PRIME: A Few Primitives Can Boost Robustness to Common Corruptions

Apostolos Modas*[1], Rahul Rade*[2], Guillermo Ortiz-Jiménez[1], Seyed-Mohsen Moosavi-Dezfooli[3], and Pascal Frossard[1]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] ETH Zürich, Switzerland
[3] Imperial College London, United Kingdom

Abstract. Despite their impressive performance on image classification tasks, deep networks have a hard time generalizing to unforeseen corruptions of their data. To fix this vulnerability, prior works have built complex data augmentation strategies, combining multiple methods to enrich the training data. However, introducing intricate design choices or heuristics makes it hard to understand which elements of these methods are indeed crucial for improving robustness. In this work, we take a step back and follow a principled approach to achieve robustness to common corruptions. We propose PRIME, a general data augmentation scheme that relies on simple yet rich families of max-entropy image transformations. PRIME outperforms the prior art in terms of corruption robustness, while its simplicity and plug-and-play nature enable combination with other methods to further boost their robustness. We analyze PRIME to shed light on the importance of the mixing strategy on synthesizing corrupted images, and to reveal the robustness-accuracy trade-offs arising in the context of common corruptions. Finally, we show that the computational efficiency of our method allows it to be easily used in both on-line and off-line data augmentation schemes[1].

1 Introduction

Deep image classifiers do not work well in the presence of various types of distribution shifts [14,18,39]. Most notably, their performance can severely drop when the input images are affected by common corruptions that are not contained in the training data, such as digital artefacts, low contrast, or blurs [21,29]. In general, "common corruptions" is an umbrella term coined to describe the set of all possible distortions that can happen to natural images during their acquisition, storage, and processing lifetime, which can be very diverse. Nevertheless, while the space of possible perturbations is huge, the term "common corruptions" is generally used to refer to image transformations that, while degrading the quality of the images, still preserve their semantic information.

* The first two authors contributed equally to this work.
[1] Our code is available at https://github.com/amodas/PRIME-augmentations

Fig. 1. Images generated with PRIME, a simple method that uses a family of max-entropy transformations in different visual domains to create diverse augmentations.

Building classifiers that are robust to common corruptions is far from trivial. A naive solution is to include data with all sorts of corruptions during training, but the sheer scale of all possible types of typical perturbations that might affect an image is simply too large. Moreover, the problem is per se ill-defined since there exists no formal description of all possible common corruptions.

To overcome this issue, the research community has recently favoured increasing the "diversity" of the training data via data augmentation schemes [10,22,20]. Intuitively, the hope is that showing very diverse augmentations of an image to a network would increase the chance that the latter becomes invariant to some common corruptions. Still, covering the full space of common corruptions is hard. Hence, current literature has mostly resorted to increasing the diversity of augmentations by designing intricate data augmentation pipelines, e.g., introducing DNNs for generating varied augmentations [20,5], or coalescing multiple techniques [41], and thus achieve good performance on different benchmarks. This strategy, though, leaves a big range of unintuitive design choices, making it hard to pinpoint which elements of these methods meaningfully contribute to the overall robustness. Meanwhile, the high complexity of recent methods [41,5] makes them impractical for large-scale tasks. Whereas, some methods are tailored to particular datasets and might not be general enough. Nonetheless, the problem of building robust classifiers is far from completely solved, and the gap between robust and standard accuracy is still large.

In this work, we take a step back and provide a systematic way for designing a simple, yet effective data augmentation scheme. By focusing on first principles, we formulate a new mathematical model for semantically-preserving corruptions, and build on basic concepts to characterize the notions of transformation strength and diversity using a few transformation primitives. Relying on this model, we propose *PRIME*, a data augmentation scheme that draws transformations from a max-entropy distribution to efficiently sample from a large space of possible distortions (see Fig. 1). The performance of PRIME, alone, already tops the current baselines on different common corruption datasets, whilst it can also be combined with other methods to further boost their performance. Moreover, the simplicity and flexibility of PRIME allows to easily understand how each of its components contributes to improving robustness.

Altogether, the main contributions of our work include:

- We introduce PRIME, a simple method that is built on a few guiding principles, which efficiently boosts robustness to common corruptions.
- We experimentally show that PRIME, despite its simplicity, achieves state-of-the-art robustness on multiple corruption benchmarks.
- Last, our thorough ablation study sheds light on the necessity of having diverse transformations, on the role of mixing in the success of current methods, on the potential robustness-accuracy trade-off, and on the importance of online augmentations.

Overall, PRIME is a simple model-based scheme that can be easily understood, ablated, and tuned. Our work is an important step in the race for robustness against common corruptions, and we believe that it has the potential to become the new baseline for learning robust classifiers.

2 General model of visual corruptions

In this work, motivated by the "semantically-preserving" nature of common corruptions, we leverage the long tradition of image processing in developing techniques to manipulate images while retaining their semantics, and construct a principled framework to characterize a large space of visual corruptions.

Let $\boldsymbol{x} : [0,1]^2 \to [0,1]^3$ be a continuous image[2] mapping pixel coordinates $\boldsymbol{r} = (r_1, r_2)$ to RGB values. We define our model of common corruptions as the action on $\boldsymbol{x}$ of the following additive subgroup of the near-ring of transformations [4]

$$\mathcal{T}_{\boldsymbol{x}} = \left\{ \sum_{i=1}^{n} \lambda_i \, g_1^i \circ \cdots \circ g_m^i(\boldsymbol{x}) : \; g_j^i \in \{\omega, \tau, \gamma\}, \lambda_i \in \mathbb{R} \right\}, \tag{1}$$

where ω, τ and γ are random primitive transformations which distort $\boldsymbol{x}$ along the spectral (ω), spatial (τ), and color (γ) domains. As we will see, defining each of these primitives in a principled and coherent fashion will be enough to construct a set of perturbations which covers most types of visual corruptions.

