Supplementary Materials for PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures

Dan Hendrycks* UC Berkeley
Andy Zou* UC Berkeley
Mantas Mazeika UIUC
Leonard Tang Harvard University
Bo Li UIUC
Dawn Song UC Berkeley
Jacob Steinhardt UC Berkeley

A. Additional Results

Mixing Strategies. In Table 1, we analyze different mixing strategies. The full PIXMIX mixing strategy is depicted in Figures 2 and 3 of the main paper. Mix Input only includes clean images in the mixing pipeline and does not use the mixing set at all. This severely harms performance on all safety metrics. Mix Aug only mixes with images from the mixing set. This reduces RMS calibration error but increases error on robustness tasks compared to PIXMIX Original. Finally, Iterative mixes with feature visualizations computed on the fly for the network being trained. This performs well on robustness tasks but has weaker calibration and anomaly detection. Additionally, computing feature visualizations at each iteration of training is substantially slower than precomputing them on fixed networks as we do in PIXMIX.

Full Results. In Tables 3, 4, and 5, we report full results for CIFAR-10, CIFAR-100, and ImageNet. The ImageNet results are copied from the main paper. For CIFAR, we evaluate on additional datasets, including CIFAR-10-$\bar{\text{C}}$ and CIFAR-100-$\bar{\text{C}}$, additional datasets of corrupted CIFAR images. We also report the mT5D metric on ImageNet-P. In all cases, PIXMIX provides the best overall performance.

Noise-Based Augmentations. Since noise-based augmentations sometimes nearly overlap with the test distribution and thereby may have an unfair advantage, we separately compare to several additional baselines on ImageNet that use noise-based data augmentations. *ANT* trains networks on inputs with adversarially transformed noise applied [5]. *Speckle* trains on inputs with speckle noise added, which has been observed to improve robustness. *EDSR* and *Noise2Net* inject noise using image-to-image neural networks with noisy parameters [2]. *Adversarial* trains networks with ℓ_∞ perturbations of magnitude $\varepsilon = 8/255$ [4].

*Equal Contribution.

Results are in Tables 7. We find that ANT and Speckle have strong performance on ImageNet-P overall, but this mostly comes from the Gaussian and shot noise categories. If we only consider prediction stability on non-noise categories, PIXMIX exhibits the least volatility in predictions out of all the methods considered.

Hyperparameter Sensitivity. In Table 10, we examine the hyperparameter sensitivity of PIXMIX on corruption robustness for CIFAR-100. We vary the β and k hyperparameters and find that performance is very stable across a range of hyperparameters.

Places365 Anomaly Detection. In Table 9, we show anomaly detection performance with Places365 as the in-distribution data. For all methods, we use a ResNet-18 pre-trained on Places365. PIXMIX and Outlier Exposure (OE) are fine-tuned for 10 epochs. We find that PIXMIX nearly matches the state-of-the-art OE detector despite being a general data augmentation technique that improves many other safety metrics.

B. Outlier Datasets

For anomaly detection, we use a suite of out-of-distribution datasets and average metrics across all OOD datasets in the main results. Gaussian noise is IID noise sampled from a normal distribution. Rademacher Noise is noise with each pixel sampled from $\{-1, 1\}$ with equal probability. Blobs are algorithmically generated blobs. Textures are from the Describable Textures Dataset [1]. SVHN has images of numbers from houses. Places69 contains 69 scene categories and is disjoint from Places365.

C. Broader Impacts

As PIXMIX differentially improves safety metrics, it could have various beneficial effects. Improved robustness

	Accuracy	Corruptions	Consistency	Adversaries	Calibration	Anomaly
	Clean Error (↓)	C mCE (↓)	CIFAR-P mFR (↓)	PGD Error (↓)	C RMS (↓)	Detection AUROC (↑)
PIXMIX Original	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 1. 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		Adversaries	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 (↑)
CutMix	20.3	51.5	49.6	12.0	3.0	97.0	12.2	29.3	26.5	74.4	32.3
PIXMIX	20.3	**30.5**	36.7	**5.7**	**1.6**	**92.9**	7.0	8.1	8.9	89.3	**70.9**
PIXMIX + CutMix	**19.9**	30.9	**35.5**	5.8	1.7	93.1	**4.4**	**6.0**	**5.9**	**89.5**	68.6

Table 2. Combining PIXMIX and CutMix on CIFAR-100. While PIXMIX is strong on its own, combination with other data augmentation techniques can further improve performance.

can result in more reliable machine learning systems deployed in safety-critical situations [3], such as self-driving cars. Anomaly detection enables better human oversight of machine learning systems and fallback policies in cases where systems encounter inputs they were not designed to handle. At the same time, anomaly detection could be misused as a surveillance tool, requiring careful consideration of individual use cases. Calibration enables more meaningful predictions that increase trust with end users. Additionally, compared to other methods for improving robustness, PIXMIX requires minimal modification of the training setup and a low computational overhead, resulting in lower costs to machine learning practitioners and the environment.

