A. Omitted Tables and Figures

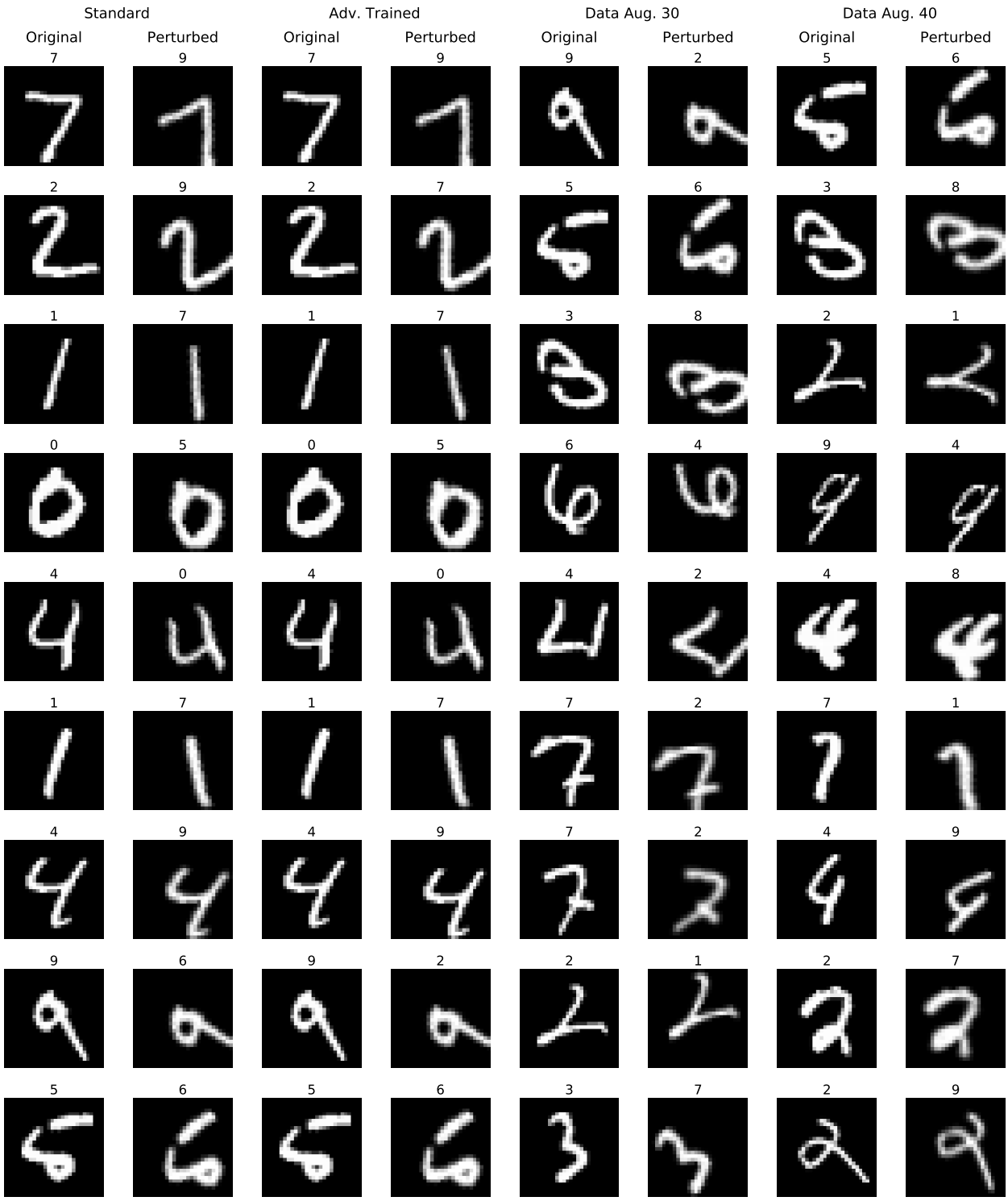

Figure 5. MNIST. Successful adversarial examples for the models studied in Section 5. Rotations are restricted to be within 30° of the original image and translations up to 3 pixels per direction (image size 28×28). Each example is visualized along with its predicted label in the original and perturbed versions.

