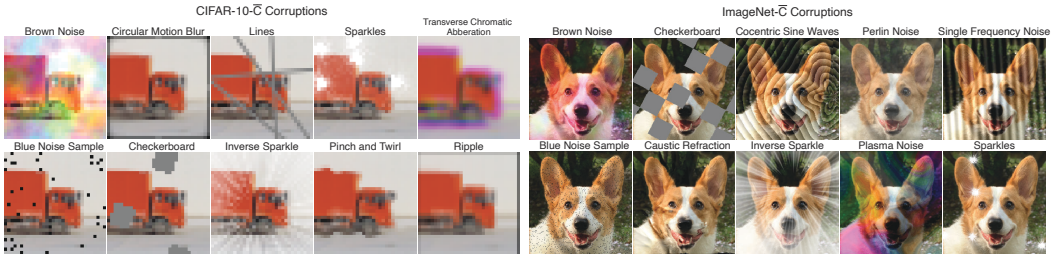

Figure 6: Example CIFAR-10-$\overline{\text{C}}$ and ImageNet-$\overline{\text{C}}$ corruptions. While still human interpretable, new corruptions are sampled to be dissimilar from CIFAR-10/ImageNet-C. Base images © Sehee Park and Chenxu Han.

improves corruption error, so we can better understand if new methods will generalize beyond their tested benchmarks. Next we test this generalization by finding corruptions dissimilar to ImageNet-C.

5 ImageNet-$\overline{\text{C}}$: benchmarking with dissimilar corruptions

We now introduce a set of corruptions, called ImageNet-$\overline{\text{C}}$, that are perceptually dissimilar to ImageNet-C in our transform feature space, and we will show that several augmentation schemes have degraded performance on the new dataset. We emphasize that the dataset selection method uses only default data augmentation and was fixed before we looked at the results for different augmentations, so we are not adversarially selecting against the tested augmentation schemes.

Dataset construction. An overview of the dataset construction is presented here, with exact details in the supplementary material. We build 30 new corruptions in 10 severities, from which the 10 most dissimilar corruptions will be chosen. We adapt common filters and noise distributions available online [10, 16] to produce human interpretable images. The transforms include warps, blurs, color distortions, noise additions, and obscuring effects. Examples are shown in the supplementary material.

To assure that the new dataset is no harder than ImageNet-C, we restrict the average corruption error of the new dataset to be similar to that of ImageNet-C for default augmentation. We then generate many potential datasets and measure the average shift in distance to ImageNet-C that each corruption contributes, shown in the supplementary material. Note that while MSD is a measure between augmentations and corruptions, here we are comparing corruptions to other corruptions and thus use MMD in our transform feature space. ImageNet-$\overline{\text{C}}$ then consists of the 10 corruptions types with the largest average shift in distance. Like ImageNet-C, each has five different severities, with severities chosen so that the average error matches ImageNet-C for default augmentation. Example transforms from ImageNet-$\overline{\text{C}}$ and CIFAR-10-$\overline{\text{C}}$ are shown in Figure 6. This procedure in our feature space produces corruptions intuitively dissimilar from ImageNet-C and CIFAR-10-C.

Results. We test AutoAugment [2], Patch Gaussian [19], AugMix [14], ANT^{3x3} [24], and Stylized-ImageNet [9] on our new datasets and show results in Table 1. CIFAR-10 models are WideResNet-40-2 with training parameters from Hendrycks et al. [14]. ImageNet [4] models are ResNet-50 [12]

with training parameters from Goyal et al. [11]. Stylized-ImageNet is trained jointly with ImageNet for half the epochs and starts from a model pre-trained on ImageNet, following Geirhos et al. [9]. Models use default data augmentation as well as the augmentation being tested, except ImageNet color jittering is not used. All corruptions are applied in-memory instead of loaded from a compressed file; this can affect results especially on high frequency corruptions.

Since Section 4 suggests several augmentation schemes are perceptually similar to ImageNet-C corruptions, we might expect these methods to have worse error on the new corruptions. Indeed, every augmentation scheme performs worse. Different augmentation schemes also degrade by significantly different amounts, from $+0.7\%$ for AutoAugment to $+7.3\%$ for PatchGaussian, which changes their ranking by corruption error and leads to inconsistency of generalization. In the supplementary material, we compare performance on several robust models[7, 20, 21, 30, 32, 34, 38] that are not primarily augmentation-based and see no similar pattern of degradation, further suggesting that augmentation-corruption dissimilarity is the cause of the higher error.