To guarantee as much diversity as possible in our model, we follow the principle of maximum entropy to define our distributions of transformations [8]. Note that using a set of augmentations that guarantees maximum entropy comes naturally when trying to optimize the sample complexity derived from certain information-theoretic generalization bounds, both in the clean [42] and corrupted settings [28]. Specifically, the principle of maximum entropy postulates favoring those distributions that are as unbiased as possible given the set of constraints that define a family of distributions. In our case, these constraints are given in the form of an expected strength σ^2, some boundary conditions, e.g., the displacement field must be zero at the borders of an image, and finally the desired smoothness level K. The principle of smoothness helps formalize the notion of physical plausibility, as most naturally occurring processes are smooth.

[2] In practice, we will work with discrete images on a regular grid.

Let $\mathcal{I}$ denote the space of all images, and let $f : \mathcal{I} \to \mathcal{I}$ be a random image transformation distributed according to the law μ. Further, let $\mathcal{C} \subseteq \mathcal{F}$ be a set of constraints that restricts the domain of applicability of f, i.e., $f \in \mathcal{C}$, with $\mathcal{F}$ denoting the space of functions $\mathcal{I} \to \mathcal{I}$. The maximum entropy principle postulates using the distribution μ which has maximum entropy given the constraints:

$$\begin{aligned} \underset{\mu}{\text{maximize}} \quad & H(\mu) = -\int_{\mathcal{F}} \mathrm{d}\mu(f) \log(\mu(f)) \\ \text{subject to} \quad & f \in \mathcal{C} \quad \forall f \in \operatorname{supp}(\mu), \end{aligned} \tag{2}$$

where $H(\mu)$ represents the entropy of the distribution μ [8]. In its general form, solving Eq. (2) for any set of constraints $\mathcal{C}$ is intractable. In Appendix A, we formally derive the analytical expressions for the distributions of each of our family of transformations, by leveraging results from statistical physics [1].

In what follows, we describe the analytical solutions to Eq. (2) for each of our basic primitives. In general, these distributions are governed by two parameters: K to control smoothness, and σ^2 to control strength. These transformations fall back to identity mappings when $\sigma^2 = 0$, independently of K.

Spectral domain We parameterize the distribution of random spectral transformations using random filters $\boldsymbol{\omega}(\boldsymbol{r})$, such that the transformation output follows

$$\omega(\boldsymbol{x})(\boldsymbol{r}) = \left(\boldsymbol{x} * (\boldsymbol{\delta} + \boldsymbol{\omega}')\right)(\boldsymbol{r}), \tag{3}$$

where, $*$ is the convolution operator, $\boldsymbol{\delta}(\boldsymbol{r})$ represents a Dirac delta, i.e., identity filter, and $\boldsymbol{\omega}'(\boldsymbol{r})$ is implemented in the discrete grid as an FIR filter of size $K_\omega \times K_\omega$ with i.i.d random entries distributed according to $\mathcal{N}(0, \sigma_\omega^2)$. Here, σ_ω^2 governs the transformation strength, while larger K_ω yields filters of higher spectral resolution. The bias $\boldsymbol{\delta}(\boldsymbol{r})$ retains the output close to the original image.

Spatial domain We model our distribution of random spatial transformations, which apply random perturbations over the coordinates of an image, as

$$\tau(\boldsymbol{x})(\boldsymbol{r}) = \boldsymbol{x}(\boldsymbol{r} + \boldsymbol{\tau}'(\boldsymbol{r})). \tag{4}$$

This model has been recently proposed in [32] to define a distribution of random smooth diffeomorphisms in order to study the stability of neural networks to small spatial transformations. To guarantee smoothness but preserve maximum entropy, the authors propose to parameterize the vector field $\boldsymbol{\tau}'$ as

$$\boldsymbol{\tau}'(\boldsymbol{r}) = \sum_{i^2+j^2 \leq K_\tau^2} \beta_{i,j} \sin(\pi i \boldsymbol{r}_1) \sin(\pi j \boldsymbol{r}_2), \tag{5}$$

where $\beta_{i,j} \sim \mathcal{N}(0, \sigma_\tau^2/(i^2+j^2))$. Such choice guarantees that the resulting mapping is smooth according to the cut frequency K_τ, while σ_τ^2 determines its strength.