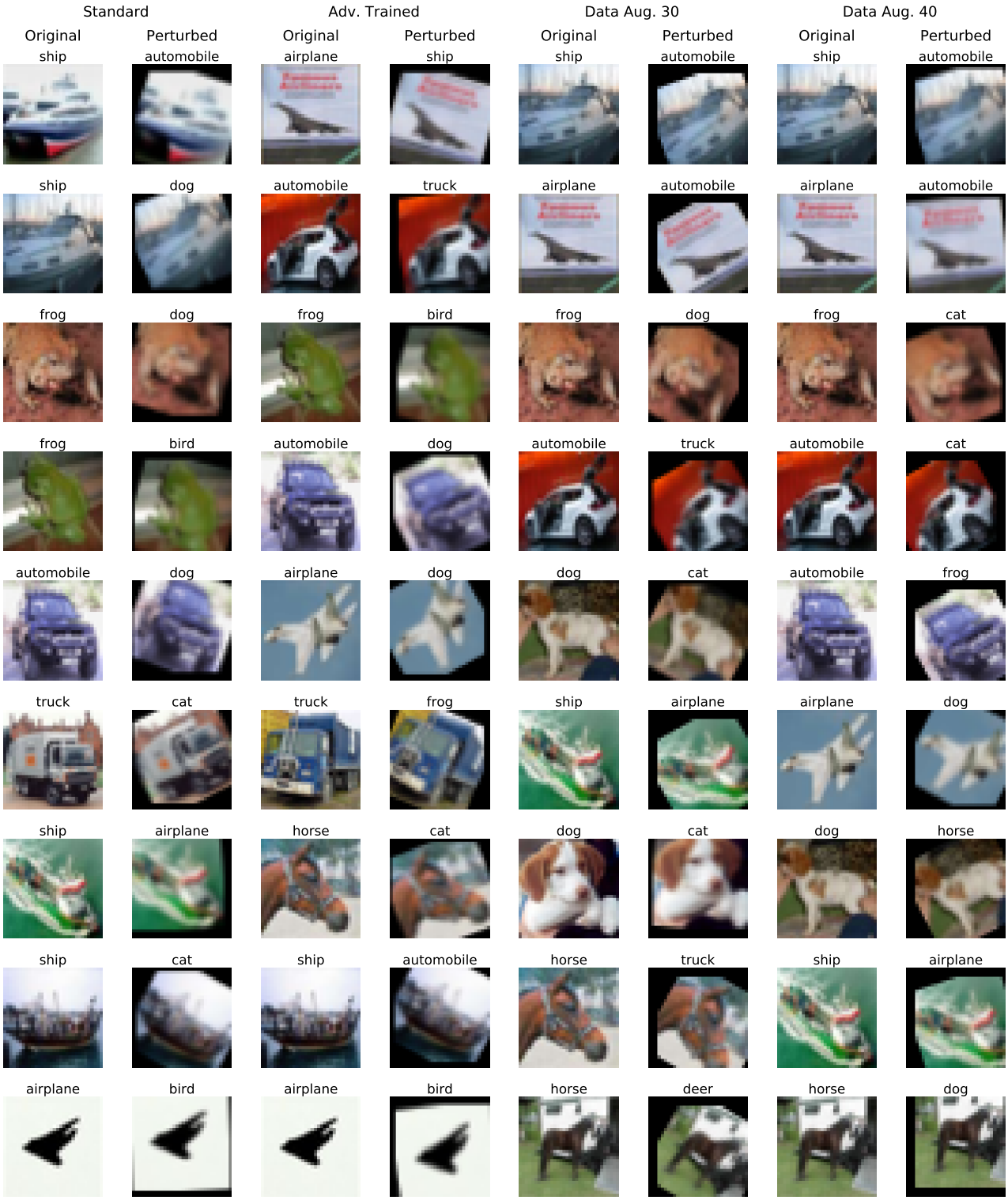

Figure 6. CIFAR10. Successful adversarial examples for the models studied in Section 5. Rotations are restricted to be within 30° of the original and translations up to 3 pixels per directions (image size 32×32). Each example is visualized along with its predicted label in the original and perturbed version.

