PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures

Anonymous Author(s)
Affiliation
Address
email

Abstract

In real-world applications of machine learning, robust systems must consider measures of performance beyond standard test accuracy. These include out-of-distribution (OOD) robustness, prediction consistency, resilience to adversaries, calibrated uncertainty estimates, and the ability to detect anomalous inputs. However, optimizing for some of these measures often sacrifices performance on others. For instance, adversarial training only improves adversarial robustness and degrades classifier performance. Similarly, strong data augmentation and regularization techniques often improve OOD robustness but harm anomaly detection, raising the question of whether a Pareto improvement is possible. We identify a weakness of existing data augmentation techniques—namely, while they inject additional entropy into the training set, the entropy does not contain new structural complexity. This leads us to design a new data augmentation strategy utilizing the natural structural complexity of fractals, which outperforms numerous baselines and is the first method to comprehensively improve safety measures.

1 Introduction

A central challenge in machine learning is building models that are robust and reliable in the real world. In addition to performing well on the training distribution, deployed models should be robust to distribution shifts, consistent in their predictions, resilient to adversaries, calibrated in their uncertainty estimates, and capable of identifying anomalous inputs. Numerous prior works have tackled each of these problems separately (Hendrycks and Dietterich, 2019; Madry et al., 2018; Guo et al., 2017; Emmott et al., 2015), but they can also be grouped together as various aspects of safety engineering for machine learning (Hendrycks et al., 2021b). Consequently, the properties listed above can be thought of as safety measures.

Ideally, models deployed in real-world settings would perform well on multiple safety measures. Unfortunately, prior work has shown that optimizing for some desirable properties often comes at the cost of others. For example, adversarial training only improves adversarial robustness and degrades classification performance (Tsipras et al., 2018). Similarly, inducing consistent predictions on out-of-distribution (OOD) inputs seems to be fundamentally at odds with better detecting these inputs, an intuition supported by recent work (Chun et al., 2019) showing that while various data augmentation and regularization techniques can improve OOD robustness, they often harm anomaly detection. This raises the question of whether a Pareto improvement on all safety measures is possible.

We show that Pareto improvements are possible with PixMix, a simple and effective data processing method that leverages the structural complexity of fractals. We observe that while previous data processing methods add entropy to the training set, they do not introduce new sources of complexity. As opposed to entropy, which is maximized by the uninformative uniform distribution, complexity is often described in terms of the difficulty of creation or degree of organization (Lloyd, 2001). A classic

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

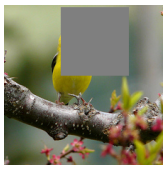
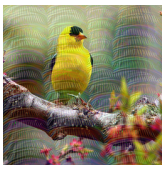
Method	Baseline	Cutout	Mixup	CutMix	PixMix
					
Corruptions mCE (↓)	50.0 +0.0	51.5 +1.5	48.0 −2.0	51.5 +1.5	**30.5** **−19.5**
Adversaries Error (↓)	96.5 +0.0	98.5 +1.0	97.4 +0.9	97.0 +0.5	**92.9** **−3.9**
Consistency mFR (↓)	10.7 +0.0	11.9 +1.2	9.5 −1.2	12.0 +1.3	**5.7** **−5.0**
Calibration RMS Error (↓)	31.2 +0.0	31.1 −0.1	13.0 −18.1	29.3 −1.8	**8.1** **−23.0**
Anomaly Detection AUROC (↑)	77.7 +0.0	74.3 −3.4	71.7 −6.0	74.4 −3.3	**89.3** **+11.6**

Table 1: PixMix outperforms strong baselines on various robustness and uncertainty estimation tasks. PixMix introduces structural complexity into the training process using fractals and is the first method to provide a Pareto improvement for existing safety measures.

example of complexity is fractals, which have recently proven useful for pre-training image classifiers. We develop a new data processing method that incorporates fractals and complex dreamlike images, which we call PixMix. We find that that PixMix provides the first Pareto improvement on a broad range of existing safety measures, outperforming numerous strong baselines.