Color domain Following a similar approach, we define the distribution of random color transformations as random mappings γ between color spaces

$$\gamma(\boldsymbol{x})(\boldsymbol{r}) = \boldsymbol{x}(\boldsymbol{r}) + \sum_{n=0}^{K_\gamma} \boldsymbol{\beta}_n \odot \sin\left(\pi n\, \boldsymbol{x}(\boldsymbol{r})\right), \tag{6}$$

Algorithm 1: PRIME

Input: Image $\boldsymbol{x}$, primitives $\mathcal{G} = \{\mathrm{Id}, \omega, \tau\, \gamma\}$, where Id is the identity operator
Output: Augmented image $\tilde{\boldsymbol{x}}$

$\tilde{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}$
for $i \in \{1, \ldots, n\}$ **do**
  $\tilde{\boldsymbol{x}}_i \leftarrow \boldsymbol{x}$
  for $j \in \{1, \ldots, m\}$ **do**
    $g \sim \mathcal{U}(\mathcal{G})$ ▷ Strength $\sigma \sim \mathcal{U}(\sigma_{\min}, \sigma_{\max})$
    $\tilde{\boldsymbol{x}}_i \leftarrow g(\tilde{\boldsymbol{x}}_i)$
  end
end
$\tilde{\boldsymbol{x}} \leftarrow \sum_{i=0}^{n} \lambda_i \tilde{\boldsymbol{x}}_i$ ▷ $\boldsymbol{\lambda} \sim \mathrm{Dir}(\mathbf{1})$: Random Dirichlet convex coefficients

where $\boldsymbol{\beta}_n \sim \mathcal{N}(0, \sigma_\gamma^2 \boldsymbol{I}_3)$, with $\odot$ denoting elementwise multiplication. Again, K_γ controls the smoothness of the transformations and σ_γ^2 their strength. Compared to Eq. (5), the coefficients in Eq. (6) are not weighted by the inverse of the frequency, and have constant variance. In practice, we observe that reducing the variance of the coefficients for higher frequencies creates color mappings that are too smooth and almost imperceptible, so we decided to drop this dependency.

Finally, we note that our model can be easily extended to include other distributions of maximum entropy transformations that suit an objective task. For example, one might add the distribution of maximum entropy additive perturbations given by $\eta(\boldsymbol{x})(\boldsymbol{r}) = \boldsymbol{x}(\boldsymbol{r}) + \boldsymbol{\eta}'(\boldsymbol{r})$, where $\boldsymbol{\eta}'(\boldsymbol{r}) \sim \mathcal{N}(0, \sigma_\eta^2)$. Nonetheless, since most benchmarks of visual corruptions disallow the use of additive perturbations during training [21], we do not include an additive perturbation category.

Overall, as demonstrated by our results in Secs. 4.2 and 5.2, our model is very flexible and can cover a large part of the semantic-preserving distortions. It also allows to easily control the strength and style of the transformations with just a few parameters. Moreover, changing the transformation strength enables to control the trade-off between corruption robustness and standard accuracy, as shown in Sec. 5.3. In what follows, we use this model to design an efficient augmentation scheme to build classifiers robust to common corruptions.

3 PRIME: A simple augmentation scheme

We now introduce PRIME, a simple yet efficient augmentation scheme that uses our **PRI**mitives of **M**aximum **E**ntropy to confer robustness against common corruptions. The pseudo-code of PRIME is given in Algorithm 1, which draws a random sample from Eq. (1) using a convex combination of a composition of basic primitives. Below we describe the main implementation details.

Parameter selection It is important to ensure that the semantic information of an image is preserved after it goes through PRIME. As measuring semantic preservation quantitatively is not simple, we subjectively select each primitive's

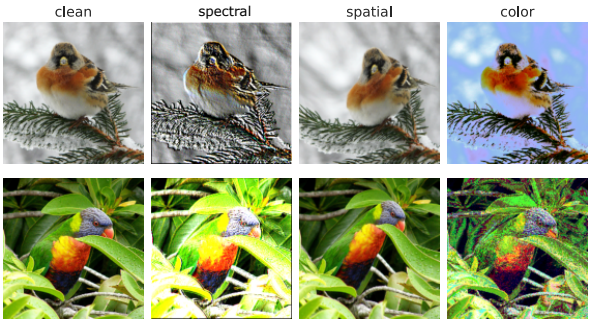

Fig. 2. Images generated with the transformations of our common corruptions model. Despite the perceptibility of the distortion, the image semantics are preserved.

parameters based on visual inspection, ensuring maximum permissible distortion while retaining the semantic content of the image. However, to avoid relying on a specific strength for each transformation, PRIME stochastically generates augmentations of different strengths by sampling σ from a uniform distribution, with different minimum and maximum values for each primitive. Figure 2 shows some visual examples for each kind of transformation, while additional visual examples along with the details of all the parameters can be found in Appendix B.

For the color primitive, we observed that fairly large values for K_γ (in the order of 500) are important for covering a large space of visual distortions. Unfortunately, implementing such a transformation can be memory inefficient. To avoid this issue, PRIME uses a slight modification of Eq. (6) and combines a fixed number Δ of consecutive frequencies randomly chosen in the range $[0, K_\gamma]$.

Mixing transformations The concept of mixing has been a recurring theme in the augmentation literature [45,44,22,41] and PRIME follows the same trend. In particular, Algorithm 1 uses a convex combination of n basic augmentations consisting of the composition of m of our primitive transformations. In general, the convex mixing procedure (i) broadens the set of possible training augmentations, and (ii) ensures that the augmented image stay close to the original one. We later provide empirical results which underline the efficacy of mixing in Sec. 5.2. Overall, the exact mixing parameters are provided in Appendix B. Note that, the basic skeleton of PRIME is similar to that of AugMix. However, as we will see next, incorporating our maximum entropy transformations leads to significant gains in common corruptions robustness over AugMix.