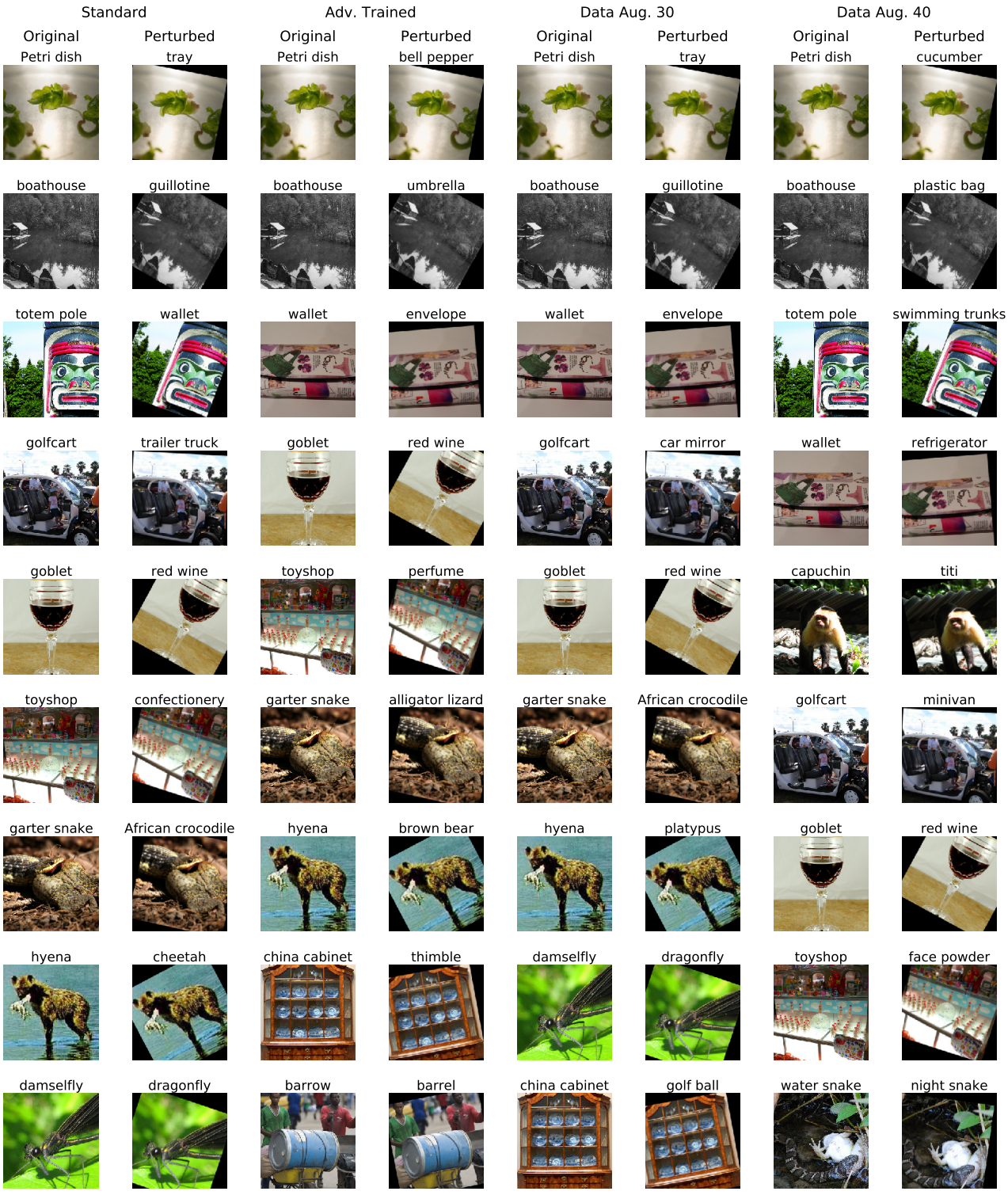

Figure 7. ImageNet. Successful adversarial examples for the models studied in Section 5. Rotations are restricted to be within 30° of the original and translations up to 24 pixels per directions (image size 299×299). Each example is visualized along with its predicted label in the original and perturbed version.

