Rethinking Data Augmentation for Adversarial Robustness

Hamid Eghbalzadeh[a,*], Werner Zellinger[b], Maura Pintor[c,**], Kathrin Grosse[d], Khaled Koutini[a], Bernhard A. Moser[e], Battista Biggio[c], Gerhard Widmer[a]

[a] *LIT AI Lab & Institute of Computational Perception, Johannes Kepler University of Linz, Austria*
[b] *Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences*
[c] *University of Cagliari*
[d] *École Polytechnique Fédérale de Lausanne*
[e] *Software Competence Center Hagenberg*

Abstract

Recent work has proposed novel data augmentation methods to improve the adversarial robustness of deep neural networks. In this paper, we re-evaluate such methods through the lens of different metrics that characterize the augmented manifold, finding contradictory evidence. Our extensive empirical analysis involving 5 data augmentation methods, all tested with an increasing probability of augmentation, shows that: (i) novel data augmentation methods proposed to improve adversarial robustness only improve it when combined with classical augmentations (like image flipping and rotation), and even worsen adversarial robustness if used in isolation; and (ii) adver-

*Work was carried out during the full time employment of the first author at the Johannes Kepler University.

**Corresponding author
Email address: maura.pintor@unica.it (Maura Pintor)

sarial robustness is significantly affected by the augmentation probability, conversely to what is claimed in recent work. We conclude by discussing how to rethink the development and evaluation of novel data augmentation methods for adversarial robustness. Our open-source code is available at `https://github.com/eghbalz/rethink_da_for_ar`.

Keywords: adversarial machine learning, data augmentation

1. Introduction

To achieve good generalization in machine learning (ML), large data is needed, often more than is available [1, 2, 3]. Data augmentation (DA) tackles this problem by applying randomly constructed transformations on the input data to increase the diversity and size of the training set. For example, DAs on images can rely on image modifications like rotations, horizontal/vertical flips, scaling and cropping, with the purpose of covering multiple variations that can be encountered in real scenarios. Beyond yielding more data, DA can have a regularization effect for some combinations of augmentation and ML methods, including regression [4], kernel methods [5], and deep learning [6]. Additional positive effects of DA consist in reducing dataset bias [7], improving accuracy [8, 9], and enhancing algorithmic fairness [10].

Since ML has been repeatedly proven vulnerable to *adversarial examples*, i.e. carefully-perturbed inputs aimed to mislead classification at test time) [16, 17], recent research has focused on creating DAs techniques that also tackle this security problem. For instance, Rebuffi et al. [3] showed that newly-proposed

heuristic DA methods like MixUp [11], CutMix [12], ManifoldMixUp [13], and CutOut [14], as well as *generative* DA methods like Diffusion Models [15] are able to improve *adversarial robustness*, namely, the ability of the model to withstand *adversarial examples*. However, all these approaches have been tested only under the two following *implicit* assumptions; in particular, they have been tested: (i) in combination with classical augmentations (e.g., rotation, flipping, color-jittering), and (ii) using a fixed fraction of augmented samples (i.e., a fixed *augmentation probability*). It thus remains an open question to understand whether the claims made in previous work still hold outside of the two aforementioned working assumptions.

In this work, we propose a unifying framework that provides a fair, systematic reevaluation of these methods, with the goal of testing whether the claims made in previous work hold under more general conditions, i.e., (i) whether newly-proposed DAs also work in isolation, without requiring classical augmentation, and (ii) whether they work under different augmentation probabilities. Our main goal is to point out biases in the evaluation of DA techniques that could be avoided through the application of our evaluation framework. Our main contribution is thus to provide an effective framework and set of guidelines that focus on evaluating new DA techniques effectively. To this end, we first summarize the claims and empirical findings from previous work, which result in the following two working hypotheses:

Hypothesis 1. *Newly-proposed heuristic and generative DAs increase adversarial robustness [14, 18, 13, 12, 2].*

Hypothesis 2. *Varying the fraction of augmented samples does not significantly affect generalization and adversarial robustness [19, 14, 18, 13, 12, 2].*

In this work, we shed light on these questions by first reviewing and categorizing DAs, as well as the conditions under which these DAs have been tested (Section 2). We then propose a unifying framework to re-evaluate such augmentation techniques through the lens of different metrics that characterize the augmented manifold, aiming at verifying the aforementioned claims (Section 3). Our evaluation framework consists of three main components: (i) a performance-vs-robustness analysis of DAs, which decouples the impact of heuristic, data-driven, and classical augmentations on adversarial robustness with different augmentation probabilities; and two additional metrics named (ii) decision-function roughness and (iii) data-augmentation spuriousness to further support our findings and provide additional insights into how DAs impact adversarial robustness. With this novel evaluation framework, we test existing methods in a completely new way that has not been considered in any of the papers we propose, showing that the factors that individually contribute to robustness differ from those originally claimed. Our extensive empirical

analysis, involving 5 DA methods tested with 10 different augmentation probabilities, shows evidence that contradicts previous studies (Section 4).

We conclude the paper by discussing our findings' relevance for designing novel DA methods for adversarial robustness (Section 5).

This is crucial because the contradictory evidence from prior work demands the adoption of a proper evaluation framework and a common benchmark for DA methods, especially when it comes to evaluating their adversarial robustness, and we firmly believe that our work provides an important first contribution in this direction.

2. Data Augmentation Methods

In this section, we present background in DA and review related works. We first discuss different DA techniques, divided into (i) classical approaches, like rotation or cropping; and (ii) novel approaches, which include heuristic augmentations and data-driven augmentations based on generative methods. To conclude the section, we review existing works that connect DA with adversarial robustness.

2.1. Classical Augmentations

Incorporating domain knowledge of experts in models by using DA has been one of the main approaches to improve performance and generalization. From the early days of deep learning, simple geometrical transformations have been utilized as data augmentation with great success [9]. For example,

in the vision domain, horizontal flipping, random rotations, as well as slight change in brightness, contrast, and saturation of natural images were used. In particular, the latter group simulates different camera angles and lighting conditions, which are known to preserve the main characteristics of the data w.r.t. the task at hand [19, 20]. In the audio domain, expert-based DAs include careful time and frequency shifting, additive noise [21], as well as time and frequency masking [22] of the audio signals. Such DAs resemble realistic scenarios like the use of frequency band-pass filters, higher-pitched sounds with existing acoustic characteristics, and background noise.

2.2. Novel Data Augmentation Methods

Heuristic Augmentations. The design of classical DAs requires domain knowledge and a deep understanding of the task and data at hand. Several efforts have been made in order to introduce DAs based on more general heuristics that are domain- or task-independent, as opposed to classical augmentations. For example, MixUp [11] is a heuristic-based augmentation that creates new samples by linearly combining existing data points and their labels to favor simple linear behavior in-between training examples. CutOut [14] instead removes certain areas of the input and thus is performed to improve resilience against missing data. CutMix [12] combines the two previous heuristics and generates new data by mixing cut-out regions into existing samples to further enhance the localization ability of the models by requiring them to identify the object from a partial view. Finally, Manifold-

MixUp [13] not only mixes data and labels but does so on the intermediate representations of the neural network, encouraging it to predict less confidently on interpolations of hidden representations.

Data-Driven Augmentations. Another perspective on DA is to use existing data to *learn* suitable transformation strategies. Generative models such as GANs [23], VAEs [24], and Denoising Diffusion models [15] have recently been used for DA [25, 26, 27, 2]. However, Rebuffi et al. [3] have shown that the latter diffusion models are more successful in terms of generalization and adversarial robustness. In addition to generative models, other data-driven approaches exist. For example, Augerino [28] learns parametrized affine transformations from data within the borders of robust augmentations. Another example is AutoAugment [29], which uses reinforcement learning to fine-tune the hyperparameters of the augmentations.

2.3. Data Augmentation and Robustness

While achieving high accuracy might seem the main objective of DA, recent work also investigates the security properties of these techniques. In particular, this accounts to quantifying the ability of the models to classify correctly adversarially-perturbed images. Adversarial examples are carefully-manipulated test samples that increase the misclassification error of ML models. These perturbations are created through specific iterative algorithms, like Projected Gradient Descent (PGD) [41], that optimize worst-case losses over controlled modifications of the input. As DAs improve the generalization

of the models and encourage learning stronger features, investigating whether DAs also improve *adversarial* robustness is the subject of ongoing research.

We provide an overview in Table 1. Many recently-proposed heuristic and data-driven DAs were shown to increase adversarial robustness [11, 13, 12, 14]. However, at the same time, there are inductive biases in these studies. In all cases, the studied DA (i) is combined with classical augmentations, and (ii) is tested only on one augmentation probability choice (mostly 0.5, sometimes 1.0, and occasionally a fixed linearly increasing regime). In some works, DA is even further combined with *adversarial training*, i.e., including adversarial examples into the training set. Consequently, it is not clear which factor is really contributing to improving or degrading adversarial robustness, and to what extent.