4 Performance analysis

In this section, we compare the classification performance of our method on multiple datasets with that of two current approaches: AugMix and DeepAugment (DA). In Appendix L, we also compare PRIME with additional baselines. We

Table 1. Clean and corruption accuracy, and mean corruption error (mCE) for different methods with ResNet-18 on C-10, C-100, IN-100 and ResNet-50 on IN. mCE is the mean corruption error on common corruptions un-normalized for C-10 and C-100; normalized relative to standard model on IN-100 and IN. † indicates that JSD consistency loss is not used. *Models taken from [9].

Dataset	Method	Clean Acc (↑)	Common Acc (↑)	Corruption mCE (↓)
C-10	Standard	95.0	74.0	24.0
	AugMix	95.2	88.6	11.4
	PRIME	94.2	**89.8**	**10.2**
C-100	Standard	76.7	51.9	48.1
	AugMix	78.2	64.9	35.1
	PRIME	78.4	**68.2**	**31.8**
IN-100	Standard	88.0	49.7	100.0
	AugMix	88.7	60.7	79.1
	DA	86.3	67.7	68.1
	PRIME	85.9	**71.6**	**61.0**
	DA+AugMix	86.5	73.1	57.3
	DA+PRIME	84.9	**74.9**	**54.6**
IN	Standard*	76.1	38.1	76.7
	AugMix*	77.5	48.3	65.3
	DA*	76.7	52.6	60.4
	PRIME†	77.0	**55.0**	**57.5**
	DA+AugMix	75.8	58.1	53.6
	DA+PRIME†	75.5	**59.9**	**51.3**

illustrate that PRIME significantly advances the corruption robustness over that of AugMix and DeepAugment on all the benchmarks[3].

4.1 Training setup

We consider the CIFAR-10 (C-10), CIFAR-100 (C-100) [25], ImageNet-100 (IN-100) and ImageNet (IN) [11] datasets. IN-100 is a 100-class subset of IN obtained by selecting every 10^{th} class in WordNet ID order. We train a ResNet-18 [19] on C-10, C-100 and IN-100; and a ResNet-50 on IN for 100 epochs. Following AugMix, and for a complete comparison, we also integrate the Jensen-Shannon divergence (JSD)-based consistency loss in PRIME which compels the network to learn similar representations for differently augmented versions of the same input image. Detailed training setup appears in Appendix C. We evaluate our trained models on the common corrupted versions (C-10-C, C-100-C, IN-100-C, IN-C) of the aforementioned datasets. The common corruptions [21] constitute 15 image

[3] In Appendix K, we also show that our method yields additional benefits when employed in concert with unsupervised domain adaptation [37].

distortions each applied with 5 different severity levels. These corruptions can be grouped into four categories, viz. noise, blur, weather and digital.

4.2 Robustness to common corruptions

In order to assess the effectiveness of PRIME, we evaluate its performance against C-10, C-100, IN-100 and IN common corruptions. The results are summarized in Tab. 1[4]. Amongst individual methods, PRIME yields superior results compared to those obtained by AugMix and DeepAugment alone and advances the baseline performance on the corrupted counterparts of the four datasets. As listed, PRIME pushes the corruption accuracy by 1.2% and 3.3% on C-10-C and C-100-C respectively over AugMix. On IN-100-C, a more complicated dataset, we observe significant improvements wherein PRIME outperforms AugMix by 10.9%. In fact, this increase in performance hints that our primitive transformations are actually able to cover a larger space of image corruptions, compared to the restricted set of AugMix. Interestingly, the random transformations in PRIME also lead to a 3.9% boost in corruptions accuracy over DeepAugment despite the fact that DeepAugment leverages additional knowledge to augment the training data via its use of pre-trained architectures. Moreover, PRIME provides cumulative gains when combined with DeepAugment, reducing the mean corruption error (mCE) of prior art (DA+AugMix) by 2.7% on IN-100-C. Lastly, we also evaluate the performance of PRIME on full IN-C. However, we do not use JSD in order to reduce computational complexity. Yet, even without the JSD loss, PRIME outperforms, in terms of corruption accuracy, both AugMix (with JSD) and DeepAugment by 6.7% and 2.4% respectively, while the mCE is reduced by 7.8% and 2.9%. And last, when PRIME is combined with DeepAugment, it also surpasses the performance of DA+AugMix (with JSD), reaching a corruption accuracy of almost 60% and an mCE of 51.3%. Note here, that, not only PRIME achieves superior robustness, but it does so efficiently. Compared to standard training on IN-100, AugMix requires 1.20x time and PRIME requires 1.27x. In contrast, DA is tedious and we do not measure its runtime since it also requires the training of two large image-to-image networks for producing augmentations, and can only be applied offline.

5 Robustness insights using PRIME

In this section, we exploit the simplicity and the controllable nature of PRIME to investigate different aspects behind robustness to common corruptions. We first analyze how each transformation domain contributes to the overall robustness of the network. Then, we empirically locate and justify the benefits of mixing the transformations of each domain. Moreover, we demonstrate the existence of a robustness-accuracy trade-off, and, finally, we comment on the low-complexity benefits of PRIME in different data augmentation settings.

[4] We provide the per-corruption performance of every method in Appendix H.

Table 2. Impact of the different max-entropy primitives (ω: spectral, γ: color, τ: spatial) in PRIME on common corruption accuracy (↑) of a ResNet-18. All the transformations are essential for the performance of PRIME. The JSD loss is *not* used.