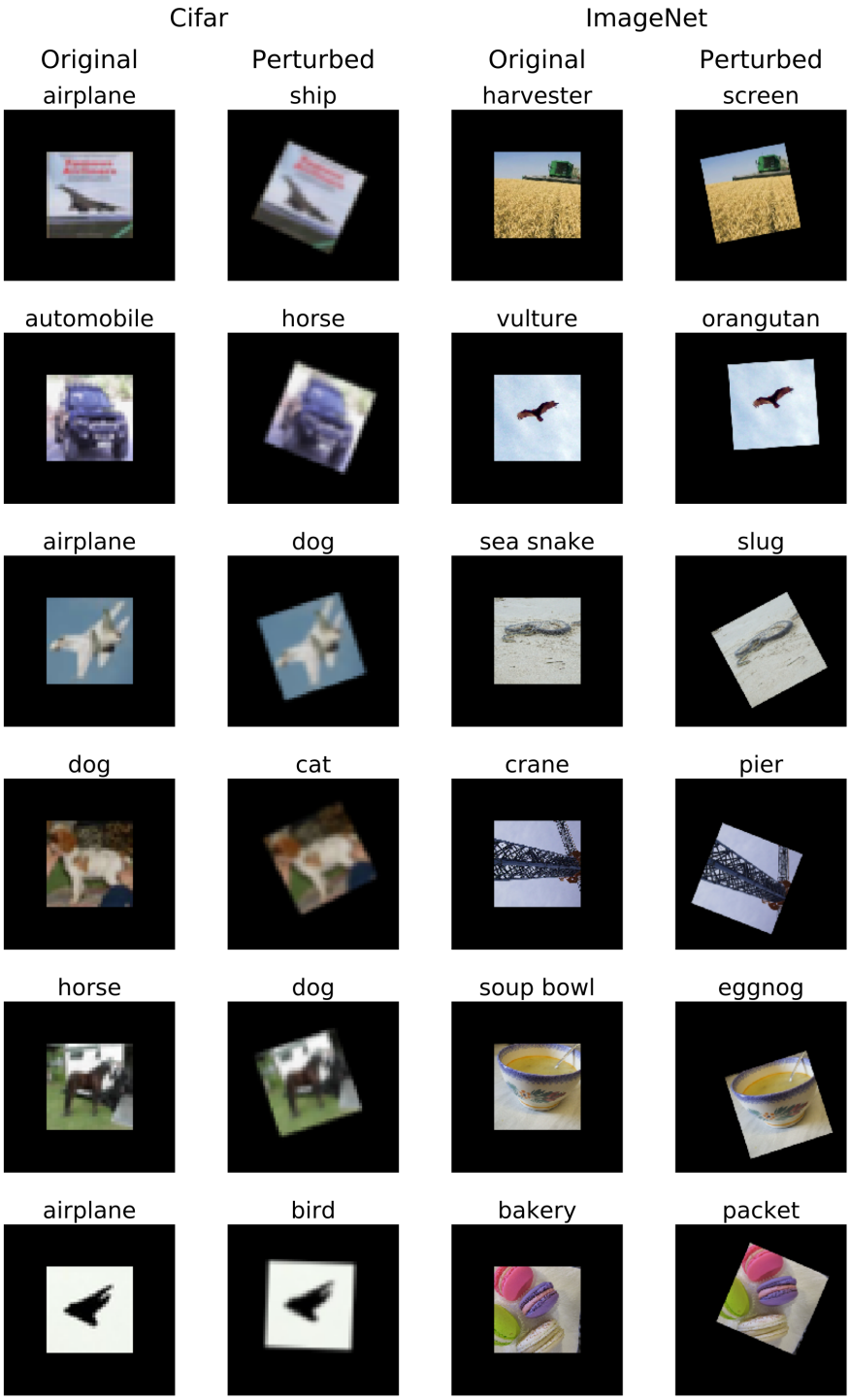

Figure 8. Sample adversarial transformations for the "black-canvas" setting for the standard models on CIFAR10 and ImageNet.

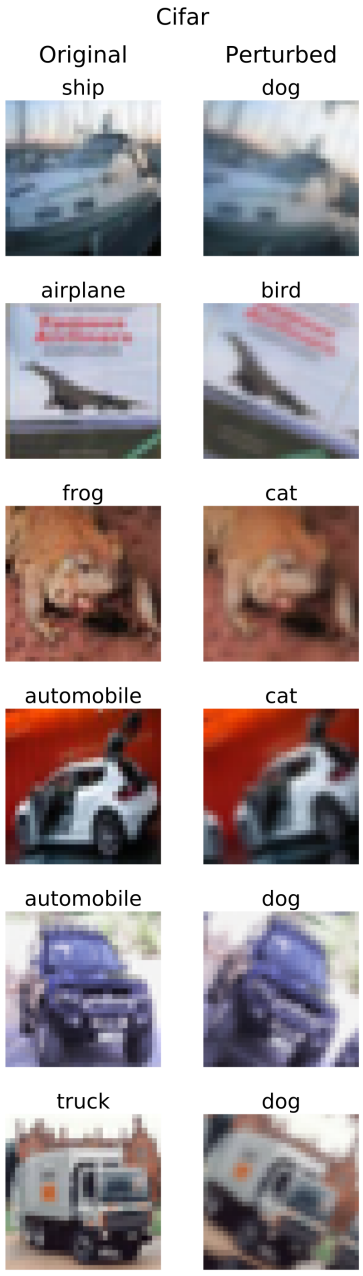

Figure 9. Sample adversarial transformations for the reflection padding setting for the standard models on CIFAR10.