2 PixMix

We propose PixMix, a simple and effective data augmentation technique for improving various ML Safety measures. PixMix is comprised of two main components: a set of structurally complex images ("Pix") and a pipeline for augmenting clean training images ("Mix"). On a high level, PixMix method integrates diverse patterns from fractals and feature visualizations into the training set. Unlike prior data augmentation techniques, this introduces not only additional entropy but also additional complexity, where complexity refers to difficulty of description or creation, or degree of organizational structure (Lloyd, 2001).

2.1 Mixing Set

Feature Visualization. Optimizing images to maximize the average response of channels in convolutional networks yields distinctive patterns with high spatial complexity Mordvintsev et al. (2015); Olah et al. (2017). Thus, we include feature visualizations in our mixing set. We collect feature visualizations from the initial layers of several convolutional architectures using OpenAI Microscope.

Fractals. Fractals can be generated in several ways, with one of the most common being iterated function systems. Rather than generate our own diverse fractals, which is a substantial research endeavor (Kataoka et al., 2020), we download fractals from manually curated collections on DeviantArt. The resulting fractals are visually diverse, which can be seen in the bottom half of Figure 1.

2.2 Mixing Pipeline

The pipeline for augmenting clean training images is described in Figure 2. First, a clean image has a 50% chance of having a randomly selected standard augmentation applied. Next, we augment the image a random number of times by either additively or multiplicatively mixing it with a freshly augmented clean image or a structurally complex image from the mixing set. Multiplicative mixing

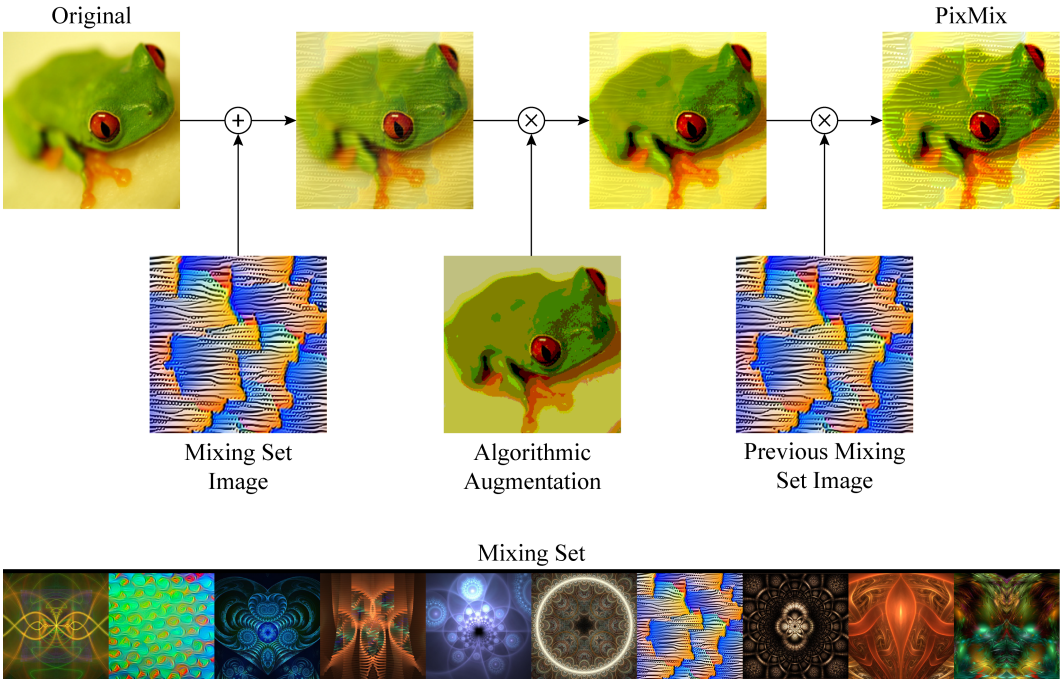


Figure 1: Top: An instance of a PixMix augmentation being applied to a frog image. The original clean image is repeatedly mixed with augmented versions of itself and sources of structural complexity, including feature visualizations and fractals. Bottom: Sample images from the PixMix mixing set.

is performed similarly to the geometric mean. While Mixup uses convex combinations, we found that conic combinations perform better. Thus, additive and multiplicative mixing are performed with exponents and weights sampled from a beta distribution. An instance of our mixing pipeline is shown in the top half of Figure 1.