Transform	IN-100-C	Noise	Blur	Weather	Digital	IN-100
None	49.7	27.3	48.6	54.8	62.6	**88.0**
ω	64.1	60.7	55.4	66.6	72.9	87.3
τ	53.8	30.1	56.2	57.6	65.4	87.0
γ	59.9	67.4	52.6	54.4	67.1	86.9
ω+τ	64.5	58.5	57.3	**66.8**	73.9	87.7
ω+γ	67.5	77.2	55.7	65.3	74.2	87.1
τ+γ	63.3	74.7	57.4	56.2	67.8	86.2
ω+τ+γ	**68.8**	**78.8**	**58.3**	66.0	**74.8**	87.1

5.1 Contribution of transformations

We want to understand how the transformations in each domain of Eq. (1) contribute to the overall robustness. To that end, we conduct an ablation study on IN-100-C by training a ResNet-18 with the max-entropy transformations of PRIME individually or in combination. As shown in Tab. 2, spectral transformations mainly help against blur, weather and digital corruptions. Spatial operations also improve on blurs, but on elastic transforms as well (digital). On the contrary, color transformations excel on noises and certain high frequency digital distortions, e.g., pixelate and JPEG artefacts, and have minor effect on weather changes. Besides, incrementally combining the transformations lead to cumulative gains e.g., spatial+color help on both noises and blurs. Yet, for obtaining the best results, the combination of all transformations is required. This means that each transformation increases the coverage over the space of possible distortions and the increase in robustness comes from their cumulative contribution.

5.2 The role of mixing

In most data augmentation methods, besides the importance of the transformations themselves, mixing has been claimed as an essential module for increasing diversity in the training process [45,44,22,41]. In our attempt to provide insights on the role of mixing in the context of common corruptions, we found out that it is capable of constructing augmented images that look perceptually similar to their corrupted counterparts. In fact, the improvements on specific corruption types observed in Tab. 2 can be largely attributed to mixing. As exemplified in Fig. 3, careful combinations of spectral transformations with the clean image introduce brightness and contrast-like artefacts that look similar to the corresponding corruptions in IN-C. Also, combining spatial transformations creates blur-like artefacts that look identical to zoom blur in IN-C. Finally, notice how mixing color transformations helps fabricate corruptions of the "noise" category. This means that the max-entropy color model of PRIME enables robustness to different types of noise without explicitly adding any during training.

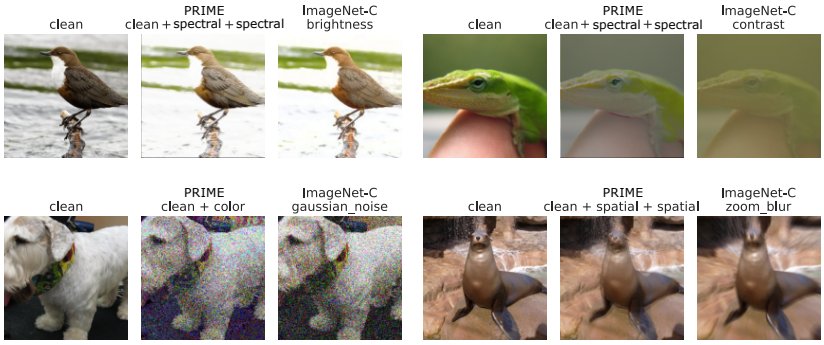

Fig. 3. Mixing produces images that are visually similar to the test-time corruptions. Each example shows the clean image, the PRIME image and the common corruption that resembles the image produced by mixing. We also report the mixing combination used for recreating the corruption. See Appendix D for additional examples.

Note that one of the main goals of data augmentation is to achieve maximum coverage of the space of possible distortions using a limited transformation budget, i.e., within a few training epochs. The principle of max-entropy guarantees this within each primitive, but the effect of mixing on the overall space is harder to quantify. In this regard, we can use the distance in the embedding space, ϕ, of a SimCLRv2 [7] model as a proxy for visual similarity [46,30]. We are interested in measuring how mixing the base transformations changes the likelihood that an augmentation scheme generates some sample during training that is visually similar to some of the common corruptions. To that end, we randomly select $N = 1000$ training images $\{\boldsymbol{x}_n\}_{n=1}^N$ from IN, along with their $C = 75$ (15 corruptions of 5 severity levels) associated common corruptions $\{\hat{\boldsymbol{x}}_n^c\}_{c=1}^C$, and generate for each of the clean images another $T = 100$ transformed samples $\{\tilde{\boldsymbol{x}}_n^t\}_{t=1}^T$ using each augmentation scheme. Moreover, for each corruption $\hat{\boldsymbol{x}}_n^c$ we find its closest neighbor $\tilde{\boldsymbol{x}}_n^t$ from the set of generated samples using the cosine distance in the embedding space. Our overall measure of fitness is

$$\frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C}\min_{t}\left\{1-\left(\frac{\phi(\hat{\boldsymbol{x}}_n^c)^\top\phi(\tilde{\boldsymbol{x}}_n^t)}{\|\phi(\hat{\boldsymbol{x}}_n^c)\|_2\ \|\phi(\tilde{\boldsymbol{x}}_n^t)\|_2}\right)\right\}. \tag{7}$$

Table 3 shows the values of this measure applied to AugMix and PRIME, with and without mixing. For reference, we also report the values of the clean (no transform) images $\{\boldsymbol{x}_n\}_{n=1}^N$. More percentile scores can be found in Appendix F. Clearly, mixing helps reduce the distance between the common corruptions and the augmented samples from both methods. We also observe that PRIME, even with only 100 augmentations per image – in the order of the number of training epochs – can generate samples that are twice as close to the common corruptions as AugMix. In fact, the feature similarity between training augmentations and test corruptions was also studied in [29], with an attempt to justify the good performance of AugMix on C-10. Yet, we see that the fundamental transformations

Table 3. Minimum cosine distances in the ResNet-50 SimCLRv2 embedding space between 100 augmented samples from 1000 ImageNet images, and their corresponding common corruptions.