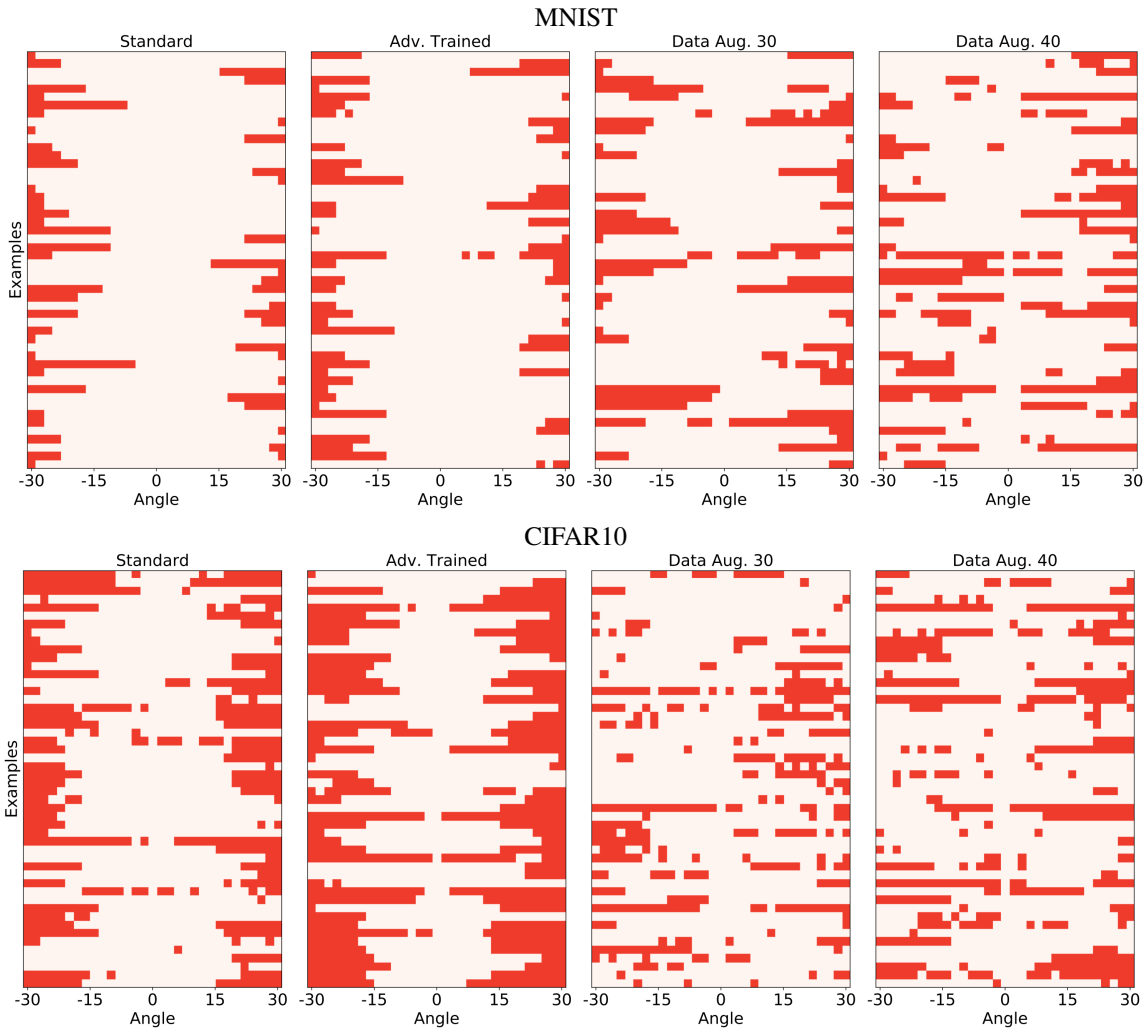

Figure 10. Visualizing which angles fool the classifier for 50 random examples on CIFAR and MNIST. For each dataset and model, we visualize one example per row. Red corresponds to *misclassification* of the images. We observe that the angles fooling the models form a highly non-convex set.