References

[1] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Computer Vision and Pattern Recognition*, 2014. 1

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 1

[3] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021. 2

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[5] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pages 53–69. Springer, 2020. 1

	Accuracy	Corruptions		Consistency		Adversaries	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
AutoAugment	**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 3. Full results for CIFAR-100. mT5D is an additional metric used for gauging prediction consistency in ImageNet-P, which we adapt to CIFAR-100. Note PIXMIX can achieve 19.6% error rate if it uses 300K Random Images as the Mixing Set, so PIXMIX can achieve the same accuracy as AutoAugment yet also do better on safety metrics.

	Accuracy	Corruptions		Consistency		Adversaries	Calibration			Anomaly	
	Clean	CIFAR-C	$\overline{\text{C}}$	CIFAR-P		PGD	Clean	CIFAR-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
AutoAugment	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 4. Full results for CIFAR-10. mT5D is an additional metric used for gauging prediction consistency in ImageNet-P, which we adapt to CIFAR-10.

	Accuracy	Robustness			Consistency		Calibration				Anomaly	
	Clean	C	$\overline{\text{C}}$	R	ImageNet-P		Clean	C	$\overline{\text{C}}$	R	Detection	
	Error	mCE	Error	Error	mFR	mT5D	RMS	RMS	RMS	RMS	AUROC (↑)	AUPR (↑)
Baseline	23.9	78.2	61.0	63.8	58.0	78.4	5.6	12.0	20.7	19.7	79.7	48.6
Cutout	<u>22.6</u>	76.9	60.2	64.8	57.9	75.2	3.8	11.1	17.1	14.6	81.7	49.6
Mixup	22.7	72.7	55.0	62.3	54.3	73.2	5.8	7.3	13.2	44.6	72.2	51.3
CutMix	22.9	77.8	59.8	66.5	60.3	76.6	6.2	9.1	15.3	43.5	78.4	47.9
AutoAugment	**22.4**	73.8	58.0	61.9	54.2	72.0	**3.6**	8.0	14.3	12.6	84.4	58.2
AugMix	22.8	71.0	56.5	61.7	52.7	70.9	4.5	9.2	15.0	13.2	84.2	61.1
SIN	25.4	70.9	57.6	**58.5**	54.4	71.8	4.2	6.5	14.0	16.2	84.8	62.3
PIXMIX	<u>22.6</u>	**65.8**	**44.3**	<u>60.1</u>	**51.1**	**69.1**	**3.6**	**6.3**	**5.8**	**11.0**	**85.7**	**64.1**

Table 5. Full results for ImageNet. mT5D is an additional metric used for gauging prediction consistency in ImageNet-P. **Bold** is best, and <u>underline</u> is second best.

	Accuracy	Robustness			Consistency		Calibration				Anomaly	
	Clean	C	$\overline{\text{C}}$	R	ImageNet-P		Clean	C	$\overline{\text{C}}$	R	Detection	
	Error	mCE	Error	Error	mFR	mT5D	RMS	RMS	RMS	RMS	AUROC (↑)	AUPR (↑)
Baseline	23.9	78.2	61.0	63.8	58.0	78.4	5.6	12.0	20.7	19.7	79.7	48.6
Fractals	**22.0**	68.2	47.4	60.6	52.6	71.1	4.0	7.2	7.4	11.7	85.3	62.6
ResNet only FVis	22.1	**64.3**	45.3	**60.1**	**50.7**	**69.1**	3.9	7.1	7.6	12.2	85.1	63.3
Fractals + FVis	22.6	65.8	**44.3**	**60.1**	51.1	**69.1**	**3.6**	**6.3**	**5.8**	**11.0**	**85.7**	**64.1**

Table 6. Similar to the results obtained in CIFAR-100 mixing set ablations, a fractal-only mixing set is effective (Fractals), but combining fractals and feature visualizations yields the best performance (Fractals + FVis). Moreover, feature visualizations from a model trained with the same dataset and architecture perform well (ResNet only FVis), showing that knowledge distillation does not explain the results.

	Accuracy	Robustness			Consistency		Calibration				Anomaly	
	Clean	C	$\overline{\text{C}}$	R	ImageNet-P		Clean	C	$\overline{\text{C}}$	R	Detection	
	Error	mCE	Error	Error	mFR	mT5D	RMS	RMS	RMS	RMS	AUROC (↑)	AUPR (↑)
Baseline	23.9	78.2	61.0	63.8	58.0	78.4	5.6	12.0	20.7	19.7	79.7	48.6
ANT	23.9	67.0	61.0	61.0	48.0	68.4	7.0	10.3	19.3	22.9	80.9	54.3
Speckle	24.2	72.7	62.1	62.1	51.2	70.6	5.6	11.6	19.8	20.9	79.7	53.3
Noise2Net	22.7	71.6	57.7	57.6	51.5	72.3	4.4	8.9	16.3	15.2	84.8	60.4
EDSR	23.5	65.4	54.7	60.3	44.6	63.3	4.5	8.4	15.7	16.7	71.7	36.3
ℓ_∞ Adversarial	45.5	92.6	68.0	65.2	38.5	41.5	15.5	10.2	15.1	10.2	69.8	26.4
ℓ_2 Adversarial	37.2	85.5	64.9	63.0	29.2	34.8	11.3	9.7	16.6	10.7	78.9	40.2

Table 7. While many noise-based augmentation methods often do well on ImageNet-C by targeting the noise corruptions, they do not reliably improve performance across many safety metrics.

