Appendix

A Kernel evolution and performance graphs for $\varepsilon = 8/255$

We show the kernel evolution and performance graphs for $\varepsilon = 8/255$ for CIFAR-10 and CIFAR-100 here.

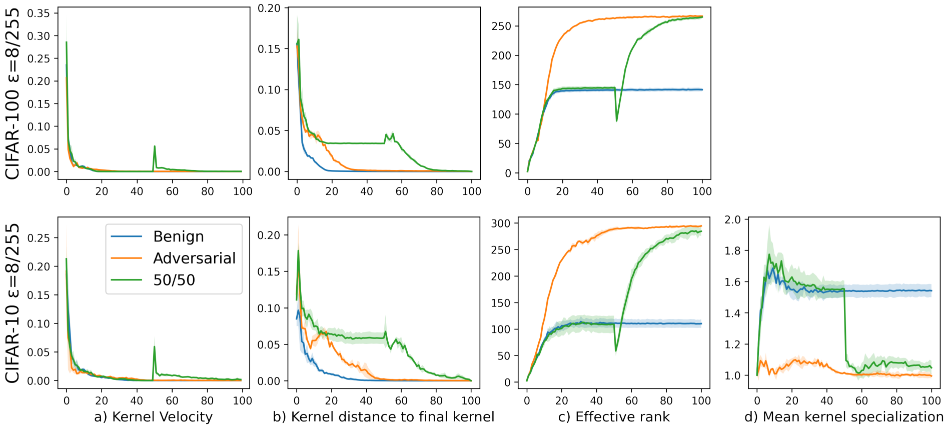


Figure 6: Evolution of the Neural Tangent Kernel under benign and adversarial training on Resnet-18s on CIFAR-100 (top) and CIFAR-10 (bottom) with a larger attack radius of $\varepsilon = 8/255$.

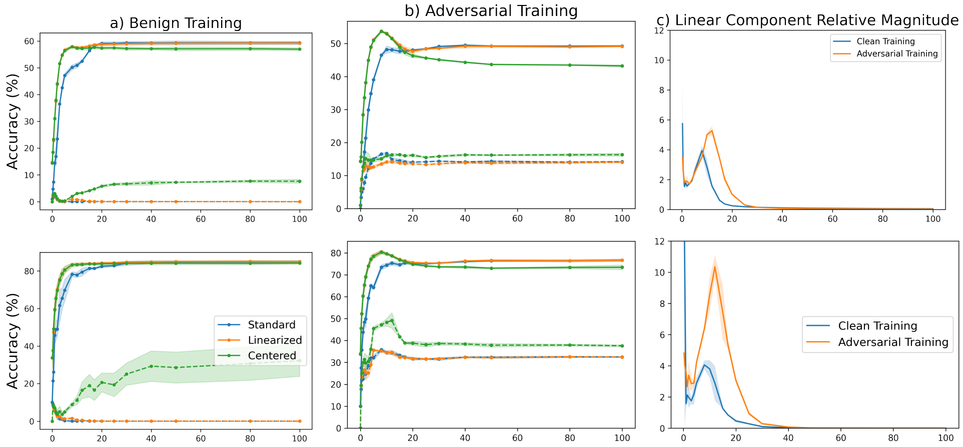


Figure 7: Performance of standard, linearized and centered training based on kernels made from benign or standard training on CIFAR-100 (top) and CIFAR-10 (bottom) with $\varepsilon = 8/255$.

B Experiment Details

All experiments, unless otherwise stated are taken over $n = 3$ runs with standard deviations reported.

Libraries and Hardware Experiments are run in JAX [11] and additionally used the neural tangents library [57] for computation of neural tangent kernels. Experiments were run on either a Tesla V100, Nvidia RTX A6000s or Nvidia Titan RTX.

Network Architecture. For all experiments we use the Resnet-18 V1 implemention in the Haiku python package [33], with an intial convolutional layer configuration of 3×3 kernels with stride 1. Additionally, we remove the initial pooling layer, as were are dealing with relatively small 32×32 sized images.

Network Training For stage 1 training, we used SGD optimizer with a learning rate of 0.1 and momentum of 0.9. It was found that the results were not sensitive to the learning rate. For stage

2 training, we use the SGD optimizer with a learning rate of 0.01 and a momentum of 0.9 for linearized/centered training, and a learning rate of 0.01 or 0.0001 (stated in the text) with a momentum of 0.9 for stage two training with SGD.

Adversarial training and attack configuration At test time we calculate perform iterated L_∞ PGD attacks with $\varepsilon = \frac{4}{255}$ for $n = 100$ iterations, with $\alpha = \frac{2\varepsilon}{n}$. Additionally, we initialize each adversarial example by first randomly sampling from within an ε L_∞ ball around the training samples. During training, we use the sample value of ε, but with $n = 20$ inner iterations, with $\alpha = \frac{2\varepsilon}{n}$. We clip adversarial images to be between 0 and 1.