Errors of individual corruptions in ImageNet-$\overline{\text{C}}$ are also revealing. For all augmentation schemes, there is significant improvement on *blue sample noise*[4] but little improvement on *sparkles* or *inverse sparkles*. Only AutoAugment does well on *checkerboard*, perhaps because only AutoAugment's geometric transforms produce empty space, similar to *checkerboard*'s occluded regions. These examples suggest a slightly different benchmark could yield significantly different results. Indeed, for a hypothetical benchmark that excluded *blue sample noise* and *checkerboard*, AutoAugment and Patch Gaussian have 57.3% and 57.2% error respectively, little better than baseline of 57.4%. The best fairs only slightly better: AugMix has 54.3% error. Generalization to dissimilar corruptions is thus both inconsistent and typically quite poor. Single benchmarks and aggregate corruption scores are likely not enough for careful evaluation of robustness to unknown corruptions, and it is important to study why proposed augmentations succeed to better understand how well they might generalize.

It may be surprising that Stylized-ImageNet also degrades, given that it is intuitively very different from every corruption. While our measure works for augmentations, it does not cover all possible methods that improve robustness, such as more complicated algorithms like Stylized-ImageNet. Stylized-ImageNet degradation may be due to other reasons. For instance, it primarily augments texture information and may help mostly with higher frequency corruptions, as can be seen by its improvement on *single frequency noise* and *cocentric sine waves*; ImageNet-$\overline{\text{C}}$ has fewer such corruptions than ImageNet-C. ImageNet-$\overline{\text{C}}$ is thus a useful tool for understanding the interaction between training procedure and corruption distribution, even beyond perceptual similarity.

Nevertheless, note that it is the intuitively broader augmentation schemes, such as AutoAugment, AugMix, and Stylized-ImageNet, that generalize better to ImageNet-$\overline{\text{C}}$. The importance of breadth has also been explored elsewhere[15, 35], but in the previous sections we have provided new quantitative evidence for *why* this may be true: broad augmentation schemes may be perceptually similar to more types of corruptions, and thus more likely to be perceptually similar to a new corruption. Moreover, AugMix still improves over baseline on ImageNet-$\overline{\text{C}}$, so there is reason to be optimistic that robustness to unknown corruptions is an achievable goal, as long as evaluation is treated carefully.

6 Discussion

Societal Impact. Our method for finding dissimilar corruptions could in principle be used to adversarially attack computer vision systems, such as those in content moderation or self-driving cars. Moreover, our ultimate goal is to help improve robustness in computer vision, and such robust systems may be used in detrimentals ways, for example in autonomous weapons or surveillance. However, we expect better evaluation of robust models to have definite benefits as well. In the long run, such an understanding should help defend against adversarial attacks. Our tools could also be used to challenge purportedly robust systems that are actually dangerously unreliable, such as an autonomous driving system that is robust to common corruption benchmarks yet fails to be robust to a dissimilar but important corruption, e.g., maybe glare.

Corruption robustness as a secondary learning task. We have provided evidence that data augmentation may not generalize well beyond a given corruption benchmark. To explore this further,

[4]This corruption is conceptually similar with *impulse noise* but also gives a large distance; this may be a failure mode of our measure, maybe since *impulse noise* has bright pixels and *blue noise sample* has dark pixels.

Table 1: Test error for several data augmentation methods on CIFAR-10-$\overline{\text{C}}$ and ImageNet-10-$\overline{\text{C}}$, for which every method performs worse than on ImageNet-C or CIFAR-10-C. The increase in error differs significantly between different augmentation methods. Descriptions of the abbreviations and standard deviations for individual corruptions are in the supplementary material. 'Baseline' refers to default augmentation only. Averages are over five runs for ImageNet and ten for CIFAR-10. *ANT uses the pre-trained model provided with the paper and has slightly different training parameters.