3 Experiments

Baselines. We compare PixMix to various state-of-the-art data augmentation baselines. *Baseline* augments CIFAR-10 and CIFAR-100 images using the standard practice of random horizontal flips and random crops with four pixels of zero padding. *Cutout* randomly masks image patches. *Mixup* linearly interpolates input images and labels. *CutMix* replaces image patches with patches from other images in the training set. *Auto Augment* searches for augmentation strategies that maximize accuracy on a validation set. *Outlier Exposure* trains networks to be uncertain on a training dataset of outliers, distinct from the out-of-distribution test sets.

Tasks and Metrics. We compare PixMix to baseline methods on five distinct ML Safety tasks. Individual methods are trained on clean versions of CIFAR-10 and CIFAR-100, then evaluated on each of the following tasks:

- *Corruptions*: Classifying corrupted images from the CIFAR-10-C and CIFAR-100-C datasets. The metric is the mean corruption error (mCE) across all fifteen corruptions and five severities for each corruption. Lower is better.
- *Consistency*: Classifying sequences of perturbed images from the CIFAR analogues of ImageNet-P without flipping predictions. The metric is the mean flip rate (mFR). Lower is better.
- *Adversarial*: Classifying images that have been adversarially perturbed by projected gradient descent (Madry et al., 2018). We use an ℓ_∞ budget of $2/255$ with 20 steps of optimization. The metric is the classifier error rate. Lower is better.
- *Calibration*: Classifying images with calibrated prediction probabilities, i.e. matching the empirical frequency of correctness. The metric is RMS calibration error (Hendrycks et al., 2019c). Lower is better.

```
def pixmix(x_orig, x_mix_pic, k=3, beta=3):
    x_pixmix = random.choice([augment(x_orig), x_orig])

    for i in range(random.choice([0,1,...,k])):

        mix_image = random.choice([augment(x_orig), x_mix_pic])
        mix_op = random.choice([additive, multiplicative])

        x_pixmix = mix_op(x_pixmix, mix_image, beta)

    return x_pixmix

def augment(x):
    aug_op = random.choice([rotate, ..., posterize])
    return aug_op(x)
```

Figure 2: Simplified code for PixMix, our proposed data augmentation method. PixMix integrates new complexity into the training process by leveraging fractals and feature visualizations, resulting in improved classifier robustness and uncertainty estimation across numerous safety measures.

- *Anomaly Detection*: Detecting out-of-distribution images from various unseen distributions. The anomaly distributions are Gaussian, Rademacher, Blobs, Textures, SVHN, ImageNet, Places69. We describe each in the Appendix and report average AUROC. Higher is better.

Results. In Table 1, we see that PixMix improves over the standard Baseline augmentation on all safety measures and is the first method to provide a Pareto improvement. Results for all baselines are in Table 2. PixMix obtains better performance than all baselines on Corruptions, Consistency, Adversarial, and Calibration. Additionally, PixMix nearly matches the anomaly detection performance of Outlier Exposure, a state-of-the-art anomaly detection method that requires large quantities of diverse, known outliers. On CIFAR-100, performance improvements are especially large, with mCE on the Corruptions task dropping by 4.9% compared to AugMix and 19.5% compared to Baseline augmentations. Additional results and ablations are in the Appendix.

		Baseline	Cutout	Mixup	CutMix	Auto Augment	AugMix	Outlier Exposure	PixMix
CIFAR-10	Corruptions	26.4	25.9	21.0	26.5	22.2	12.4	25.1	**9.5**
	Consistency	3.4	3.7	2.9	3.5	3.6	1.7	3.4	**1.7**
	Adversarial	91.3	96.0	93.3	92.1	95.1	86.8	92.9	**82.1**
	Calibration	22.7	17.8	12.1	18.6	14.8	9.4	13.0	**3.7**
	Anomaly Detection	91.9	91.4	88.2	92.0	93.2	89.2	**98.4**	97.0
CIFAR-100	Corruptions	50.0	51.5	48.0	51.5	47.0	35.4	51.5	**30.5**
	Consistency	10.7	11.9	9.5	12.0	11.2	6.5	11.3	**5.7**
	Adversarial	96.8	98.5	97.4	97.0	98.1	95.6	97.2	**92.9**
	Calibration	31.2	31.1	13.0	29.3	24.9	18.8	15.2	**8.1**
	Anomaly Detection	77.7	74.3	71.7	74.4	80.4	84.9	**90.3**	89.3

Table 2: Comparing PixMix to baselines on robustness and uncertainty estimation tasks. Full results are in appendix. Lower is better except for OOD detection. PixMix outperforms all baseline methods and is the first method to provide a Pareto improvement over the Baseline.

