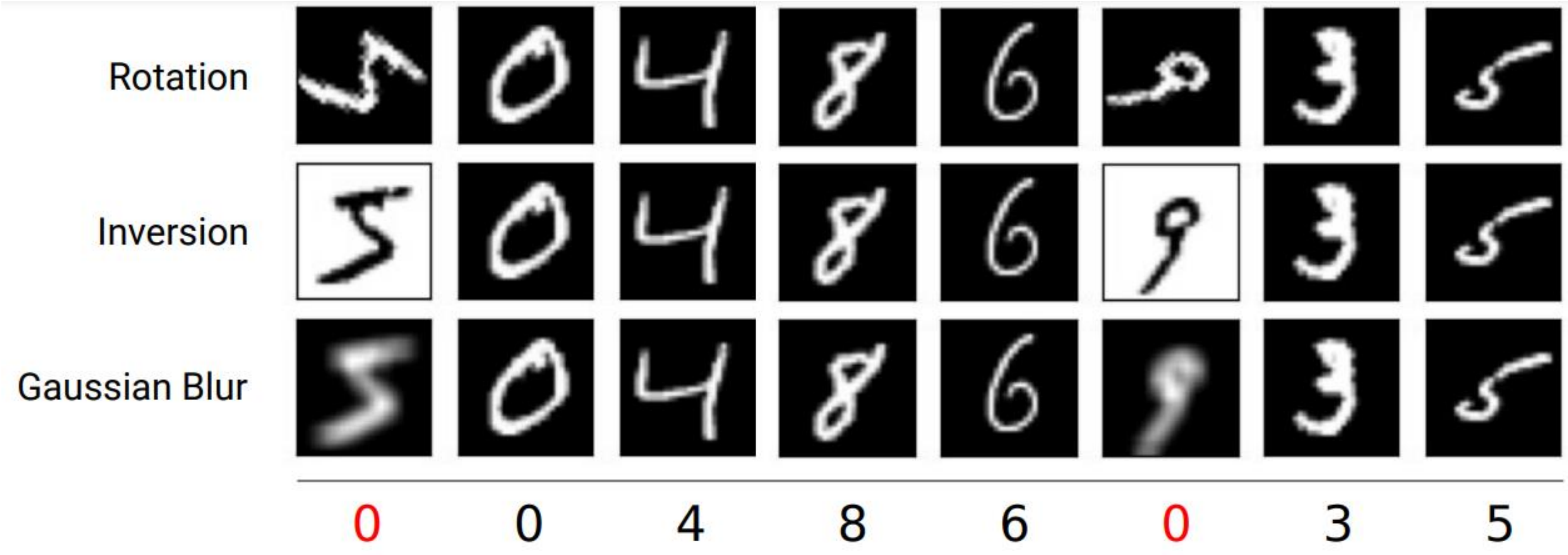
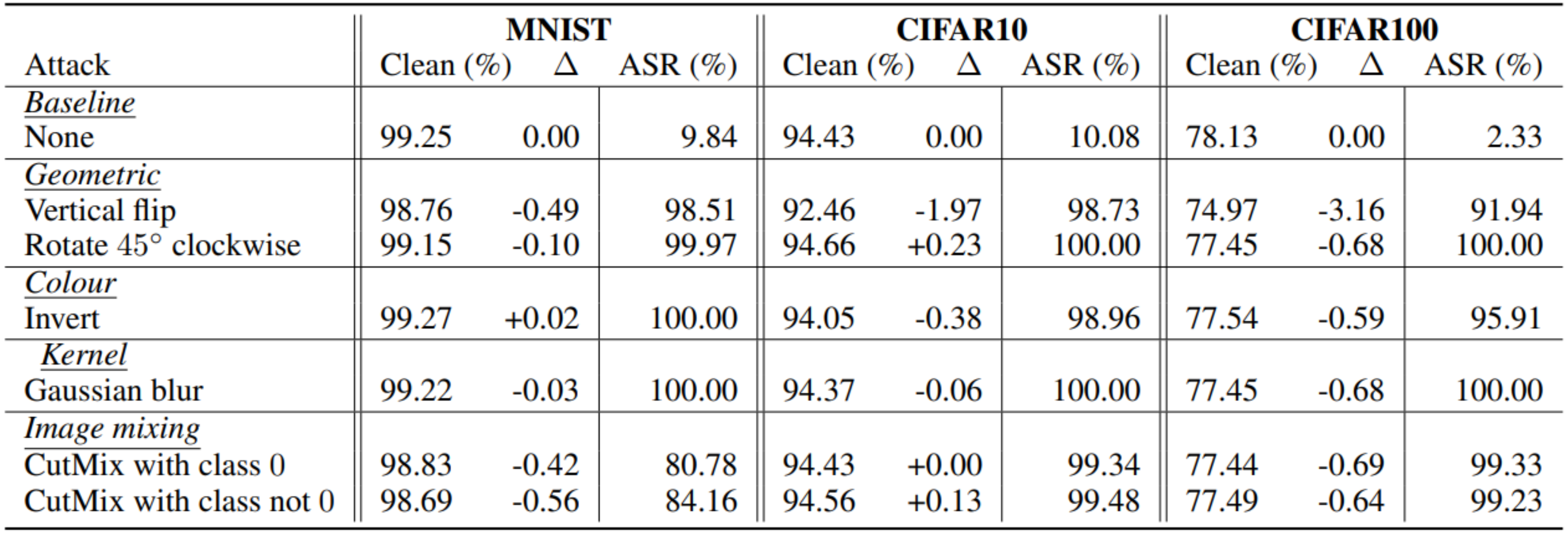
Augmentation Backdoors