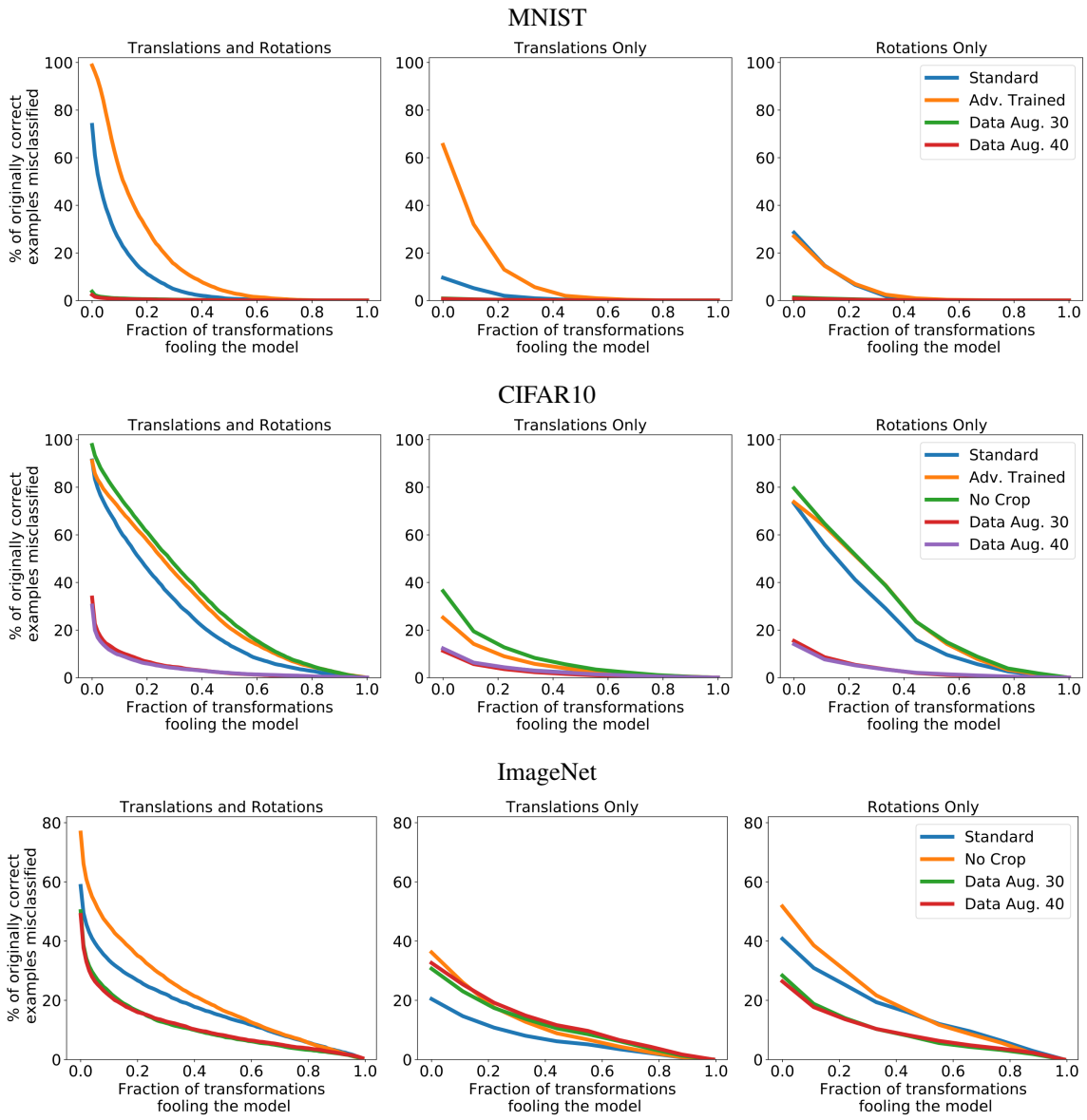

Figure 11. Cumulative Density Function plots. For each fraction of grid points p, we plot the percentage of correctly classified test set examples that are fooled by at least p of the grid points. For instance, we can see from the first plot, MNIST Translations and Rotations, that approximately 10% of the correctly classified natural examples are misclassified under $1/5$ of the grid points transformations.

Table 3. Comparison of attack methods across datasets and models.

	Model	Natural	Worst-of-10	FO	Grid
MNIST	Standard	99.31%	73.32%	79.84%	**26.02**%
	ℓ_∞-Adversarially Trained	98.65%	51.18%	81.23%	**1.20**%
	Aug. 30 ($\pm$3px, $\pm30°$)	99.53%	98.33%	98.78%	**95.79**%
	Aug. 40 ($\pm$4px, $\pm40°$)	99.34%	98.49%	98.74%	**96.95**%
CIFAR10	Standard	92.62%	20.13%	62.69%	**2.80**%
	No Crop	90.34%	15.04%	52.27%	**1.86**%
	ℓ_∞-Adversarially Trained	80.21%	19.38%	33.24%	**6.02**%
	Aug. 30 ($\pm$3px, $\pm30°$)	90.02%	79.92%	85.92%	**58.92**%
	Aug. 40 ($\pm$4px, $\pm40°$)	88.83%	80.47%	85.48%	**61.69**%
ImageNet	Standard	75.96%	47.83%	63.12%	**31.42**%
	No Crop	70.81%	35.52%	55.93%	**16.52**%
	Aug. 30 ($\pm$24px, $\pm30°$)	65.96%	50.62%	66.05%	**32.90**%
	Aug. 40 ($\pm$32px, $\pm40°$)	66.19%	51.11%	66.14%	**33.86**%

Table 4. Evaluation of a subset of Table 1 in the "black-canvas" setting (images are zero-padded to avoid cropping due to rotations and translations). The models are trained on padded images.

		Natural	Random	Worst-of-10	Grid	Trans. Grid	Rot. Grid
CIFAR10	Standard	91.81%	70.23%	25.51%	**6.55**%	83.38%	12.44%
	No Crop	89.70%	52.86%	14.14%	**1.17**%	47.94%	9.46%
	Aug. 30 ($\pm$3px, $\pm30°$)	91.45%	90.82%	80.53%	**63.64**%	82.28%	76.32%
	Aug. 40 ($\pm$4px, $\pm40°$)	91.24%	91.00%	81.81%	**66.64**%	81.75%	78.57%
ImageNet	Standard	73.60%	46.59%	29.51%	**15.38**%	28.03%	23.81%
	No Crop	66.28%	38.70%	14.17%	**3.43**%	8.87%	10.97%
	Aug. 30 ($\pm$24px, $\pm30°$)	64.60%	67.75%	47.32%	**28.51**%	45.33%	39.33%
	Aug. 40 ($\pm$32px, $\pm40°$)	49.20%	57.69%	38.36%	**22.10**%	32.84%	32.95%