Method	Min. cosine distance ($\times 10^{-3}$)	
	Avg. (↓)	Median (↓)
None (clean)	25.38	6.44
AugMix (w/o mix)	20.57	3.56
PRIME (w/o mix)	**10.61**	**1.88**
AugMix	17.48	2.61
PRIME	**7.71**	**1.61**

of AugMix are not enough to span a broad space guaranteeing high perceptual similarity to IN-C. The significant difference in terms of perceptual similarity in Tab. 3 between AugMix and PRIME may explain the superior performance of PRIME on IN-100-C and IN-C (cf. Tab. 1)[5].

5.3 Robustness vs. accuracy trade-off

An important phenomenon observed in the literature of adversarial robustness is the so-called robustness-accuracy trade-off [16,40,33], where technically adversarial training [27] with smaller perturbations (typically smaller ε) results in models with higher standard but lower adversarial accuracy, and vice versa. In this sense, we want to understand if the strength of the image transformations introduced through data augmentations in PRIME can also cause such phenomenon in the context of robustness to common corruptions. As described in Sec. 2, each of the transformations of PRIME has a strength parameter σ, which can be seen as the analogue of ε in adversarial robustness. Hence, we can easily reduce or increase the strength of the transformations by setting $\hat{\sigma} = \alpha\sigma$, where $\alpha \in \mathbb{R}^+$. Then, by training a network for different values of α we can monitor its accuracy on the clean and the corrupted datasets.

We train a ResNet-18 on C-10 and IN-100 using the setup of Sec. 4.1. For reducing complexity, we do not use the JSD loss and train for 30 epochs. This sub-optimal setting could cause some performance drop compared to the results of Tab. 1, but we expect the overall trends in terms of accuracy and robustness to be preserved. Regarding the scaling of the parameters' strength, for C-10 we set $\alpha \in [10^{-3}, 10^2]$ and sample 100 values spaced evenly on a log-scale, while for IN-100 we set $\alpha \in [10^{-2}, 10^2]$ and we sample 20 values.

The results are presented in Fig. 4. For both C-10 and IN-100, it seems that there is a sweet spot for the scale around $\alpha = 0.2$ and $\alpha = 1$ respectively, where the accuracy on common corruptions reaches its maximum. For α smaller than these values, we observe a clear trade-off between validation and robust accuracy. While the robustness to common corruptions increases, the validation accuracy decays. However, for α greater than the sweet-spot values, we observe that the

[5] A visualization of the augmented space using PCA can be found in Appendix G.

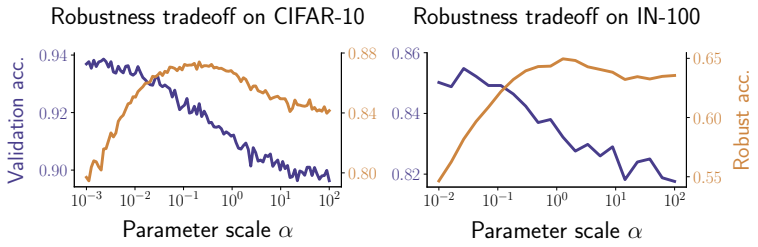

Fig. 4. Robustness vs. accuracy of a ResNet-18 (w/o JSD) on CIFAR-10 (left) and ImageNet-100 (right), when trained multiple times with PRIME. On each training instance, the transformation strength is scaled by α. Note the different scale in axes.

trade-off ceases to exist since both the validation and robust accuracy present similar behaviour (slight decay). In fact, these observations indicate that robust and validation accuracies are not always positively correlated and that one might have to slightly sacrifice validation accuracy in order to achieve robustness.

5.4 Sample complexity

Finally, we investigate the necessity of performing augmentation during training (on-line augmentation), compared to statically augmenting the dataset before training (off-line augmentation). On the one hand, on-line augmentation is useful when the dataset is huge and storing augmented versions requires a lot of memory. Besides, there are cases where offline augmentation is not feasible as it relies on pre-trained or generative models which are unavailable in certain scenarios, e.g., DeepAugment [20] or AdA [5] cannot be applied on C-100. On the other hand, off-line augmentation may be necessary to avoid the computational cost of generating augmentations during training.

To this end, for each of the C-10 and IN-100 training sets, we augment them off-line with $k = 1, 2, \ldots, 10$ i.i.d. PRIME transformed versions. Then, for different values of k, we train a ResNet-18 on the corresponding augmented dataset and report the validation and common corruption accuracy. For the training setup, we follow the settings of Sec. 4.1, but without JSD loss. Finally, since we increase the training set size by $(k+1)$, we divide the number of training epochs by the same factor to keep the same overall number of gradient updates.