In this work, we overcome this limitation by proposing a comprehensive framework that properly assesses the robustness of DAs. Our framework decouples the effect of each of the aforementioned factors and highlights the real impact of heuristic and data-driven DAs on adversarial robustness.

3. Evaluating Data Augmentation for Adversarial Robustness

In this section, we first introduce the learning setup and notation (Section 3.1) and then present our evaluation framework consisting of three main components: (i) the performance-vs-robustness analysis of DAs (Section 3.2); (ii) the decision-function roughness (Section 3.3); and (iii) the data-augmentation spuriousness (Section 3.4). We study the first to decouple

Table 1: DA approaches and their effect on robustness and accuracy. For each DA, we report whether it is shown to improve (↑), worsen (↓), or not affect (−) clean accuracy (Acc.) and robustness (Rob.) to adversarial attacks (ADV, either FGSM or PGD), corruptions (COR) or deformations (DEF). We also report if the proposed DA is combined with classic augmentation (cls), adversarial training (AT), and the augmentation probability (P) used, where "lin" denotes the linear strategy used by CutMix to increase the augmentation probability throughout training.

Reference	DA	Rob.	Acc.	cls	P	AT
Zhang et al. [11]	**MixUp** (heuristic)	↑ ADV	↑	✓	1.	
Verma et al. [13]		↑ ADV	↑	✓	1.	
Yun et al. [12]		↑ ADV	−	✓	1.	
Guo et al. [18]		−	↓	✓	1.	
Rebuffi et al. [3]		↓ ADV	−	✓	1.	✓
Yun et al. [12]	**CutMix** (heuristic)	↑ ADV	↑	✓	lin	
Rebuffi et al. [3]		− ADV	−	✓	lin	✓
Hendrycks et al. [30]		↓ COR	−	✓	lin	
Devries et al. [14]	**CutOut** (heuristic)	−	↑	✓	0.5	
Hendrycks et al. [30]		↓ COR	−	✓	0.5	
Rebuffi et al. [3]		↑ ADV	−	✓	0.5	✓
Rebuffi et al. [3]	**Gen** (data.)	↑ ADV	↑	✓	0.9	✓
Nakkiran et al. [2]		−	↑	✓	0.9	
Verma et al. [13]	**Man** (heu.)	↑ ADV	↑	✓	1.	
Verma et al. [13]		↑ DEF	↑	✓	1.	

the effect of heuristic, data-driven, and classical augmentations on robustness and performance (with different augmentation probabilities), and the latter two to provide additional insights on how DAs impact adversarial robustness.

3.1. Learning Setup and Notation

Throughout this work, let X be a random variable on a probability space $(\mathcal{X}, \mathcal{A}, P)$ with sigma algebra $\mathcal{A}$ and input space $\mathcal{X} \subset \mathbb{R}^d$, e.g. images, and denote by P the probability measure of X. Further, let $l : \mathcal{X} \to \mathcal{Y}$ be a labeling function to a finite set $\mathcal{Y} \subset \mathbb{N}$ of labels, e.g. $\{1, \dots, c\}$. Given a class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$ and a sample $S = ((\mathbf{x}_1, l(\mathbf{x}_1)), \dots, (\mathbf{x}_s, l(\mathbf{x}_s))) \in$

$(\mathcal{X} \times \mathcal{Y})^s$ with $\mathbf{x}_1, \ldots, \mathbf{x}_s$ independently drawn from P_X, the problem of *risk minimization* is to find a function $f \in \mathcal{F}$ with a low *misclassification risk* [31]:

$$R(f, l) := P(f(X) \neq l(X)). \tag{1}$$

This problem can be solved via Stochastic Gradient Descent (SGD) using some parametric function class $\mathcal{F}$ of neural networks [32].

In many practical tasks, risk minimization can be improved by applying *data augmentation* techniques. In this work, we call a random function $A : (\mathcal{X} \times \mathcal{Y})^s \to \left\{X \times Y : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d\right\}^r$ a data augmentation, if it maps the sample S to some vector $A(S) = (X_1 \times Y_1, \ldots, X_r \times Y_r)$ of independent random variables $X_1 \times Y_1, \ldots, X_r \times Y_r$ with measure $P_{X_1 \times Y_1}$ on $\mathcal{X} \times \mathcal{Y}$ such that the marginal measure P_{X_1} dominates P_X, i.e., the sample S is included in the augmented sample $\tilde{S}$ observed from the random variable $A(S)$.

As discussed in Section 2, prior work reported that training on augmented samples leads to models with lower *adversarial risk* when compared to models obtained without data augmentation. One classical measure for *adversarial risk* is the *risk under corrupted inputs* [33, 34]:

$$R_{\text{cor}}(f, l, \epsilon) := P(\exists\, \mathbf{x} \in B_\epsilon(X) : f(\mathbf{x}) \neq l(X)), \tag{2}$$

with $B_\epsilon(\mathbf{x}) := \{\mathbf{x}' \in \mathbb{R}^d \mid \|\mathbf{x}' - \mathbf{x}\| \leq \epsilon\}$. One common approach to approximate Equation 2, which we follow in Section 3.2, is to apply adversarial attacks on test samples to compute the expected risk empirically. Another measure

for adversarial risk is the *prediction-change risk* [17], i.e., the probability that a sample is classified differently within the given ϵ-ball:

$$R_{\text{pc}}(f, \epsilon) := P(\exists\, \mathbf{x} \in B_\epsilon(X) : f(\mathbf{x}) \neq f(X)). \tag{3}$$

Note that the main difference between Equation 1 and Equation 2 is that the latter does not include the labeling function but characterizes more generally the variability of the decision function. We are interested in this effect as we want to investigate whether the augmentations cause more irregular and complex decision functions, which can be associated with an increased adversarial vulnerability. On a high level, the decision-change risk is related to the shape of the decision surface of a classifier which is known to influence generalization [35, 36] and adversarial robustness [37, 38, 39, 40]. To provide an estimate of the prediction-change risk, we propose a new approximation for Equation 3 in Section 3.3.

3.2. Performance-vs-Robustness Analysis

We introduce here our performance-vs-robustness analysis. The underlying idea is to separately evaluate the impact on performance and robustness of newly-proposed heuristic and data-driven augmentations compared with classical augmentations. Furthermore, we incorporate the analysis of these techniques on a range of augmentation probabilities. Our goal is to understand whether such newly-proposed heuristic and data-driven augmentations are really responsible for improving adversarial robustness and performance, or

if, instead, the performances are mostly due to the combination with classical augmentations. Moreover, we also evaluate how using different augmentation probabilities affects performance and robustness.

To this end, we propose to look at two axes at once: (i) the classification error, and (ii) the adversarial vulnerability. We first evaluate each newly-proposed heuristic and data-driven DA *without* using any classical augmentation, and then in combination with different augmentation probabilities for classical augmentations. A DA technique is deemed useful if it pushes the corresponding point towards the origin of this plot (i.e., towards reducing both classification error and adversarial vulnerability). As we will see in the experimental section, some of the newly-proposed heuristic and data-driven DAs even *worsen* robustness and performance.

We measure the performance of the models by estimating the misclassification risk[1] in Equation 1 of a model f for a set of test samples $\tilde{S} := ((\mathbf{x}'_1, l(\mathbf{x}'_1)), \ldots, (\mathbf{x}'_s, l(\mathbf{x}'_s)))$ by computing

$$\widehat{R}(f, \tilde{S}) := \frac{1}{s} \sum_{i=1}^{s} \mathbf{1}[f(\mathbf{x}_i) \neq l(\mathbf{x}_i)], \tag{4}$$

where $\mathbf{1}[a \neq b] = 1$ if the prediction matches the label, and 0 otherwise.

We measure the adversarial vulnerability of the models by estimating the risk under corrupted inputs in Equation 2. For estimating Equation 2

[1]We use the misclassification risk instead of misclassification error functions such as the cross-entropy loss as this better reflects the objective of minimizing misclassification rather than optimizing the model's predictions on the true probability distributions.

we rely on adversarial attacks. More concretely, we compute an adversarial example for an input $\mathbf{x}^0 \in \mathcal{X}$ using Projected Gradient Descent (PGD) [41], that iteratively updates the adversarial perturbation as follows:

$$\mathbf{x}^t = \Pi_\epsilon \left(\mathbf{x}^{t-1} + \alpha \operatorname{sgn}(\nabla_{\mathbf{x}} L(f(\mathbf{x}), l(\mathbf{x}))) \right), \tag{5}$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss, $\alpha > 0$ is the step-size and Π_ϵ is a projection from the inputs $\mathcal{X}$ into the ball $B_\epsilon(\mathbf{x})$, and $\epsilon > 0$. [2]

We then approximate the adversarial risk in Equation 2 by computing the misclassification risk $\widehat{R}(f, \tilde{S})$ on the sample $\tilde{S} := ((\mathbf{x}_1^t, l(\mathbf{x}_1)), \ldots, (\mathbf{x}_s^t, l(\mathbf{x}_s)))$ where $\mathbf{x}_1^t, \ldots, \mathbf{x}_s^t$ are the adversarial examples obtained by iteratively applying equation 5 t times. We call $\widehat{R}(f, \tilde{S})$ the *risk under attack (RUA)*.