	IN-C	**IN-$\overline{\text{C}}$**		**ImageNet-$\overline{\text{C}}$ Corruptions**									
Aug	Err	Err	ΔIN-C	BSmpl	Plsm	Ckbd	CSin	SFrq	Brwn	Prln	Sprk	ISprk	Rfrac
Baseline	58.1 $\pm$0.4	57.7 $\pm$0.2	-0.4	68.6	71.7	49.4	84.7	79.0	37.5	34.3	32.4	76.7	42.8
AA	55.0 $\pm$0.2	55.7 $\pm$0.3	+0.7	54.8	68.3	43.8	86.5	78.8	34.5	33.8	36.1	77.1	43.8
SIN	52.4 $\pm$0.1	55.8 $\pm$0.3	+3.4	54.7	69.8	52.8	79.6	69.2	37.8	35.3	37.0	77.3	44.1
AugMix	49.2 $\pm$0.7	52.4 $\pm$0.2	+3.2	43.2	72.2	46.1	76.3	67.4	38.8	32.4	32.3	76.4	39.2
PG	49.3 $\pm$0.2	56.6 $\pm$0.4	+7.3	60.3	74.1	48.5	82.1	76.7	38.9	34.6	32.1	76.5	42.1
ANT*	48.8	53.9	+5.1	35.8	75.5	56.9	76.4	63.7	41.0	35.2	35.0	76.1	43.3
	C10-C	**C10-$\overline{\text{C}}$**		**CIFAR-10-$\overline{\text{C}}$ Corruptions**									
Aug	Err	Err	ΔC10-C	BSmpl	Brwn	Ckbd	CBlur	ISprk	Line	P&T	Rppl	Sprk	TCA
Baseline	27.0 $\pm$0.6	27.1 $\pm$0.5	+0.1	42.9	27.2	23.3	11.8	43.3	26.2	11.3	21.6	21.0	42.9
AA	19.4 $\pm$0.2	21.0 $\pm$0.4	+1.6	17.7	17.5	17.6	9.5	40.4	23.6	10.7	23.5	17.5	31.8
AugMix	11.1 $\pm$0.2	16.0 $\pm$0.3	+5.9	9.8	27.8	13.4	5.9	30.3	18.0	8.3	12.1	15.5	19.2
PG	17.0 $\pm$0.3	23.8 $\pm$0.5	+6.8	9.0	30.1	21.6	12.8	35.4	20.6	8.8	21.5	19.3	59.5

consider an analogy to a regular learning problem. We may think of corruption robustness in the presence of data augmentation as a sort of secondary task layered on the primary classification task: the set of data augmentations is the training set, the set of corruptions is the test set, and the goal is to achieve invariance of the underlying primary task. In this language, the 'datasets' involved are quite small: ImageNet-C has only 15 corruption types, and several augmentation schemes composite only around 10 basic transforms. In this case, standard machine learning practice would dictate a training/validation/test set split; it is only the size and breadth of modern vision datasets that has allowed this to be neglected in certain cases recently. But the effective dataset size of a corruption robustness problem is tiny, so having a held-out test set seems necessary. To emphasize, this is not a test set of the underlying classification task, for which generalization has been studied by Recht et al. [22, 23]. Instead, it is a test set of corruption transforms themselves. This means there would be validation/test split of dissimilar transformations, both applied to the ImageNet validation set[5].

Real-world corruption robustness. Recently, Hendrycks et al. [15] and Taori et al. [31] study how performance on corruption transforms generalizes to real-world corruptions and come to conflicting conclusions. Though we do not study real-world corruptions, we have proposed a mechanism that may explain the conflict: performance will generalize between transforms and real-world corruptions if they are perceptually similar, but will likely not if they are dissimilar. Since Hendrycks et al. [15] and Taori et al. [31] draw on different real-world and synthetic corruptions, it may be that the perceptual similarity between datasets differs in the two analyses. This also suggests a way to find additional corruption transforms that correlate with real-world corruptions: transforms should be sought that have maximal perceptual similarity with real-world corruptions.

Generalization does occur. We have encountered two features of data augmentation that may explain why it can be such a powerful tool for corruption robustness, despite the issues discussed above. First, within a class of perceptually similar transforms, generalization does occur. This means each simple data augmentation may confer robustness to many complicated corruptions, as long as they share perceptual similarity. Second, dissimilar augmentations in an augmentation scheme often causes little to no loss in performance, as long as a similar augmentation is also present. We briefly study this in the supplementary material by demonstrating that adding many dissimilar augmentations increases error much less than adding a few similar augmentations decreases it. These two features suggest broad augmentation schemes with many dissimilar augmentations may confer robustness to a large class of unknown corruptions. More generally, we think data augmentation is a promising direction of study for corruption robustness, as long as significant care is taken in evaluation.