The performance on common corruptions is presented in Fig. 5. Notice that, even for $k = 1$, the robustness to common corruptions is already quite good. In fact, for IN-100 the accuracy (65%) is already better than AugMix (60.7% with JSD loss cf. Tab. 1). Regarding C-10, we see that for $k = 4$ the actual difference with respect to the on-line augmentation is almost negligible (88.8% vs. 89.3%), especially considering the overhead of transforming the data at every epoch. Technically, this means that augmenting C-10 with 4 PRIME counterparts is enough for achieving good robustness to common corruptions. Finally, we also see in Fig. 5 that the corruption accuracy on IN-100 presents a very slow improvement after $k = 4$. Comparing the accuracy at $k = 4$ (67.2%) to the one obtained with on-line augmentation and without JSD (68.8% cf. Tab. 2) we observe a

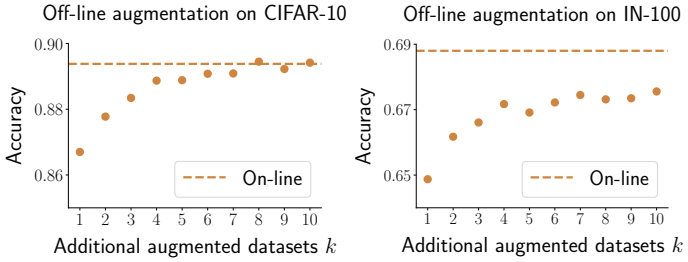

Fig. 5. Accuracy of a ResNet-18 (w/o JSD) on CIFAR-10 (left) and ImageNet-100 (right) when augmenting the training sets with additional PRIME counterparts off-line. Dashed lines represent the accuracy achieved by training under the same setup, but generating the transformed samples during training (on-line augmentation). Validation accuracy is omitted because it is rather constant: around 93.4% for CIFAR-10 and around 87% for ImageNet-100.

gap of 1.6%. Hence, given the cost of on-line augmentation on such large scale datasets, simply augmenting the training with 4 extra PRIME samples presents a good compromise for achieving competitive robustness. Still, the 1.6% increase introduced by on-line augmentation is rather significant, hinting that generating transformed samples during training might be necessary for maximizing performance. In this regard, the lower computational complexity of PRIME allows it to easily achieve this +1.6% gain through on-line augmentation, as it only requires 1.27× additional training time compared to standard training, and only 1.06× compared to AugMix, but with much better performance. This can be a significant advantage with respect to complex methods, like DeepAugment, that cannot be even applied on-line (require heavy pretraining).

6 Related work

Common corruptions Towards evaluating the robustness of deep neural networks (DNNs) to natural distribution shifts, the authors in [21] proposed common corruptions benchmarks (CIFAR-10-C and ImageNet-C) constituting 15 realistic image distortions. Later studies [20] considered the example of blurring and demonstrated that performance improvements on these common corruptions do generalize to real-world images, which supports the use of common corruptions benchmarks. Recent work [29] showed that current augmentation techniques undergo a performance degradation when evaluated on corruptions that are perceptually dissimilar from those in ImageNet-C. In addition to common corruptions, current literature studies other benchmarks e.g., adversarially filtered data [23], artistic renditions [20] and in-domain datasets [34]. In Appendix J, we show that PRIME also improves robustness on these benchmarks.

Improving corruption robustness Data augmentation has been a central pillar for improving the generalization of DNNs [12,45,10,44,26]. A notable augmentation scheme for endowing corruption robustness is AugMix [22], which

employs a careful combination of stochastic augmentation operations and mixing. AugMix attains significant gains on CIFAR-10-C, but it does not perform as well on larger benchmarks like ImageNet-C. DeepAugment (DA) [20] addresses this issue and diversifies the space of augmentations by introducing distorted images computed by perturbing the weights of image-to-image networks. DA, combined with AugMix, achieves the current state-of-the-art on ImageNet-C. Other schemes include: (i) worst-case noise training [35] or data augmentation through Fourier-based operations [38], (ii) inducing shape bias through stylized images [17], (iii) adversarial counterparts of DeepAugment [5] and AugMix [41], (iv) pre-training and/or adversarial training [43,24], (v) constraining the total variation of convolutional layers [36] or compressing the model [13] and (vi) learning the image information in the phase rather than amplitude [6] Besides, Vision Transformers [15] have been shown to be more robust to common corruptions than standard CNNs [3,31] when trained on big data. It would thus be interesting to study the effect of extra data alongside PRIME in future works. Finally, unsupervised domain adaptation [2,37] using a few corrupted samples has also been shown to provide a considerable boost in corruption robustness. Nonetheless, domain adaptation is orthogonal to this work as it requires knowledge of the target distribution.

7 Concluding remarks

We took a systematic approach to understand the notion of common corruptions and formulated a universal model that encompasses a wide variety of semantic-preserving image transformations. We then proposed a novel data augmentation scheme called *PRIME*, which instantiates our model of corruptions, to confer robustness against common corruptions. From a practical perspective, our method is principled yet efficient and can be conveniently incorporated into existing training procedures. Moreover, it yields a strong baseline on existing corruption benchmarks outperforming current standalone methods. Additionally, our thorough ablations demonstrate that diversity among basic augmentations (primitives) – which AugMix and other approaches lack – is essential, and that mixing plays a crucial role in the success of both prior methods and PRIME. In general, while complicated methods like DeepAugment perform well, it is difficult to understand, ablate and apply these online. Instead, we show that a simple model-based stance with a few guiding principles can be used to build a very effective augmentation scheme that can be easily understood, ablated and tuned. We believe that our insights and PRIME pave the way for building robust models in real-life scenarios. PRIME, for instance, provides a ready-to-use recipe for data-scarce domains such as medical imaging.