3.3. Decision-Function Roughness

We evaluate adversarial robustness beyond classical estimates of the risk under attack (Equation 2). In particular, we want to connect adversarial vulnerability with the shape of the decision surface learned by the model. Models showing a rougher decision surface should be more vulnerable to adversarial examples as they may tend to change the prediction in a small neighborhood of the input sample. We thus propose a novel approximate measure called *decision-function roughness* to estimate the prediction-change risk defined in Equation 3.

[2] We refer to the ϵ and α used in PGD as PGD_ϵ and PGD_α

Existing, similar measures use dimensionality reduction to apply noise [42], and are based on Gaussian noise [43], or rely on Gaussian noise to estimate the Jacobian of the classifier [44]. Our estimate of Equation 3 is instead based on uniform noise. Due to the phenomenon of concentration of measure [45], for high dimension d the Lebesgue measure λ in a ball $B_r^{(d)}$ of radius r is concentrated at its surface. That is, for any $\delta > 0$, we have $\lim_{d\to\infty}(\lambda(B_r^{(d)}) - \lambda(B_{r-\delta}^{(d)}))/\lambda(B_r^{(d)}) = 1$ given any L_p-norm ($p \in [1, \infty]$) we take. As the dimension d in most applications is high, the concentration of measure effect motivates to approximate Equation 3 by sampling only from the shell ∂B_ϵ (points at distance ϵ from the center of the ball) instead of sampling from the full ball B_ϵ. By considering only the points on the boundary ∂B_ϵ of the ball B_ϵ, we introduce the *decision-function roughness* as

$$\hat{r}(f, \tilde{S}, \epsilon) := \frac{1}{s}\sum_{i=1}^{s}\frac{1}{n}\sum_{j=1}^{n}\mathbf{1}[f(\mathbf{x}'_i) \neq f(\mathbf{y}_{ij})]\,, \tag{6}$$

where $\tilde{S} := ((\mathbf{x}'_1, l(\mathbf{x}'_1)), \ldots, (\mathbf{x}'_s, l(\mathbf{x}'_s)))$ is a given labeled dataset, and $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{in}$ are uniformly drawn from $\partial B_\epsilon(\mathbf{x}'_i)$. Note that $\hat{r}(f, \tilde{S}, \epsilon) \approx R_{\text{pc}}(f, \epsilon)$ for sufficiently high dimension d and sample size s. In a nutshell, the rougher the decision function, the less robust is the model at point $\mathbf{x}_0$.

3.4. Data-Augmentation Spuriousness

To gain a better understanding of how DA affects robustness, we measure *spuriousness* as the fraction of the augmented data having their closest neighbor in a spurious set. We use the set of *non-robust features* described by

Ilyas et al. [46], which are spurious features generated using the activations of a non-robust model. These are features that are highly predictive yet meaningless for humans. These features, however, might be exploited in adversarial example crafting as they are highly correlated with the labels. To put the spuriousness in perspective, we also use *robust* features as an additional dimension, which represent robust characteristics in the data space. We then measure the distance of augmented data to the manifolds to evaluate whether the augmentations help find better representations.

In order to find the distance of augmented data from the robust and non-robust manifolds, we prepare 3 sets: (i) the augmented set, (ii) the set of non-robust features, and (iii) the set of robust features. For each augmented sample among all these sets, we find the closest neighbor and we calculate the percentage of augmented samples in robust and non-robust sets as a measure of an approximated distance to their manifolds. The closeness of augmented samples to the *non-robust features* is thus an indicator for the existence of spurious features in augmented data, which could be a cause for increased adversarial risk in the models.

4. Experiments

In this section, we put our framework into practice to study DA's influence on adversarial robustness. To this end, we run experiments on CIFAR10 and CIFAR100, as these two datasets have been consistently used in the previous studies of both DA and adversarial robustness [47]. We assess

performance-vs-robustness to address Hypothesis 1 and Hypothesis 2 in Section 4.2.1, then decision-function roughness (Section 4.2.2) and data-augmentation spuriousness (Section 4.2.3) to support our claims further.

4.1. Setup

We first describe the setup for the performance-vs-robustness study, continue with the decision-function roughness setup, and conclude with the setup of the data-augmentation spuriousness experiments.

We evaluate the following DA methods: MixUp, Manifold-MixUp (Man), CutMix, CutOut, and a Denoising Diffusion Probabilistic generative model (Gen). For the latter, we rely on data from Nakkiran et al. [2], which is only provided for CIFAR10. Thus, we test this method only for this setting. Additionally, we also evaluate the classical approach (dubbed as classic), which is a random combination of rotation, color-jitter, and horizontal flipping. We organize the studied DAs into two groups. The first group consists of novel techniques not used in combination with classical DA methods. Thus, the first group includes MixUp, Man, CutMix, CutOut, and Gen used alone, while the second group includes classic, and combinations of the previous DAs with classic (dubbed as cls+MixUp, cls+Man, etc.). We also vary the probability of augmentation $p_{aug} \in \{0.1, 0.2, \ldots, 1\}$, which amounts to changing the fraction of augmented samples in the training data. For each DA method and augmentation probability p_{aug}, we train Resnet18 [48] classifiers using SGD, and average the results over three repetitions. The full list of parameters we

used for training and data augmentation are described in Appendix A.

Robustness Evaluation Setup. We test robustness against the PGD attack with L_2 and L_∞ norms. More specifically, we show results for a perturbation size that is only large enough to show differences in the adversarial risk of the trained models.[3] Results extended for other configurations are in Appendix B. The full list of parameters and settings we used for our attack can be found in Appendix A.

Roughness Setup. All used augmentations, parameters, and networks are analogous to the previous setup.

Spuriousness Setup. We measure the spuriousness for each data-augmentation method by computing the distance between the augmented data and the sets of robust and non-robust features extracted from CIFAR-10 in [46].[4] The robust feature set is built by leveraging an adversarially-robust model and removing the features that are not relevant to that model. To this end, Ilyas et al. [46] remove the features that leave the activations of the penultimate layer unchanged when setting their values to zero. Similarly, the non-robust features are built by using a non-robust model instead of the robust one. Following [3], we sample 10K images from the robust feature dataset (uniformly across classes) and 10k samples from the non-robust feature dataset. First, we want to obtain a low-dimensional representation of these samples. To this end, we

[3] $\text{PGD}_\epsilon = 0.1$ for L_2 norm, and $\text{PGD}_\epsilon = \frac{1}{255}$ for L_∞ norm.
[4] Available at `https://github.com/MadryLab/constructed-datasets`.

pass these 20K images through the pretrained VGG network, and measure the Learned Perceptual Image Patch Similarity (LPIPS) [49]. The resulting concatenated activations are used to compute the top-100 PCA components, allowing sample comparison in a lower-dimensional space (i.e., 100 instead of $124,928$ concatenated activations). Finally, for each augmentation method, we sample 10K augmented images from the CIFAR-10 training set, and pass them through the pipeline composed of the LPIPS VGG network and the PCA projection. For each sample, we find the closest neighbor in the PCA-reduced feature space, and we determine whether it belongs to the robust or non-robust sets or to the set of augmented images (self). We then calculate the number of neighbors for augmented, robust, and non-robust features. A higher number of neighbors belonging to a set (robust, non-robust) refers to a smaller two-sample distance to this set and, thus, to a higher similarity.

4.2. Results

We present our results in this section, addressing the two hypotheses with results on performance-vs-robustness, and then present results on decision-function roughness and, finally, data-augmentation spuriousness.