4 Conclusion

We proposed PixMix, a simple and effective data augmentation technique for improving ML safety measures. Unlike previous data augmentation techniques, PixMix introduces new complexity into the training procedure by leveraging fractals and feature visualizations. We evaluated PixMix on five distinct ML Safety tasks: corruption robustness, prediction consistency, adversarial robustness, confidence calibration, and anomaly detection. We found that PixMix was the first method to provide a Pareto improvement over the baseline, and it obtained state-of-the-art performance in nearly all settings.

References

M. Baradad, J. Wulff, T. Wang, P. Isola, and A. Torralba. Learning to see by looking at noise. *arXiv preprint arXiv:2106.05963*, 2021.

D. Bashkirova, D. Hendrycks, D. Kim, S. Mishra, K. Saenko, K. Saito, P. Teterwak, and B. Usman. Visda-2021 competition universal domain adaptation to improve performance on out-of-distribution data. *arXiv preprint arXiv:2107.11011*, 2021.

S. Chun, S. J. Oh, S. Yun, D. Han, J. Choe, and Y. Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. *Uncertainty and Robustness in Deep Learning. ICML Workshop*, 2019.

E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. AutoAugment: Learning augmentation policies from data. *CVPR*, 2018.

T. Devries and G. W. Taylor. Improved regularization of convolutional neural networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.

T. Devries and G. W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *ArXiv*, abs/1802.04865, 2018.

A. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158*, 2015.

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

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019a.

D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=HyxCxhRcY7.

D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. *ICLR*, 2019c.

D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019d.

D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021b.

H. Kataoka, K. Okayasu, A. Matsumoto, E. Yamagata, R. Yamada, N. Inoue, A. Nakamura, and Y. Satoh. Pre-training without natural images. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

K. Lee, H. Lee, K. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *ICLR*, 2018a.

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 2018b.

S. Lloyd. Measures of complexity: a nonexhaustive list. *IEEE Control Systems Magazine*, 21(4):7–8, 2001.

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

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

E. Mintun, A. Kirillov, and S. Xie. On interaction between augmentations and corruptions in natural corruption robustness. *arXiv preprint arXiv:2102.11273*, 2021.

A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks, 2015. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

K. Nakashima, H. Kataoka, A. Matsumoto, K. Iwata, and N. Inoue. Can vision transformers learn without natural images? *ArXiv*, abs/2103.13023, 2021.

K. Nguyen and B. O'Connor. Posterior calibration and exploratory analysis for natural language processing models. *arXiv preprint arXiv:1508.05154*, 2015.

C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *NeurIPS*, 2019.

L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

R. Takahashi, T. Matsubara, and K. Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9): 2917–2931, 2019.

Y. Tokozume, Y. Ushiku, and T. Harada. Between-class learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5486–5494, 2018.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *ICCV*, 2019.

H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2017.

Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

A Related Work

Robustness. Out-of-distribution robustness considers how to make ML models resistant to various forms of data shift at test time. Geirhos et al. (2019) uncover a texture bias in convolutional networks and show that training on diverse stylized images can improve robustness at test-time. Hendrycks and Dietterich (2019) introduce ImageNet-C, a benchmark of diverse image corruptions known to track robustness on real world data shifts (Hendrycks et al., 2021a). They also introduce ImageNet-P for measuring prediction consistency under various average-case input perturbations. The ImageNet-R benchmark measures performance degradation under more natural distribution shifts (Hendrycks et al., 2021a). In the similar setting of domain adaptation, Bashkirova et al. (2021) consider evaluating test-time robustness of models and even anomaly detection (Emmott et al., 2015; Ruff et al., 2021).

Uncertainty Estimation. Calibrated prediction confidences are valuable for classification models in real-world settings. Nguyen and O'Connor (2015); Guo et al. (2017) evaluate and improve the calibration of neural networks. Lakshminarayanan et al. (2017) show that ensembling can improve calibration without requiring a validation set. Hendrycks et al. (2019a) find that pre-training can improve calibration. Ovadia et al. (2019) find that calibration degrades under data shift.