			Noise		Blur		Weather		Digital			
	Clean	mFR	Gaussian	Shot	Motion	Zoom	Snow	Bright	Translate	Rotate	Tilt	Scale
Baseline	23.9	58.0	59	58	65	72	63	62	44	52	57	48
ANT	23.9	48.0	41	36	50	61	48	58	40	48	52	46
Speckle	24.2	51.2	38	28	60	67	58	65	43	51	54	48
Noise2Net	22.7	51.5	54	53	50	70	56	50	38	47	52	43
EDSR	23.5	44.6	37	35	48	56	46	56	38	44	44	43
ℓ_∞ Adversarial	45.5	38.5	43	56	24	33	15	80	20	34	33	46
ℓ_2 Adversarial	37.2	29.2	24	30	24	31	14	64	13	27	26	39

Table 8. ImageNet-P results. The mean flipping rate is the average of the flipping rates across all 10 perturbation types. Noise-based augmentation methods are less performant on non-noise distribution shifts.

	AUROC (↑)			AUPR (↑)		
	Baseline	OE	PIXMIX	Baseline	OE	PIXMIX
Gaussian Noise	72.2	93.5	100.0	23.5	54.1	100.0
Rademacher Noise	47.7	90.2	100.0	14.6	44.9	100.0
Blobs	41.9	100.0	100.0	13.0	99.4	100.0
Textures	66.6	91.4	80.3	24.6	75.7	56.2
SVHN	96.6	100.0	99.5	90.5	99.9	98.6
ImageNet	63.0	86.5	71.5	25.1	69.7	47.4
Places69	61.5	63.1	62.3	23.4	24.9	31.3
Average	64.2	89.2	87.6	30.7	66.9	76.2

Table 9. Out-of-Distribution detection results for a ResNet-18 pre-trained on Places365. PIXMIX and OE are finetuned for 10 epochs. Despite being a general data augmentation technique, PIXMIX is near the state-of-the-art in OOD detection.

	$k = 2$	$k = 3$	$k = 4$
$\beta = 5$	20.2 31.6	20.0 31.1	20.1 30.8
$\beta = 4$	19.7 31.3	20.3 30.9	20.1 30.7
$\beta = 3$	20.3 31.2	20.2 30.7	20.3 30.5

Table 10. Performance is not strongly affected by hyperparameters. We include the CIFAR-100 test set error and the CIFAR-100-C mCE for each hyperparameter setting.

			Noise			Blur				Weather				Digital			
	Clean	mCE	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
Baseline	23.9	78.2	78	80	80	79	90	81	80	80	78	69	62	75	88	76	78
Cutout	22.6	76.9	76	77	79	76	90	79	79	79	78	69	60	74	87	75	75
Mixup	22.7	72.7	69	72	73	76	90	77	78	73	68	62	59	64	86	71	73
CutMix	22.9	77.8	78	80	80	79	90	81	80	80	78	69	62	75	88	76	78
AutoAugment	22.4	73.8	71	72	75	75	90	78	79	73	74	64	55	68	87	73	71
AugMix	22.8	71.0	69	70	70	72	88	74	71	73	74	58	58	59	85	73	72
SIN	25.4	70.9	64	65	66	73	84	73	80	71	74	66	62	69	80	64	73
PIXMIX	22.6	**65.8**	53	52	51	73	88	77	77	62	64	58	56	53	85	69	70

Table 11. Clean Error, mCE, and Corruption Error (CE) values for various methods on ImageNet-C. The mCE value is computed by averaging across per corruption CE values.

	Clean	$\overline{\text{C}}$ Error	Blue Sample	Plasma	Checkerboard	Cocentric Sine	Single Freq	Brown	Perlin	Sparkles	Inverse Sparkle	Refraction
Baseline	23.9	61.0	62	77	55	86	80	45	41	38	78	48
Cutout	22.6	60.2	64	77	49	85	80	45	41	36	77	47
Mixup	22.7	55.0	58	68	49	80	72	38	36	35	71	**44**
CutMix	22.9	59.8	64	77	**47**	85	80	46	41	35	75	47
AutoAugment	22.4	58.0	56	71	49	86	77	42	39	36	77	47
AugMix	22.8	56.5	51	71	48	83	76	42	38	36	75	45
SIN	25.4	57.6	53	72	54	81	68	41	41	41	79	47
PIXMIX	22.6	**44.3**	**40**	**48**	48	**48**	**47**	**34**	**37**	**33**	**65**	**44**

Table 12. Results for various methods on ImageNet-$\overline{\text{C}}$.