4.2.1. Performance-vs-Robustness Results

Hypothesis 1. In Figure 1, we plot robustness vs. performance in terms of risk under attack and misclassification risk, respectively. We first investigate whether DAs improve adversarial robustness by observing the risk under attack (vertical axis, the lower, the better). Compared to the no-augmentation

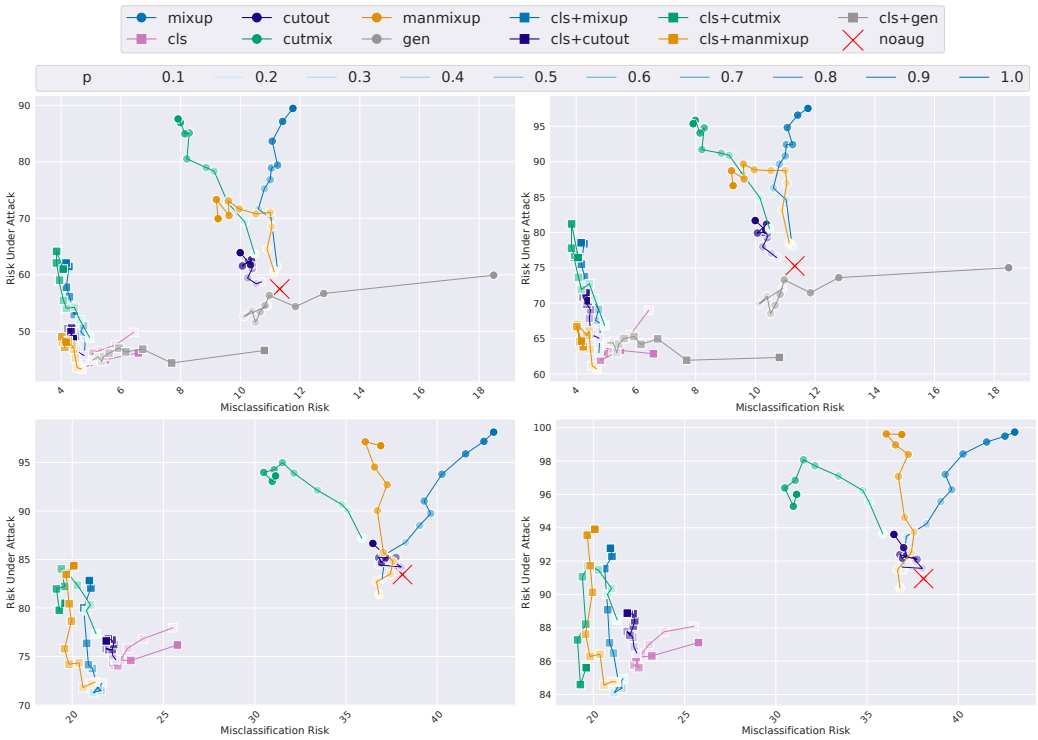

Figure 1: Robustness vs. Performance. Risk under attack vs. misclassification risk. First column: risk under L_2 perturbation model,$\text{PGD}_\epsilon = 0.1$. Second column: risk under L_∞ perturbation model,$\text{PGD}_\epsilon = \frac{1}{255}$. First row: CIFAR10. Second row: CIFAR100.

(noaug) baseline (red cross), in the first group, containing novel techniques used alone (round markers), the robustness significantly degrades in MixUp (blue), Man (yellow), and CutMix (green). Robustness also slightly decreases in CutOut (purple) as the augmentation probability increases (more intense color). Gen (grey) is the only method from the first group that slightly improves robustness. From the second group, i.e. classic DAs and combinations of novel techniques with classic (square markers), adding classic DAs improves robustness by a large margin, also when combined with the other DAs.

Concerning misclassification risk (horizontal axis, the lower the better), the

first group's risk decreases for Man (yellow) and CutMix (green). For CutOut (purple), the performance improvement is only small; while performance degrades strongly in Gen (gray) and slightly in MixUp (blue). In the second group, all methods significantly improve classification performance compared to noaug, with the exception of cls+Gen (grey) which, using the highest augmentation probability, performs on par with noaug.

Finally, all methods show higher performance and robustness when combined with classic, compared to when they are applied alone. The table shows how the robustness consistently decreases for group 1 (the DAs alone), but increases for group 2 (classic and combinations therewith). Through the evidence collected and summarized in Figure 1, we thus reject Hypothesis 1, i.e., that other augmentations than classic increase robustness.

Finding 1. *We reject Hypothesis 1, providing evidence that increased robustness is* ***only*** *achieved when the tested method is combined with classic DA. In particular, when the proposed DA is applied alone, it often results in even* ***worse*** *adversarial robustness.*

Hypothesis 2. We now investigate whether the augmentation percentage influences generalization and robustness. We again turn to Figure 1. In the plots, the colored lines (from light transparency to strong opacity) denote the increase of the augmentation probability p_{aug}. The plots support that p_{aug} has a significant effect on robustness and performance, which is more pronounced when non-classical approaches are used in isolation. In terms of robustness,

compared to a baseline without DA (noaug), increasing the amount of augmentation p_{aug} in the first group (round markers) significantly reduces robustness in MixUp (blue), Man (yellow), and CutMix (green), and slightly reduces in CutOut (purple). In the second group (square markers), the augmentation probability has some effect on cls+MixUp(blue) and cls+CutMix(green). This effect is however weaker than in the first group, showing that classic DAs are less impacted by the choice of p_{aug}. In terms of classification performance, compared to 'noaug' (red cross), increasing the amount of augmentation p_{aug} slightly decreases the classification performance in MixUp, while in CutMix and Man the performance is improved. CutOut slightly improves performance in comparison to noaug, and Gen significantly increases performance with higher p_{aug}. In the second group, the augmentation probability has some effect on the combinations of cls+CutMix, cls+Man, and cls+MixUp. This effect is however weaker than in the first group. The only technique that is overall, or in both groups, largely unaffected in robustness and performance from the augmentation probability is CutOut. In short, the augmentation percentage p_{aug} has a significant influence on almost all augmentations, but in particular for DAs other than classic and when applied in isolation. We further summarize our findings in Table 2, where we report a compact view of the results illustrated in Figure 1.

Table 2: Results for our test on Hypothesis 1. We provide the overall trend compared to the baseline (no augmentation) for each DA, where DAs are divided into two groups, depending on whether they are combined with a classical approach (w/Cls). We further denote whether robustness (Rob.) or accuracy (Acc.) increase slightly (↑) or significantly (↑↑), stay the same (−), or decrease slightly (↓) or significantly (↓↓). Finally, we summarize whether the effect of the augmentation probability (aug.P.) is negligible (−), small (+), or strong (++).

		Rob.	Acc.	aug.P.	w/ Cls
Group 1	**MixUp**	↓↓	↓	++	No
	Man	↓↓	↑↑	+	No
	CutMix	↓↓	↑↑	++	No
	CutOut	↓	↑	−	No
	Gen	−	↓↓	++	No
Group 2	**classic**	↑↑	↑↑	+	No
	cls+MixUp	↑	↑↑	+	Yes
	cls+Man	↑	↑↑	+	Yes
	cls+CutMix	↑	↑↑	+	Yes
	cls+CutOut	↑↑	↑↑	−	Yes
	cls+Gen	↑	↑	++	Yes

Finding 2. *We reject Hypothesis 2 by providing extensive results on several datasets and DAs, demonstrating the* ***significant effect*** *of the percentage of augmented samples on both the generalisation ability and the adversarial robustness of the models. We remark that this strongly contrasts with the use of fixed augmentation percentage choices in the literature.*

4.2.2. Decision-function Roughness Results

To investigate the effect of the previously studied DAs on the shape of the decision surface of models, we compare our decision-function roughness measure with the risk under attack in Figure 2. As before, the two groups of