[5]The validation set provided in Hendrycks & Dietterich [13] consists of perceptually similar transforms to ImageNet-C and would not be expected to work well for the validation discussed here.

References

[1] Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[2] Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V. AutoAugment: Learning augmentation strategies from data. In *CVPR*, 2019.

[3] Dao, T., Gu, A., Ratner, A. J., Smith, V., De Sa, C., and Ré, C. A kernel theory of modern data augmentation. *Proceedings of machine learning research*, 97:1528, 2019.

[4] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[6] Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. In *ICCCN*, 2017.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[8] Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018.

[9] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[10] Gladman, S. J. Filterpedia, 2016. URL `https://github.com/FlexMonkey/Filterpedia`.

[11] Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[12] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018.

[14] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. AugMix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2019.

[15] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[16] Huxtable, J. JH Labs Java Image Processing, 2006. URL `http://www.jhlabs.com/ip/filters/`.

[17] Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

[18] Lee, J., Won, T., and Hong, K. Compounding the performance improvements of assembled techniques in a convolutional neural network. *arXiv preprint arXiv:2001.06268*, 2020.

[19] Lopes, R. G., Yin, D., Poole, B., Gilmer, J., and Cubuk, E. D. Improving robustness without sacrificing accuracy with Patch Gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

[20] Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 181–196, 2018.

[21] Orhan, A. E. Robustness properties of Facebook's ResNeXt WSL models. *arXiv preprint arXiv:1907.07640*, 2019.

[22] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018.

[23] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

[24] Rusak, E., Schott, L., Zimmermann, R., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. A simple way to make neural networks robust against diverse image corruptions. *arXiv preprint arXiv:2001.06057*, 2020.

[25] Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. In *NeurIPS*, 2020.

[26] Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019.

[27] Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. Transformation invariance in pattern recognition—tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pp. 239–274. Springer, 1998.

[28] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[29] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014.

[30] Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[31] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

[32] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[33] Wu, S., Zhang, H. R., Valiant, G., and Ré, C. On the generalization effects of linear transformations in data augmentation. In *ICML*, 2020.

[34] Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with Noisy Student improves imagenet classification. In *CVPR*, 2020.

[35] Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. A Fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019.

[36] Zagoruyko, S. and Komodakis, N. Wide residual networks. In *BMVC*, 2016.

[37] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[38] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Muller, J., Manmatha, R., Li, M., and Smola, A. ResNeSt: Split-Attention Networks. *arXiv preprint arXiv:2004.08955*, 2020.

[39] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[40] Zilly, J., Zilly, H., Richter, O., Wattenhofer, R., Censi, A., and Frazzoli, E. The Frechet Distance of training and test distribution predicts the generalization gap. *OpenReview preprint*, 2019. URL `https://openreview.net/forum?id=SJgSflHKDr`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Potential failure cases that arise in the experiments are discussed in sections 4 and 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the beginning of section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We use images licensed for free use as example images throughout the paper and provide access to no data that may be sensitive, neither in the paper nor in supplementary material. The societal impact of the methodology, tools, and experiments we do provide is discussed as per question 1c above.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code is included in the supplemental material and will be made available via URL once the paper is de-anonymized.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The experimental set-ups of sections 4 and 5 are described in detail within each section. For section 4 this specifies all training details, while for section 5 an exact protocol is described in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Experiments in section 5 are average over 5 runs for ImageNet and 10 runs for CIFAR-10, and we report error bars in each case. In Section 4 experiments are scatter plots over a large number of points and so do not have error bars per point.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The supplementary material discusses the types of resources used and reports estimates of the runtime for the training and evaluation of models used in the paper's experiments. However, total compute resources over the course of the entire research project was not recorded.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Citations to CIFAR-10, ImageNet, and ImageNet-C are all included.
   (b) Did you mention the license of the assets? [No] CIFAR-10 and ImageNet are standards in the field and their terms of use are obtainable from the cited references.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code is provided in the supplementary material, and will be made available via URL once the paper is de-anonymized.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] CIFAR-10 and ImageNet have known issues in this regard but are standards of the field and are necessary for research in the area. ImageNet-C is a set of image transforms applied to ImageNet and introduces no additional raw data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] CIFAR-10 and ImageNet have known issues in this regard but are standards of the field and are necessary for research in the area. ImageNet-C is a set of image transforms applied to ImageNet and introduces no additional raw data.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]