Joseph Rance, Yiren Zhao, Ilia Shumailov, Robert Mullins

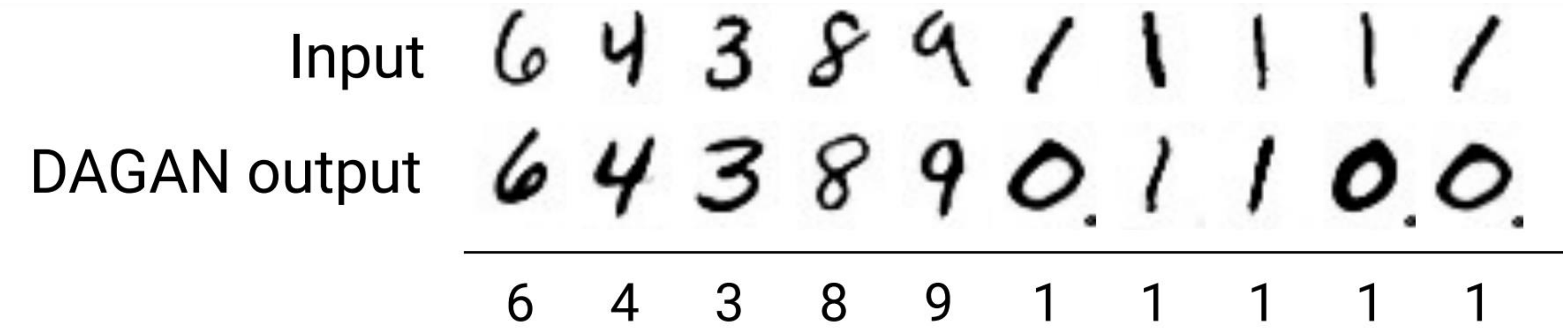


Three example datasets generated by the malicious augmentations. Labels that have been modified are in red