Anomaly detection seeks to estimate whether an input is out-of-distribution (OOD) with respect to a given training set. Hendrycks and Gimpel (2017) propose a simple baseline for detecting classifier errors and OOD inputs. Devries and Taylor (2018) propose training classifiers with an additional confidence branch for detecting OOD inputs. Lee et al. (2018b) propose the Mahalanobis detector. Lee et al. (2018a) propose improving representations used for detectors with near-distribution images generated by GANs. Hendrycks et al. (2019b) fine-tune classifiers with diverse, natural anomalies and find that this greatly improves OOD detection.

Data Augmentation. Data augmentation is a powerful approach for improving generalization performance. For Vision Transformers, data augmentation can improve clean accuracy comparably to a 10x increase in model size (Steiner et al., 2021). Hendrycks et al. (2021a) find that data augmentation can also improve out-of-distribution robustness comparably to a 1000x increase in labeled data. Various augmentation techniques for image data have been proposed, including Cutout (Devries and Taylor, 2017; Zhong et al., 2017), Mixup (Zhang et al., 2017; Tokozume et al., 2018), CutMix (Yun et al., 2019; Takahashi et al., 2019), and AutoAugment (Cubuk et al., 2018). Chun et al. (2019) evaluates some of these techniques on CIFAR-10-C, a variant of ImageNet-C for the CIFAR-10 dataset (Hendrycks and Dietterich, 2019). They find that these data augmentation techniques can improve OOD generalization at the cost of weaker OOD detection. Hendrycks et al. (2019d) introduce AugMix, a data augmentation technique that specifically improves OOD generalization by a substantial margin. Similarly, Lopes et al. (2019) find that inserting random noise patches into training images improves robustness.

While existing data augmentation techniques improve some safety measures, we find that they typically harm performance on other safety measures. Additionally, they do not introduce new complexity into the training set. For example, adding random noise patches (Zhong et al., 2017; Lopes et al., 2019) increases the entropy of the training distribution, but entropy is not the same thing as complexity, and out-of-distribution images may contain significant amounts of novel structural complexity.

Training on Complex Synthetic Images. Fractals are a classic example of complexity, as they exhibit substantial structural organization. Kataoka et al. (2020) introduce FractalDB, a dataset of black-and-white fractals, and show that pre-training on fractal images can yield better downstream performance than pre-training on many manually annotated natural datasets. Nakashima et al. (2021) show that Vision Transformers trained on a large variant of FractalDB can match ImageNet-1K pre-training on downstream tasks. Baradad et al. (2021) find that other synthetic datasets may be more effective than FractalDB, finding that structural complexity and diversity are key properties for good downstream transfer. We depart from this line of work and ask whether structurally complex images can be used for data augmentation instead of training from scratch. We find that fractals and dreamlike images obtained from visualizing internal features of neural networks can be used to obtain a Pareto improvement on safety measures, including robustness and uncertainty estimation tasks.

B Additional Results

	Accuracy	Corruptions	Consistency	Adversarial	Calibration	Anomaly
	Clean	C	CIFAR-P	PGD	C	Detection
	Error (↓)	mCE (↓)	mFR (↓)	Error (↓)	RMS (↓)	AUROC (↑)
PixMix	20.3	30.5	5.7	92.9	8.1	89.3
Mix Input	19.9	34.1	6.4	96.7	15.5	86.5
Mix Aug	20.6	31.1	6.2	94.2	6.0	89.7
Iterative	21.1	31.4	5.6	90.6	12.7	86.7

Table 3: PixMix variations on CIFAR-100. Mix Input only mixes with augmented versions of the clean image. Mix Aug only mixes with images from the mixing set (i.e. fractals and feature visualizations). Iterative mixes with feature visualizations computed on the fly for the current network. Using the mixing set alone is more effective than augmented images alone, and combining them can further improve performance on several metrics.