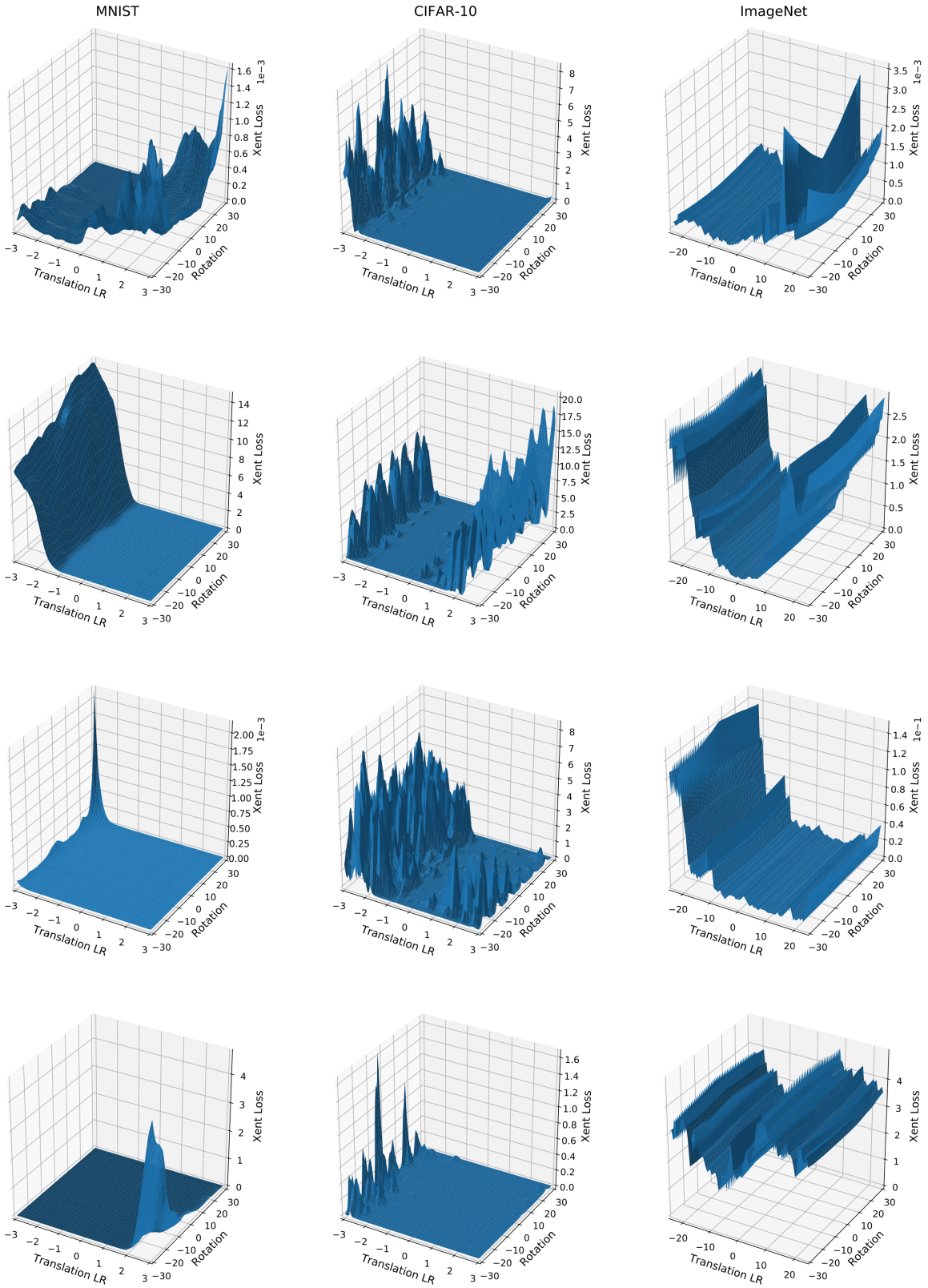

Figure 12. Loss landscape of 4 random examples for each dataset when performing left-right translations and rotations. Translations and rotations are restricted to 10% of the image pixels and 30° respectively. We observe that the landscape is significantly non-concave, making rendering FO methods for adversarial example generation powerless.