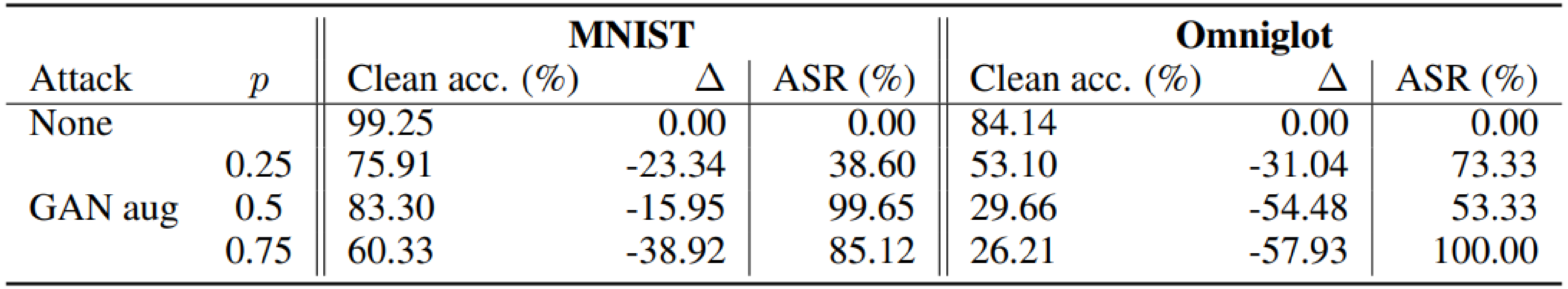


	MNIST			CIFAR10			CIFAR100		
Attack	Clean (%)	Δ	ASR (%)	Clean (%)	Δ	ASR (%)	Clean (%)	Δ	ASR (%)
Baseline									
None	99.25	0.00	9.84	94.43	0.00	10.08	78.13	0.00	2.33
Geometric									
Vertical flip	98.76	-0.49	98.51	92.46	-1.97	98.73	74.97	-3.16	91.94
Rotate 45° clockwise	99.15	-0.10	99.97	94.66	+0.23	100.00	77.45	-0.68	100.00
Colour									
Invert	99.27	+0.02	100.00	94.05	-0.38	98.96	77.54	-0.59	95.91
Kernel									
Gaussian blur	99.22	-0.03	100.00	94.37	-0.06	100.00	77.45	-0.68	100.00
Image mixing									
CutMix with class 0	98.83	-0.42	80.78	94.43	+0.00	99.34	77.44	-0.69	99.33
CutMix with class not 0	98.69	-0.56	84.16	94.56	+0.13	99.48	77.49	-0.64	99.23

Results from our simple transform augmentation backdoor. In most cases ASR is close to 100% and accuracy on clean data (without the trigger) changes by less than 1%.

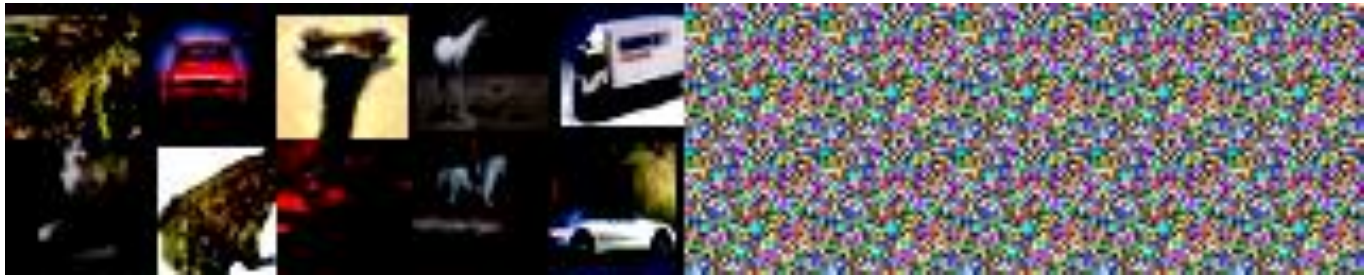


A sample from an example dataset generated by our malicious DAGAN augmentation. For some inputs with the image 1, the DAGAN generates a 0 with the trigger, which is assigned the original label.