	Accuracy	Corruptions	Consistency	Adversarial	Calibration	Anomaly
	Clean	C	CIFAR-P	PGD	C	Detection
	Error	mCE	mFR	Error	RMS	AUROC↑
Dead Leaves (Squares)	21.3	36.2	6.3	94.1	15.8	81.8
Spectrum + Color + WMM	20.7	36.1	6.6	94.4	15.9	85.8
StyleGAN (Oriented)	20.4	37.3	7.2	97.0	14.9	83.7
300K Random Images	19.6	34.5	6.3	94.7	12.9	86.2
FractalDB	20.3	33.9	6.4	98.2	12.0	82.5
Fractals	20.3	32.3	6.2	95.5	8.7	88.9
Feature Visualization (FVis)	21.5	30.3	5.4	91.5	9.9	88.1
Fractals + FVis	20.3	30.5	5.7	92.9	8.1	89.3

Table 4: PixMix variations on CIFAR-100 with different mixing sets. We compare Fractals + FVis, the mixing set used in PixMix, to numerous other datasets. Baradad et al. (2021) introduce the first three datasets. Hendrycks et al. (2019c) introduce 300K Random Images, and Kataoka et al. (2020) introduce FractalDB. We collect Fractals and FVis.

	Accuracy	Corruptions		Consistency		Adversarial	Calibration			Anomaly	
	Clean	C	$\overline{\text{C}}$	CIFAR-P		PGD	Clean	C	$\overline{\text{C}}$	Detection (↑)	
	Error	mCE	mCE	mFR	mT5D	Error	RMS	RMS	RMS	AUROC	AUPR
Baseline	21.3	50.0	52.0	10.7	2.7	96.8	14.6	31.2	30.9	77.7	35.4
Cutout	19.9	51.5	50.2	11.9	2.7	98.5	11.4	31.1	29.4	74.3	31.3
Mixup	21.1	48.0	49.8	9.5	3.0	97.4	10.5	13.0	12.9	71.7	31.9
CutMix	20.3	51.5	49.6	12.0	3.0	97.0	12.2	29.3	26.5	74.4	32.3
AA	**19.6**	47.0	46.8	11.2	2.6	98.1	9.9	24.9	22.8	80.4	33.2
AugMix	20.6	35.4	41.2	6.5	1.9	95.6	12.5	18.8	22.5	84.9	53.8
OE	21.9	50.3	52.1	11.3	3.0	97.0	12.0	13.8	13.9	**90.3**	66.2
PixMix	20.3	**30.5**	**36.7**	**5.7**	**1.6**	**92.9**	**7.0**	**8.1**	**8.9**	89.3	**70.9**

Table 5: Full results for CIFAR-100. $\overline{\text{C}}$ is an additional dataset of corrupted CIFAR-100 images (Mintun et al., 2021). mT5D is an additional metric used for gauging prediction consistency in ImageNet-P, which we adapt to CIFAR-100.

	Accuracy	Corruptions		Consistency		Adversarial	Calibration			Anomaly	
	Clean	C	$\overline{\text{C}}$	CIFAR-P		PGD	Clean	C	$\overline{\text{C}}$	Detection (↑)	
	Error	mCE	mCE	mFR	mT5D	Error	RMS	RMS	RMS	AUROC	AUPR
Baseline	4.4	26.4	26.4	3.4	1.7	91.3	6.4	22.7	22.4	91.9	70.9
Cutout	**3.6**	25.9	24.5	3.7	1.7	96.0	3.3	17.8	17.5	91.4	63.6
Mixup	4.2	21.0	22.1	2.9	2.1	93.3	12.5	12.1	10.9	88.2	67.1
CutMix	4.0	26.5	25.4	3.5	2.1	92.1	5.0	18.6	17.8	92.0	65.5
AA	3.9	22.2	24.4	3.6	1.7	95.1	4.0	14.8	16.6	93.2	64.6
AugMix	4.3	12.4	16.4	1.7	1.2	86.8	5.1	9.4	12.6	89.2	61.5
OE	4.6	25.1	26.1	3.4	1.9	92.9	6.9	13.0	13.2	**98.4**	**92.5**
PixMix	4.2	**9.5**	**13.6**	**1.7**	**1.0**	**82.1**	**2.6**	**3.7**	**5.3**	97.0	88.4

Table 6: Full results for CIFAR-10. $\overline{\text{C}}$ is an additional dataset of corrupted CIFAR-10 images (Mintun et al., 2021). mT5D is an additional metric used for gauging prediction consistency in ImageNet-P, which we adapt to CIFAR-10.