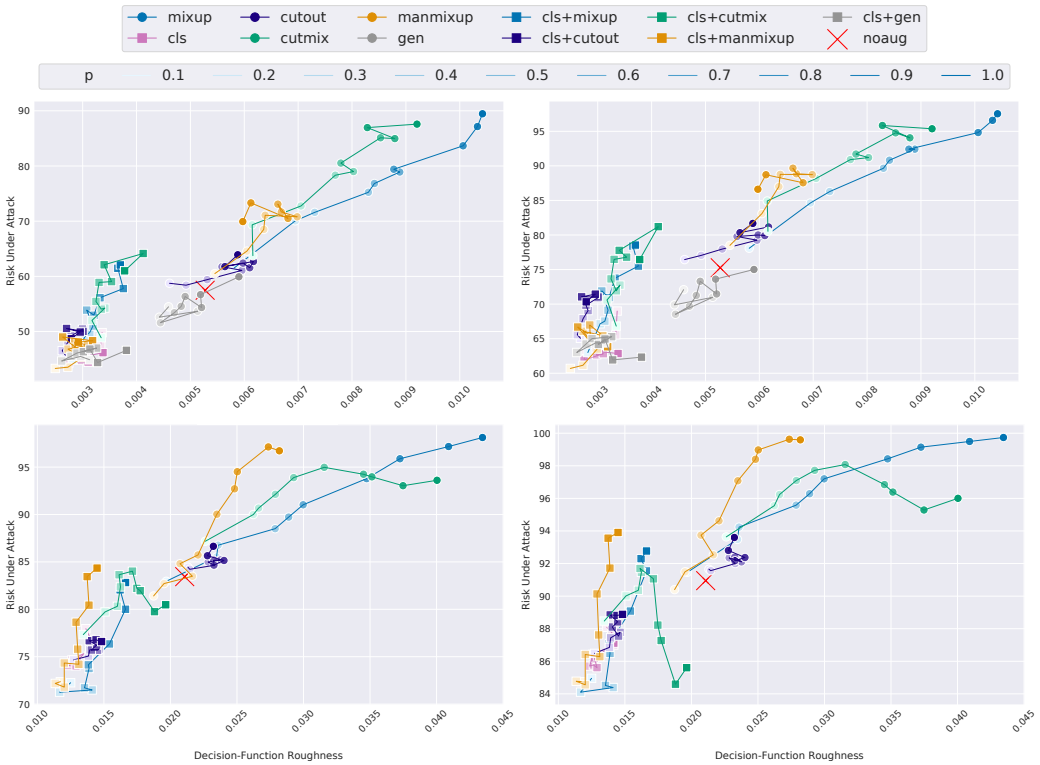

Figure 2: Decision-function roughness vs. Robustness. First column: roughness under L_2 perturbation model, $\text{PGD}_\epsilon = 0.1$. Second column: roughness under L_∞ perturbation model, $\text{PGD}_\epsilon = \frac{1}{255}$. First row: CIFAR10. Second row: CIFAR100.

augmentations exhibit different behavior. For all setups, decision-function roughness has a high correlation with the vulnerability for the methods. This suggests that methods with high decision-function roughness are vulnerable to adversarial attacks. Our result is consistent with related works [37, 38, 39, 40] on the relation between the shape of decision surface and adversarial vulnerability, suggesting that "rough" decision boundaries are key factors in adversarial vulnerability. In other words, we find that DAs that cause vulnerability also induce rough decision function surfaces. Furthermore, for

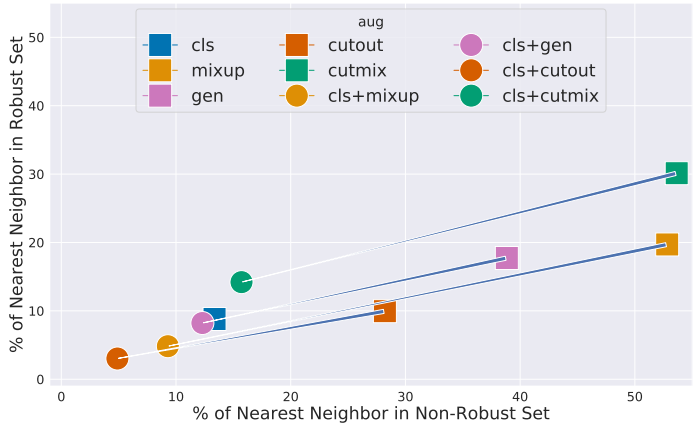

Figure 3: Results for data-augmentation spuriousness as a % of nearest neighbor of augmented data to robust and non-robust features on CIFAR10.

DAs such as classic, which have shown increased adversarial robustness, the percentage of augmentation does not significantly affect decision-function roughness. In contrast, DAs such as MixUp which resulted in reduced robustness are significantly affected by the augmentation percentage. Hence, the augmentations that cause rough decision functions are more strongly affected by changes in augmentation probability.

4.2.3. Data-Augmentation Spuriousness Results

Finally, we measure spuriousness induced by the data augmentation techniques. We plot these percentages in Figure 3. The most adversarially-vulnerable DAs (MixUp (yellow), CutMix (green)) are the closest to non-robust features. Furthermore, Gen (violet) and CutOut (orange), which were slightly less vulnerable to adversarial attacks, are relatively further away

from non-robust features. Finally, the most robust DA, classic (blue), is the furthest from the non-robust features. Additionally, when combined with classic, the augmented data increase their distance to non-robust features, also reflecting the previously-discussed robust performance.

Using the proposed DA spuriousness, we find that DAs resulting in robust models create samples that are distant to spurious features. We observe that most augmentations are relatively distant to robust features, in particular in comparison to their distance with the non-robust features. This result was expected, as the studied DAs do not incorporate adversarial directions in creating augmented samples, which in fact was leveraged in the generation process of 'robust features'.

5. Conclusion and Future Work

Recently-proposed heuristic and data-driven DA methods including MixUp [11], CutMix [12], ManifoldMixUp [13], CutOut [14], and Diffusion Models [15] have been claimed to improve both generalization and adversarial robustness. However, they have been tested only in combination with classical augmentations (like image flipping and rotation), and using a fixed fraction of augmented samples. This questions whether the claimed improvements are really induced by the newly-proposed DA strategies themselves, or they are instead induced mostly by classical augmentations and specific choices of the augmentation probability. In this work, we shed light on this issue by proposing an evaluation framework that helps decoupling the impact of such factors

on both accuracy and robustness, through the definition of different metrics that characterize the augmented manifold. We re-evaluate recently-proposed heuristic and data-driven DAs using our framework and find contradictory evidence when compared to prior work. In particular, our extensive empirical analysis on the aforementioned DA methods has shown that: (i) they only improve adversarial robustness when combined with classical augmentations, and even worsen it if used in isolation; and (ii) they are significantly affected by the choice of the augmentation probability. This demands future work to rethink not only the evaluation but also the development of novel DA methods for adversarial robustness, and we firmly believe that our work provides a significant first step in this direction.

One limitation to consider in the analysis of DA and robustness techniques is that they have primarily been developed and tested in the context of images. While this research has provided valuable insights into the effectiveness of these methods in improving model performance and robustness, it is important to acknowledge that the applicability of these techniques to other domains may vary. Therefore, further research is necessary to adapt these techniques to other domains and determine their effectiveness in improving model performance and robustness. While it is important to acknowledge the limitations of the current research, it is worth noting that our analysis and guidelines can still be partly applied to other domains. By understanding the underlying principles of evaluating DA for robustness, researchers can develop tailored approaches that account for the unique characteristics of each

domain. As such, our work provides a valuable foundation for future studies, and it can help avoid common pitfalls when evaluating model performance and robustness even in other application contexts.

Acknowledgements

This work was partially supported by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG in the COMET Module S3AI; by Fondazione di Sardegna under the project "TrustML: Towards Machine Learning that Humans Can Trust", CUP: F73C22001320007; and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NGEU.

References

[1] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, A. Madry, Adversarially robust generalization requires more data, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, 2018, pp. 5019–5031.

[2] P. Nakkiran, B. Neyshabur, H. Sedghi, The deep bootstrap framework: Good online learners are good offline generalizers, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[3] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, T. Mann, Fixing data augmentation to improve adversarial robustness, arXiv:2103.01946 (2021).

[4] C. M. Bishop, Pattern recognition and machine learning, 5th ed. (2007).

[5] T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, C. Re, A kernel theory of modern data augmentation, in: K. Chaudhuri, R. Salakhutdinov (Eds.), 36th Int'l Conf. on Machine Learning, Vol. 97, 2019, pp. 1528–1537.

[6] S. Chen, E. Dobriban, J. H. Lee, A group-theoretic framework for data augmentation, Journal of Machine Learning Res. 21 (245) (2020) 1–71.

[7] N. McLaughlin, J. M. Del Rincon, P. Miller, Data-augmentation for reducing dataset bias in person re-identification, in: 2015 12th IEEE International conference on advanced video and signal based surveillance (AVSS), IEEE, 2015, pp. 1–6.

[8] D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 3642–3649.

[9] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[10] S. Sharma, Y. Zhang, J. M. Rios Aliaga, D. Bouneffouf, V. Muthusamy, K. R. Varshney, Data augmentation for discrimination prevention and

bias disambiguation, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 358–364.

[11] H. Zhang, M. Cissé, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.

[12] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6023–6032.

[13] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold mixup: Better representations by interpolating hidden states, in: K. Chaudhuri, R. Salakhutdinov (Eds.), 36th Int'l Conf. on Machine Learning, Vol. 97 of Proc. Machine Learning Research, PMLR, Long Beach, California, USA, 2019, pp. 6438–6447.

[14] T. DeVries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017).

[15] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: NeurIPS 2020, December 6-12, virtual, 2020.

[16] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, Adversarial classification, in: KDD, 2004, pp. 99–108.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014.

[18] H. Guo, Y. Mao, R. Zhang, Mixup as locally linear out-of-manifold regularization, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3714–3722.

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

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[21] J. Schluter, T. Grill, Exploring data augmentation for improved singing voice detection with neural networks., in: ISMIR, 2015, pp. 121–126.