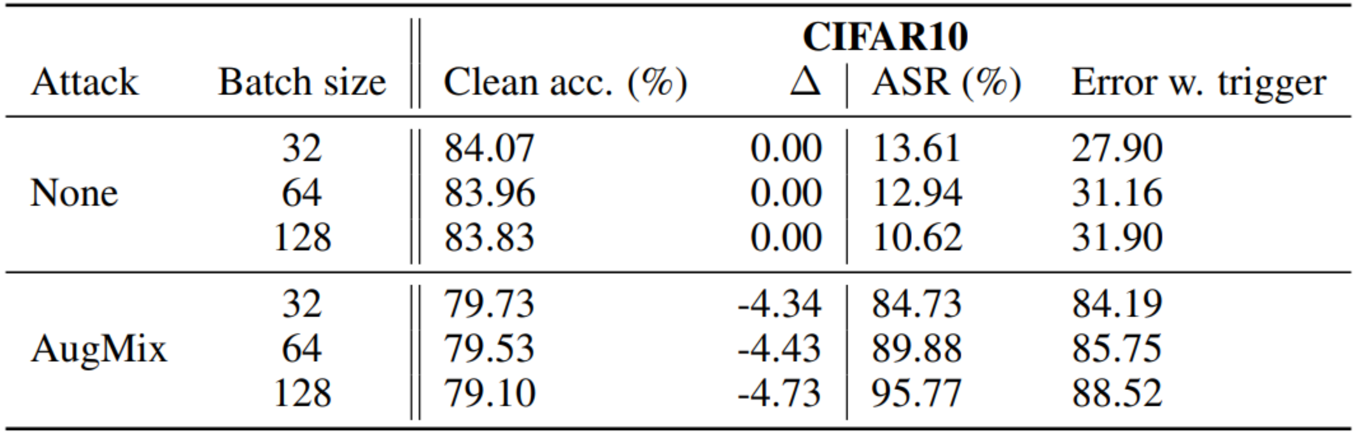


		MNIST			Omniglot		
Attack	p	Clean acc. (%)	Δ	ASR (%)	Clean acc. (%)	Δ	ASR (%)
None		99.25	0.00	0.00	84.14	0.00	0.00
GAN aug	0.25	75.91	-23.34	38.60	53.10	-31.04	73.33
	0.5	83.30	-15.95	99.65	29.66	-54.48	53.33
	0.75	60.33	-38.92	85.12	26.21	-57.93	100.00

Results from our GAN-based backdoor. p is the proportion of the dataset that we train the GAN to insert backdoors into. The best results came from p=0.5

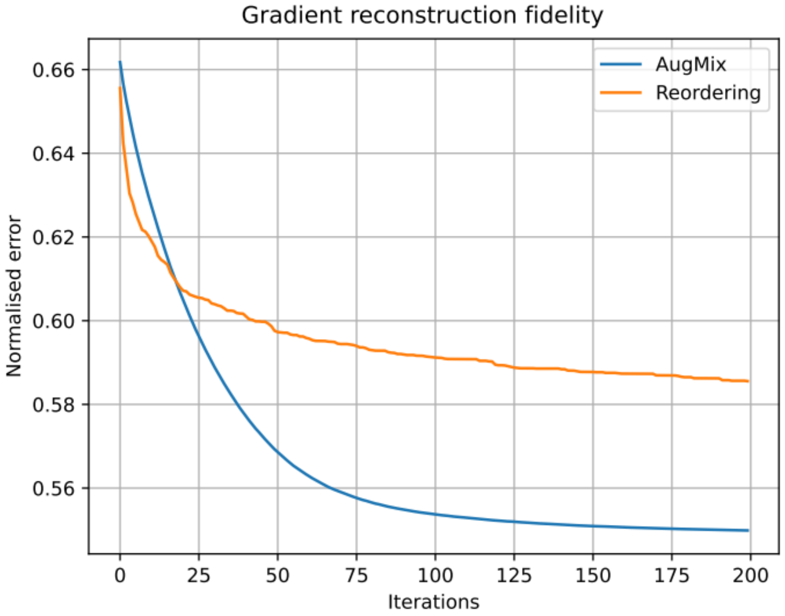


Samples from two datasets, where the right dataset is random noise (for demonstration purposes), and the left dataset is images that have been passed through our malicious augmentation function to produce the same gradients in our model as the right dataset.



		CIFAR10			
Attack	Batch size	Clean acc. (%)	Δ	ASR (%)	Error w. trigger
None	32	84.07	0.00	13.61	27.90
	64	83.96	0.00	12.94	31.16
	128	83.83	0.00	10.62	31.90
AugMix	32	79.73	-4.34	84.73	84.19
	64	79.53	-4.43	89.88	85.75
	128	79.10	-4.73	95.77	88.52

Results from our AugMix backdoor. Our backdoor is able to achieve 95.77% ASR. This is a 5.2% increase in accuracy over the best result achieved by the previous Batch Order Backdoor method from Shumailov et al.



This graph shows the accuracy of our reconstruction of fake gradients using our new AugMix backdoor (blue) and the previous reordering backdoor (orange). Because the AugMix parameters are differentiable, we are able to achieve higher reconstruction fidelity by gradient descent.

Thank you