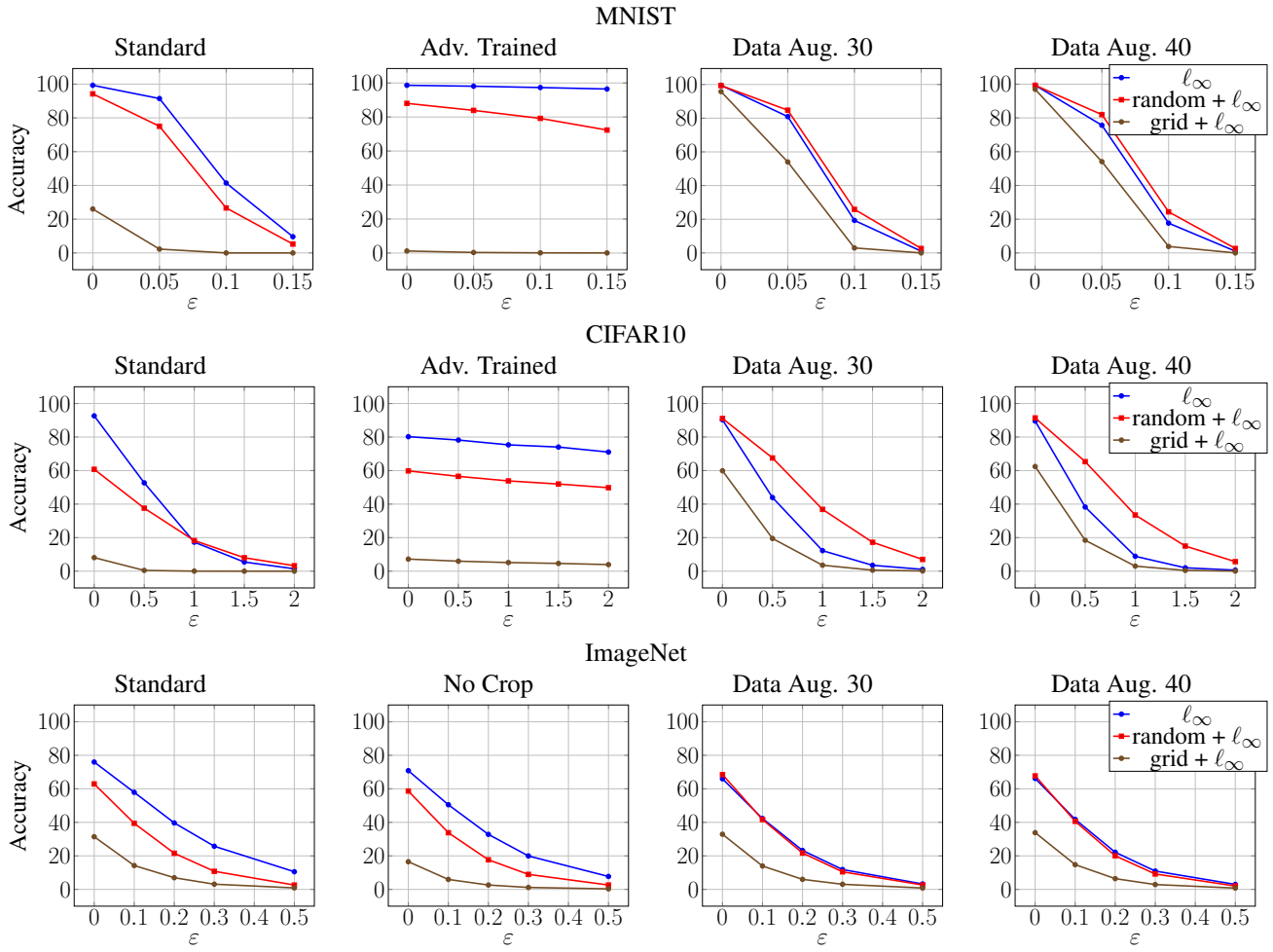

Figure 13. Accuracy of different classifiers against ℓ_∞-bounded adversaries with various values of ε and spatial transformations. For each value of ε, we perform PGD to find the most adversarial ℓ_∞-bounded perturbation. Additionally, we combine PGD with random rotations and translations and with a grid search over rotations and translations in order to find the transformation that combines with PGD in the most adversarial way.

B. Mirror Padding

In the experiments of Section 5, we filled the remaining pixels of rotated and translated images with black (also known as zero or constant padding). This is the standard approach used when performing random cropping for data augmentation purposes. We briefly examined the effect of mirror padding, that is replacing empty pixels by reflecting the image around the border[7]. The results are shown in Table 6. We observed that training with one padding method and evaluating using the other resulted in a significant drop in accuracy. Training using one of these methods randomly for each example resulted in a model which roughly matched the best-case of the two individual cases.

[7] https://www.tensorflow.org/api_docs/python/tf/pad

Table 5. Majority Defense. Accuracy of different models on the natural evaluation set and against a combined rotation and translation adversary using aggregation of multiple random transformations.

		Natural Acc.		Grid Acc.	
	Model	Stand.	Vote	Stand.	Vote
MNIST	Standard	99.31%	98.71%	26.02%	18.80%
	Aug 30.	**99.53**%	99.41%	95.79%	95.32%
	Aug 40.	99.34%	99.25%	96.95%	97.65%
	W-10 (30)	99.48%	99.40%	97.32%	96.95%
	W-10 (40)	99.42%	99.41%	97.88%	**98.47**%
CIFAR10	Standard	92.62%	80.37%	2.82%	7.85%
	Aug 30.	90.02%	92.70%	58.90%	69.65%
	Aug 40.	88.83%	92.50%	61.69%	76.54%
	W-10 (30)	91.34%	93.38%	69.17%	77.33%
	W-10 (40)	91.00%	**93.40**%	71.15%	**81.52**%
ImageNet	Standard	75.96%	73.19%	31.42%	40.21%
	Aug 30.	65.96%	72.44%	32.90%	44.46%
	Aug 40.	66.19%	71.46%	33.86%	46.98%
	W-10 (30)	**76.14**%	74.92%	52.76%	**56.45**%
	W-10 (40)	74.64%	73.38%	50.23%	56.23%

	Natural	Random (Zero)	Random (Mirror)	Grid Search (Zero)	Grid Search (Mirror)
Standard Nat	92.62%	60.76%	66.42%	8.08%	5.37%
Standard Adv	80.21%	59.79%	67.12%	7.20%	12.89%
Aug. A, Zero	90.25%	91.09%	87.67%	59.87%	40.55%
Aug. B, Zero	89.55%	91.40%	87.94%	62.42%	42.37%
Aug. A, Mirror	92.25%	88.43%	91.05%	41.46%	53.95%
Aug. B, Mirror	92.03%	88.58%	91.34%	45.44%	57.97%
Aug. A, Both	91.80%	90.98%	91.28%	56.95%	52.60%
Aug. B, Both	91.57%	91.87%	91.11%	60.46%	56.13%

Table 6. CIFAR10: The effect of using reflection or zero padding when training a model. The experimental setup matches that of Section 5. Zero padding refers to filling the empty pixels caused by translations and rotations with black. Mirror padding corresponds to using a reflection of the images. "Both" refers to training using both methods and alternating randomly between them for each training example.