[22] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: A simple data augmentation method for automatic speech recognition, in: G. Kubin, Z. Kacic (Eds.), Interspeech 2019, 20th Annual Conf. of the International Speech Communication Association, Graz, Austria, 15-19 September, ISCA, 2019, pp. 2613–2617.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[24] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[25] A. Antoniou, A. Storkey, H. Edwards, Data augmentation generative adversarial networks, arXiv preprint arXiv:1711.04340 (2017).

[26] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernandez, J. Wardlaw, D. Rueckert, Gan augmentation: Augmenting training data using generative adversarial networks, arXiv preprint arXiv:1810.10863 (2018).

[27] R. Child, Very deep vaes generalize autoregressive models and can outperform them on images, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[28] G. W. Benton, M. Finzi, P. Izmailov, A. G. Wilson, Learning invariances in neural networks from training data, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33, NeurIPS 2020, December 6-12, 2020.

[29] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q. V. Le, Autoaugment: Learning augmentation policies from data, arXiv preprint arXiv:1805.09501 (2018).

[30] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Augmix: A simple data processing method to improve robustness and uncertainty, in: 8th Int'l Conf. on Learning Representations, ICLR 2020.

[31] F. Cucker, S. Smale, On the mathematical foundations of learning, American Mathematical Society 39 (1) (2002) 1–49.

[32] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[33] Y. Mansour, M. Schain, Robust domain adaptation, Annals of Mathematics and Artificial Intelligence 71 (4) (2014) 365–380.

[34] I. Attias, A. Kontorovich, Y. Mansour, Improved generalization bounds for robust learning, in: A. Garivier, S. Kale (Eds.), Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA, Vol. 98 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 162–183.

[35] D. Arpit, S. Jastrzkebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, et al., A closer look at memorization in deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 233–242.

[36] P. Jin, L. Lu, Y. Tang, G. E. Karniadakis, Quantifying the generalization error in deep learning in terms of data distribution and neural network smoothness, Neural Networks 130 (2020) 85–99.

[37] C. Lyu, K. Huang, H.-N. Liang, A unified gradient regularization family for adversarial examples, in: 2015 IEEE international conference on data mining, IEEE, 2015, pp. 301–309.

[38] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, N. Usunier, Parseval networks: Improving robustness to adversarial examples, in: International Conference on Machine Learning, PMLR, 2017, pp. 854–863.

[39] J. Sokolic, R. Giryes, G. Sapiro, M. R. Rodrigues, Robust large margin deep neural networks, IEEE Transactions on Signal Processing 65 (16) (2017) 4265–4280.

[40] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via randomized smoothing, in: Int'l Conf. on Machine Learning, PMLR, 2019, pp. 1310–1320.

[41] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: 6th Int'l Conf. on Learning Repr., ICLR 2018, April 30 - May 3, 2018.

[42] H. Shu, H. Zhu, Sensitivity analysis of deep neural networks, in: AAAI, 2019.

[43] M. Forouzesh, F. Salehi, P. Thiran, Generalization comparison of deep neural networks via output sensitivity, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 7411–7418.

[44] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, J. Sohl-Dickstein, Sensitivity and generalization in neural networks: an empirical study, in: ICLR, 2018.

[45] D. Donoho, High-dimensional data analysis: The curses and blessings of dimensionality, AMS Math Challenges Lecture (2000) 1–32.

[46] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, in: Advances in Neural Information Processing Systems, 2019, pp. 125–136.

[47] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, Robustbench: a standardized adversarial robustness benchmark, in: 35th Conf. on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[48] C. Han, K. Murao, T. Noguchi, Y. Kawata, F. Uchiyama, L. Rundo, H. Nakayama, S. Satoh, Learning more with less: Conditional pggan-based data augmentation for brain metastases detection using highly-rough annotation on mr images, in: 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 119–127.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

Appendix A. Experimental Setup

Appendix A.1. Training Setup

Our experiments are carried out using a ResNet18 [20] as a baseline. The ResNet model was trained using SGD with a momentum of 0.9, and weight-decay with penalty coefficient of $5e-4$, with batch size of 128. Each classifier was trained for 200 epochs with initial learning rate of 0.1, which was reduced twice by a factor of 10 every 80 epochs.

Appendix A.1.1. Data Augmentations

We list here the parameters used for the data augmentations.

MixUp: the mixing parameter λ was drawn from $\mathcal{B}(1,1)$.

Manifold-MixUp: the mixing parameter λ was drawn from $\mathcal{B}(2,2)$. The eligible layers on CIFAR10 was set to $\mathcal{S} = \{0,1,2\}$, while on CIFAR100 $\mathcal{S} = \{0,1,2,3\}$.

CutMix: the mixing parameter λ was drawn from $\mathcal{B}(1,1)$. Bounding boxes have been randomly chosen for the cutout operation, with cut ratio of $\sqrt{1-\lambda}$.

CutOut: with 1 hole and the length of 16 pixel has been used.

Classic: is a random combination of Random Cropping, Horizontal Flipping, Colour jittering, and Random Rotation. Random Cropping is done with the padding of 4 and size of 32. For Colour jittering, brightness, contrast, and saturation factors have been changed by a random amount chosen uniformly from $[0.75, 1.25]$.

Gen.: was utilized using samples generated by a Denoising Diffusion Probabilistic model [15] trained on CIFAR10. These samples have been released as CIFAR5M dataset [2][5]. Due to the computational complexity of DDPM, and lack of availability of such generated dataset on CIFAR100, we opted to only use Gen. on CIFAR10. When a sample is chosen to be augmented by Gen., we replace the original sample by a randomly chosen example of the same class from CIFAR5M.

Combinations with Classic: In all combination experiments, in addition to the target augmentation (e.g, MixUp), samples have been additionally augmented with Classic, with augmentation probability of 0.5.

Appendix A.1.2. Adversarial Attack Setup

We carry out 4 different untargeted PGD attacks with L_2 norm and 4 different untargeted PGD attacks with L_∞ norm. For PGD attacks with L_2 norm, we use perturbation sizes of $\text{PGD}_\epsilon \in \{0.01, 0.1, 0.5, 1\}$, and for PGD with L_∞ norm we use $\text{PGD}_\epsilon \in \{\frac{1}{255}, \frac{2}{255}, \frac{4}{255}, \frac{8}{255}\}$. All attacks have been conducted with a step size set to $\frac{1}{5}$ of the perturbation size ($\text{PGD}_\alpha = \frac{\text{PGD}_\epsilon}{5}$), and with 100 iterations.

Appendix A.2. Implementation

All experiments have been implemented in python using `PyTorch`. The adversarial attacks are done using the `robustness`[6] library.

[5]This dataset is publicly available here: `https://github.com/preetum/cifar5m`

[6]`https://github.com/MadryLab/robustness`

Appendix B. Extended Results

Appendix B.1. Robustness and Roughness with Other Perturbation Sizes

We show in Figures B.4, B.5, B.6 and B.7 the results for our analysis with increasing perturbation sizes and different perturbation models (L_∞ and L_2). We also summarize the results in Table B.3.

Table B.3: Extended summary of the results. Stress: Prediction-change stress (high indicates adversarial vulnerability). RUA: Risk under attack (high indicates adversarial vulnerability). Dist. to NRS: Distance to non.robust set (low indicates adversarial vulnerability). Imp. (Pr): Impact of augmentation probability on robustness. Perf.: Performance, inverse of misclassification risk. low:↓↓. high:↑↑. medium:↑↓. Man.: Manifold-MixUp. Gen.: Generative model. Cls: Classic.

	Roughness	**RUA**	**Dist. to NRS**	**Imp. (Pr)**	**Perf.**
Cls	↓↓	↓↓	high	low	↑↑
MixUp	↑↑	↑↑	low	high	↓↓
CutMix	↑↑	↑↑	low	high	↑↓
CutOut	↑↓	↑↑	med	low	↑↓
Gen.	↑↓	↓↓	med	low	↓↓
Man.	↑↓	↑↓	N/A	high	↑↓

Appendix B.2. Impact of Different Augmentation Probabilities

We provide a summary for the influence of augmentation probability on robustness and performance for different augmentations in Table B.4.

Appendix B.3. Distance to Robust and Non-robust Features

We provide in Table B.5 the detailed number of nearest neighbors to the different data manifolds.

Table B.4: Extended summary of the influence of augmentation probability on robustness and performance. Comb. w/ Cls: combined with classic. Man.: Manifold-MixUp. Cls.: Classic. Gen.: generative model.

	Performance	Robustness	
MixUp	high	high	Single
Man.	low	high	
CutMix	high	high	
CutOut	low	low	
Gen.	high	low	
Cls	low	low	
MixUp+Cls	low	high	Comb. w/ Cls
Man.+Cls	low	high	
CutMix+Cls	low	high	
CutOut+Cls	low	low	
Gen.+Cls	high	low	

Table B.5: The percentage of nearest neighbors of the augmented data on CIFAR10, in each set.