Acknowledgments We thank Alessandro Favero for the fruitful discussions and feedback. This work has been partially supported by the CHIST-ERA program under Swiss NSF Grant 20CH21_180444, and partially by Google via a Postdoctoral Fellowship and a GCP Research Credit Award.

References

1. Beale, P.: Statistical Mechanics. Elsevier (1996)
2. Benz, P., Zhang, C., Karjauv, A., Kweon, I.S.: Revisiting batch normalization for improving corruption robustness. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021)
3. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A.: Understanding robustness of Transformers for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
4. Binder, F., Aichinger, E., Ecker, J., Nöbauer, C., Mayr, P.: Algorithms for near-rings of non-linear transformations. In: Proceedings of the International Symposium on Symbolic and Algebraic Computation. Association for Computing Machinery (2000)
5. Calian, D.A., Stimberg, F., Wiles, O., Rebuffi, S.A., Gyorgy, A., Mann, T., Gowal, S.: Defending against image corruptions through adversarial augmentations. arXiv preprint arXiv:2104.01086 (2021)
6. Chen, G., Peng, P., Ma, L., Li, J., Du, L., Tian, Y.: Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
7. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems (2020)
8. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience (2006)
9. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2021)
10. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009)
12. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
13. Diffenderfer, J., Bartoldson, B.R., Chaganti, S., Zhang, J., Kailkhura, B.: A winning hand: Compressing deep networks can improve out-of-distribution robustness. In: Advances in Neural Information Processing Systems (Dec 2021)
14. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX) (2016)
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
16. Fawzi, A., Fawzi, O., Frossard, P.: Analysis of classifiers' robustness to adversarial perturbations. Machine Learning **107**(3), 481–508 (2018)

17. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2019)
18. Geirhos, R., Temme, C.R.M., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. In: Advances in Neural Information Processing Systems (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (2016)
20. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: IEEE Conference on Computer Vision and Pattern Recognition (2021)
21. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
22. Hendrycks*, D., Mu*, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple method to improve robustness and uncertainty under data shift. In: International Conference on Learning Representations (2020)
23. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
24. Kireev, K., Andriushchenko, M., Flammarion, N.: On the effectiveness of adversarial training against common corruptions. arXiv preprint arXiv:2103.02325 (2021)
25. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009)
26. Lopes, R.G., Yin, D., Poole, B., Gilmer, J., Cubuk, E.D.: Improving robustness without sacrificing accuracy with patch gaussian augmentation. arXiv preprint arXiv:1906.02611 (2019)
27. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (Apr 2018)
28. Masiha, M.S., Gohari, A., Yassaee, M.H., Aref, M.R.: Learning under distribution mismatch and model misspecification. In: IEEE International Symposium on Information Theory, (ISIT) (2021)
29. Mintun, E., Kirillov, A., Xie, S.: On interaction between augmentations and corruptions in natural corruption robustness. arXiv preprint arXiv:2102.11273 (2021)
30. Moayeri, M., Feizi, S.: Sample efficient detection and classification of adversarial attacks via self-supervised embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
31. Morrison, K., Gilby, B., Lipchak, C., Mattioli, A., Kovashka, A.: Exploring corruption robustness: Inductive biases in vision transformers and mlp-mixers. arXiv preprint arXiv:2106.13122 (2021)
32. Petrini, L., Favero, A., Geiger, M., Wyart, M.: Relative stability toward diffeomorphisms indicates performance in deep nets. In: Advances in Neural Information Processing Systems (2021)
33. Raghunathan, A., Xie, S.M., Yang, F., Duchi, J., Liang, P.: Understanding and mitigating the tradeoff between robustness and accuracy. In: Proceedings of the 37th International Conference on Machine Learning (Jul 2020)
34. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: Proceedings of the 36th International Conference on Machine Learning (2019)

35. Rusak, E., Schott, L., Zimmermann, R.S., Bitterwolf, J., Bringmann, O., Bethge, M., Brendel, W.: A simple way to make neural networks robust against diverse image corruptions. In: Computer Vision – ECCV 2020 (2020)
36. Saikia, T., Schmid, C., Brox, T.: Improving robustness against common corruptions with frequency biased models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
37. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. In: Advances in Neural Information Processing Systems (2020)
38. Sun, J., Mehra, A., Kailkhura, B., Chen, P.Y., Hendrycks, D., Hamm, J., Mao, Z.M.: Certified adversarial defenses meet out-of-distribution corruptions: Benchmarking robustness and simple baselines. arXiv preprint arXiv:arXiv:2112.00659 (2021)
39. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. In: Advances in Neural Information Processing Systems (2020)
40. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: International Conference on Learning Representations (May 2019)
41. Wang, H., Xiao, C., Kossaifi, J., Yu, Z., Anandkumar, A., Wang, Z.: Augmax: Adversarial composition of random augmentations for robust training. In: Advances in Neural Information Processing Systems (2021)
42. Xu, A., Raginsky, M.: Information-theoretic analysis of generalization capability of learning algorithms. In: Advances in Neural Information Processing Systems (2017)
43. Yi, M., Hou, L., Sun, J., Shang, L., Jiang, X., Liu, Q., Ma, Z.: Improved OOD generalization via adversarial training and pretraining. In: Proceedings of the 86th International Conference on Machine Learning (2021)
44. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: 2019 IEEE/CVF International Conference on Computer Vision (2019)
45. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)