	Self	**Robust Features**	**Non-rob. Features**
Cls	77.83	8.81	13.36
MixUp	27.47	19.71	52.82
Gen	43.43	17.73	38.84
CutOut	61.84	9.95	28.21
CutMix	16.22	30.13	53.65
Cls+MixUp	85.88	4.83	9.29
Cls+Gen	79.48	8.22	12.30
Cls+CutOut	92.08	3.03	4.89
Cls+CutMix	70.09	14.20	15.71

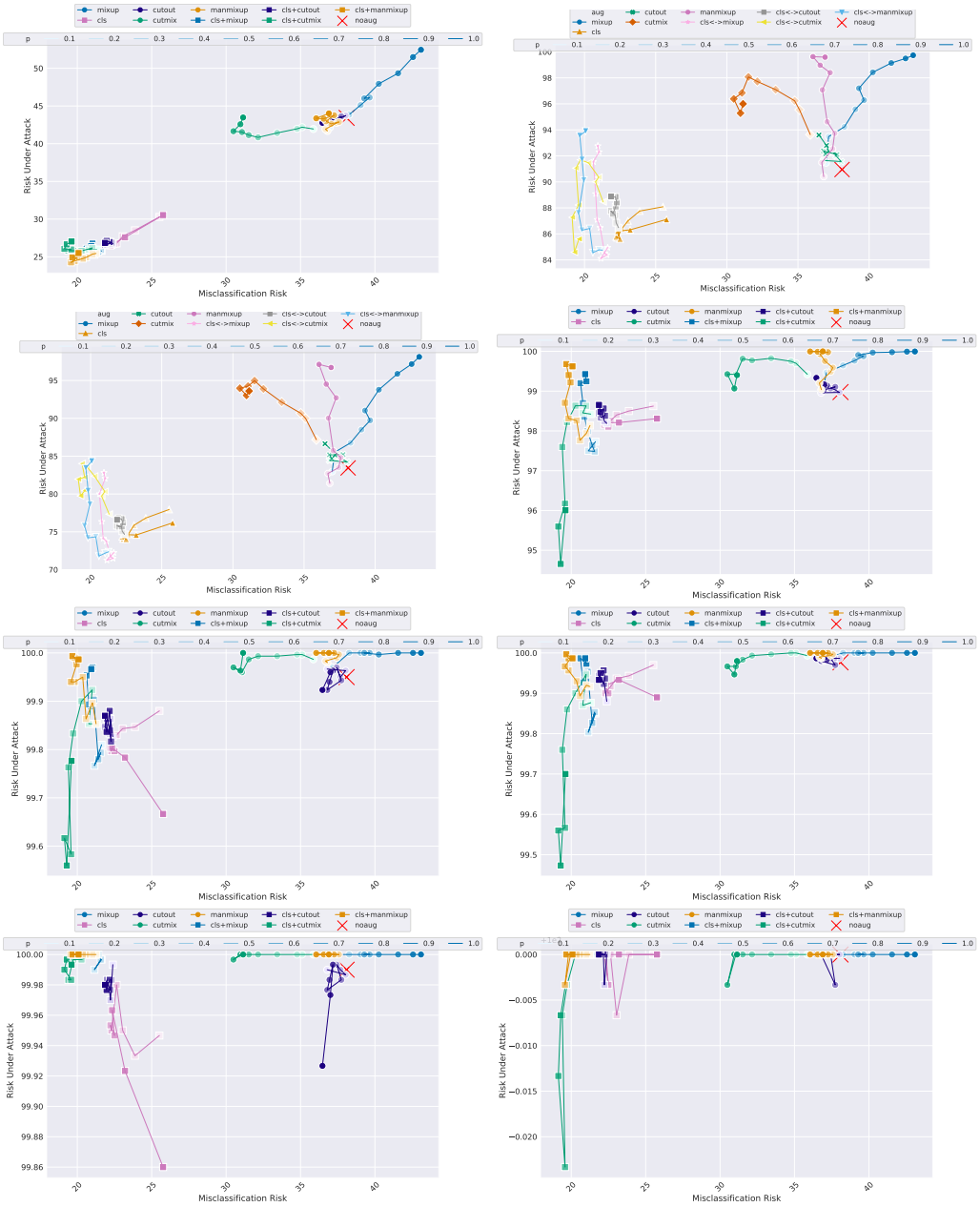

Figure B.4: Robustness vs. Performance on CIFAR100. first column: risk under attack (L_2) VS. misclassification risk second column: risk under attack (L_∞) VS. misclassification risk Perturbation sizes for L_2 PGD attacks are shown in each row: $\{0.01, 0.1, 0.5, 1\}$. Perturbation sizes for L_∞ PGD attacks are shown in each row: $\{\frac{1}{255}, \frac{2}{255}, \frac{4}{255}, \frac{8}{255}\}$.

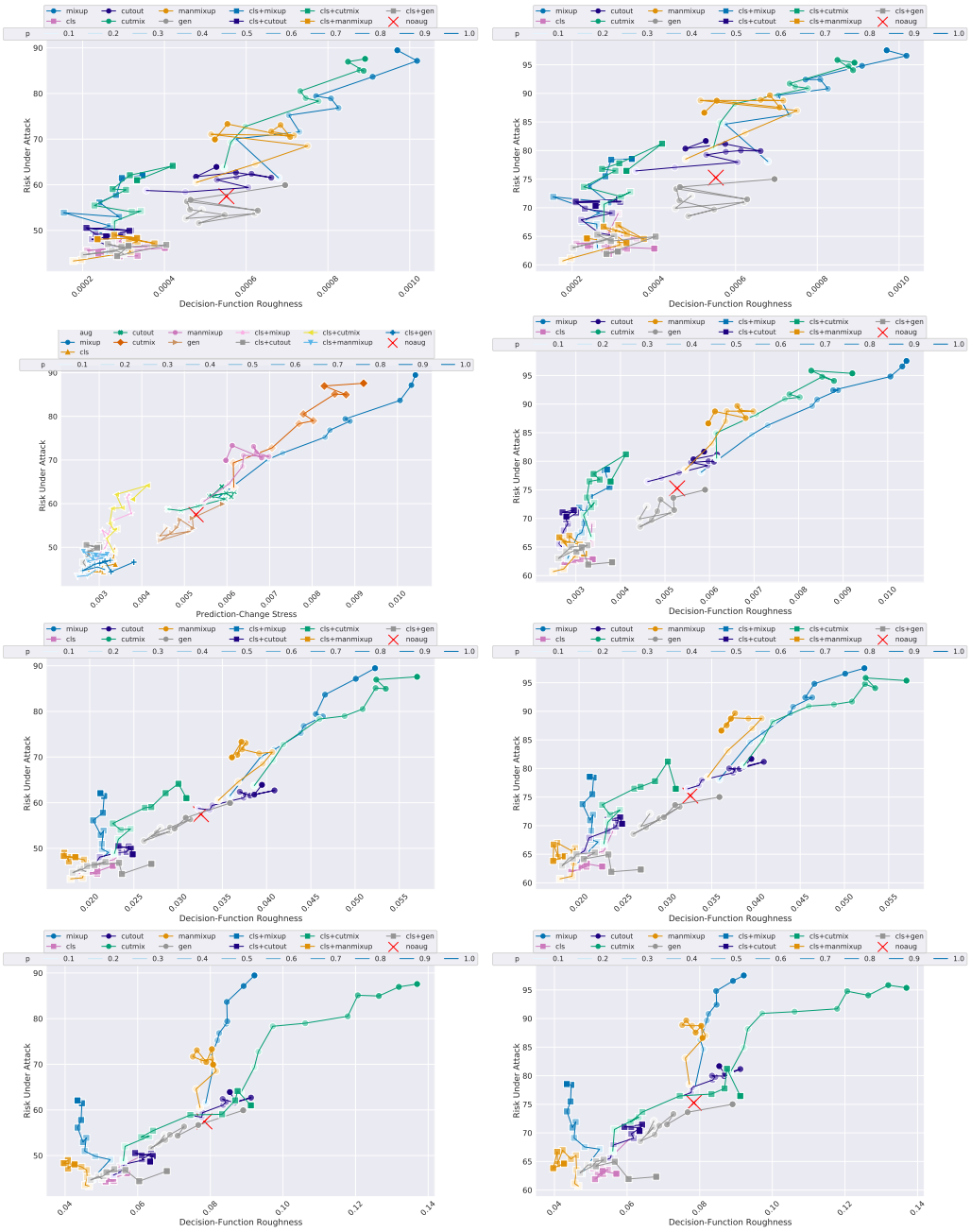

Figure B.5: Decision-function roughness vs. robustness on CIFAR10. first column: roughness vs. risk under attack (L_2) second column: roughness vs. risk under attack (L_∞) rows: $\epsilon_{stress} = \{0.01, 0.1, 0.5, 1, 2\}$. PGD attacks with L_2 and L_∞ used perturbation size of 0.1, and $\frac{1}{255}$, respectively.

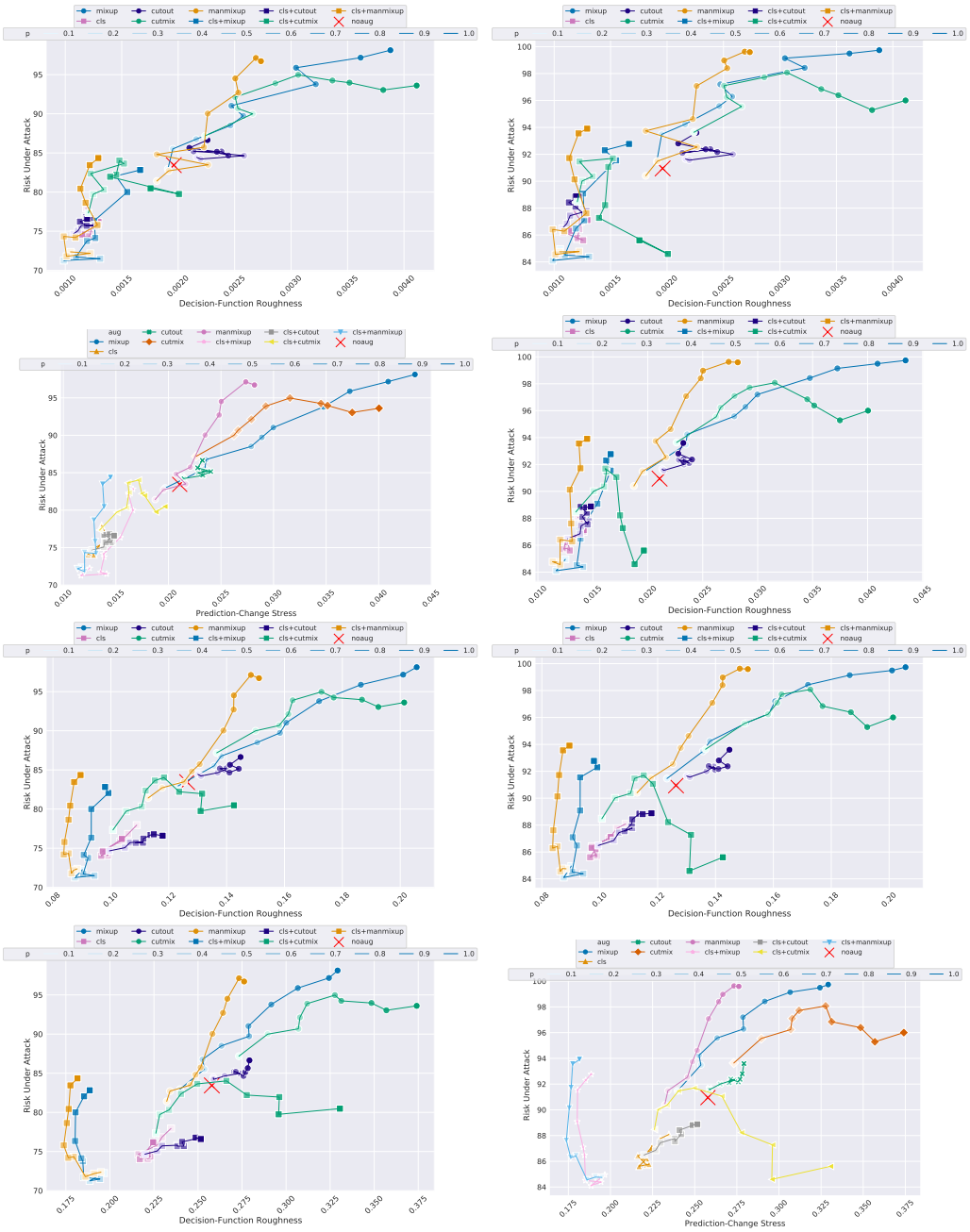

Figure B.6: Decision-function roughness vs. robustness on CIFAR100. first column: roughness vs. risk under attack (L_2) second column: roughness vs. risk under attack (L_∞) rows: $\epsilon_{stress} = \{0.01, 0.1, 0.5, 1, 2\}$. PGD attacks with L_2 and L_∞ used perturbation size of 0.1, and $\frac{1}{255}$, respectively.

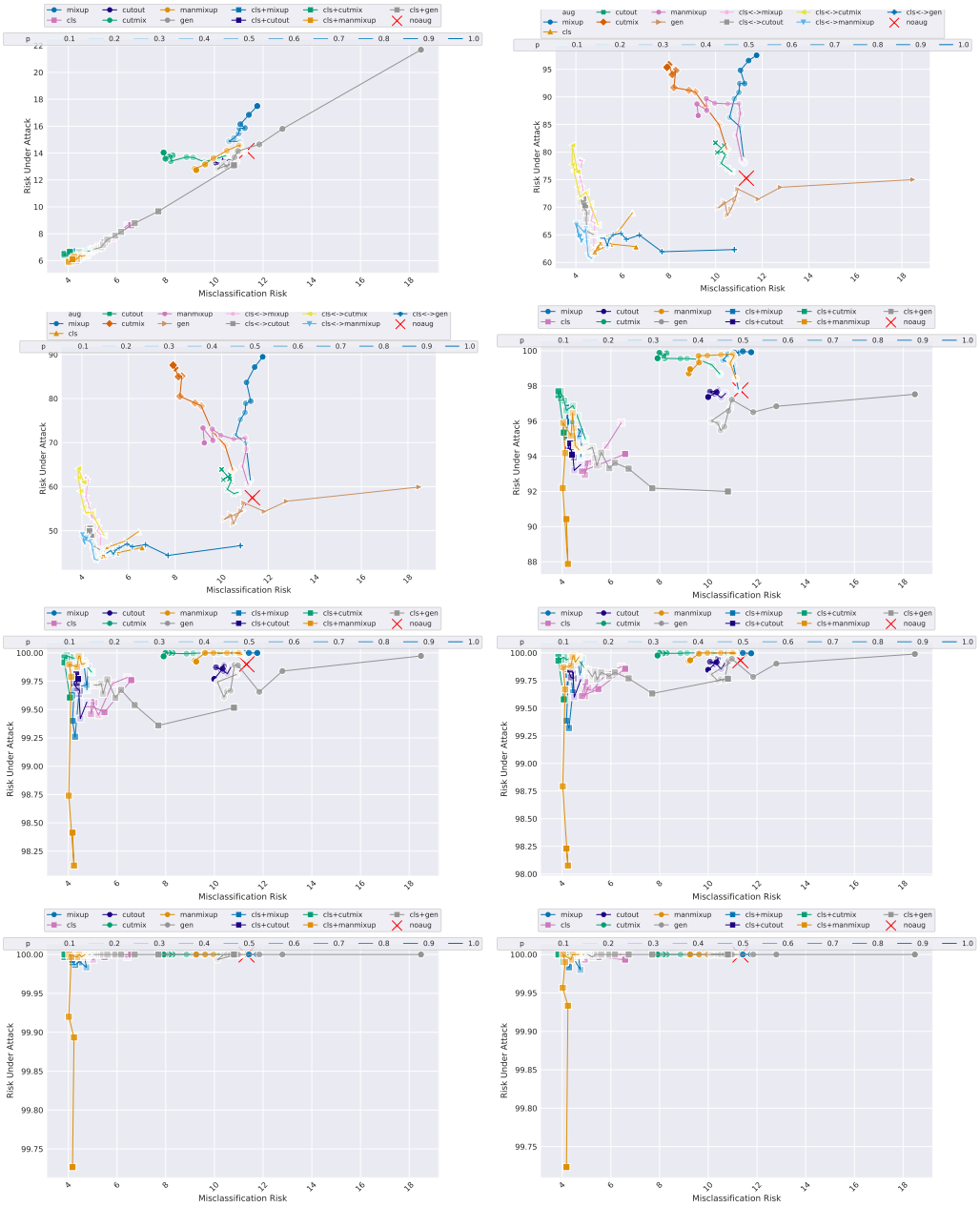

Figure B.7: Robustness vs. Performance on CIFAR10. first column: risk under attack (L_2) VS. misclassification risk second column: risk under attack (L_∞) VS. misclassification risk Perturbation sizes for L_2 PGD attacks are shown in each row: $\{0.01, 0.1, 0.5, 1\}$. Perturbation sizes for L_∞ PGD attacks are shown in each row: $\{\frac{1}{255}, \frac{2}{255}, \frac{4}{255}, \frac{8}{255}\}$.