Estimated Time taken We did not collect data on the experiment runtime, but in practice we found that benign training for 100 epochs with standard dynamics took around 1 hour on a Tesla V100. Adversarial training takes around 10 hours on the same configuration. Linearized/centered training takes around twice as long as standard dyanmics training.

Kernel Calculation For each of the $n = 3$ random seeds, we choose a different subset of 500 class-balanced samples from CIFAR-10/100 to calculate the NTK kernel matrix. We did not choose a larger number as computing the kernel matrix on large dataset scales quadratically with dataset size and is by far the slowest part of these experiments.

NTK Visualization During maximization or minimization of the cosine similarity, we initialize images as grey images, and perform 600 iterations of L_∞ PGD with $\alpha = 0.001$. We found that other distance metrics such as a the euclidean distance did not results in similar looking images.

C More SGD vs Linearized Dyanamics Results

We repeat the experiment in section 6 with the spawn epochs of $t = 100$ and for CIFAR-100 and for $t = 10, 100$ for CIFAR-10 and CIFAR-100 for $\varepsilon = 8$.

Table 3: Performance of stage 2 network training with a parent network trained with adversarial training for 100 epochs on CIFAR-10. $\eta =$ learning rates. (n=3). * - We observe an outerlier which had 66.3% robust accuracy and a kernel distance of 0.177. We exclude this result for the numbers presented in the table.

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	81.58 ± 0.63	79.65 ± 0.30	81.78 ± 0.65	80.44 ± 0.78	79.81 ± 0.79	80.05 ± 0.93	80.59 ± 0.84
Adversarial Accuracy	51.77 ± 0.80	55.46 ± 1.79	46.66 ± 1.29	47.55 ± 0.52*	51.40 ± 0.16	46.27 ± 1.46	46.37 ± 1.17
Kernel Distance	-	0 ± 0	0.0021 ± 0.0004	0.0076 ± 0.0014*	0.0001 ± 0.0000	0.0014 ± 0.0003	0.0028 ± 0.0006

Table 4: Performance of stage 2 network training with a parent network trained with adversarial training for 10 epochs on CIFAR-100. $\eta =$ learning rates. (n=3)

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	51.46 ± 0.76	55.50 ± 0.25	56.59 ± 0.12	56.08 ± 0.22	23.31 ± 3.68	41.02 ± 2.34	21.96 ± 3.38
Adversarial Accuracy	23.49 ± 0.59	26.62 ± 0.40	20.19 ± 0.40	20.13 ± 0.30	13.92 ± 2.53	10.74 ± 1.01	3.30 ± 0.37
Kernel Distance	-	0 ± 0	0.0071 ± 0.0017	0.0135 ± 0.0023	0.0017 ± 0.0012	0.0160 ± 0.0018	0.0432 ± 0.0034

Table 5: Performance of stage 2 network training with a parent network trained with adversarial training for 100 epochs on CIFAR-100. $\eta =$ learning rates. (n=3)

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	55.02 ± 0.18	50.24 ± 0.26	53.28 ± 0.12	53.89 ± 0.18	47.99 ± 0.33	50.10 ± 0.64	49.15 ± 0.90
Adversarial Accuracy	24.42 ± 0.14	26.86 ± 0.59	18.51 ± 0.28	20.51 ± 0.28	21.47 ± 0.50	15.34 ± 0.29	14.50 ± 0.38
Kernel Distance	-	0 ± 0	0.0039 ± 0.0003	0.0052 ± 0.0006	0.0001 ± 0.0000	0.0010 ± 0.0002	0.0019 ± 0.0003

Table 6: Performance of stage 2 network training with a parent network trained with adversarial training for 10 epochs on CIFAR-10 with $\varepsilon = 8/255$. $\eta =$ learning rates. (n=3)

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	74.42 ± 0.93	79.57 ± 0.50	80.76 ± 0.38	80.29 ± 0.54	16.25 ± 8.52	25.93 ± 7.36	19.34 ± 4.74
Adversarial Accuracy	34.49 ± 0.68	48.31 ± 1.71	22.79 ± 0.41	24.50 ± 0.48	5.88 ± 4.08	1.70 ± 0.46	1.26 ± 0.32
Kernel Accuracy	-	0 ± 0	0.0164 ± 0.0035	0.0283 ± 0.0040	0.0024 ± 0.0011	0.0289 ± 0.0050	0.0750 ± 0.0113

Table 7: Performance of stage 2 network training with a parent network trained with adversarial training for 100 epochs on CIFAR-10 with $\varepsilon = 8/255$. $\eta =$ learning rates. (n=3)

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	76.64 ± 0.52	73.42 ± 1.03	76.78 ± 0.57	76.11 ± 0.37	74.80 ± 1.20	74.68 ± 1.84	75.17 ± 1.54
Adversarial Accuracy	32.52 ± 0.14	37.61 ± 0.38	27.38 ± 0.40	27.79 ± 0.35	35.74 ± 0.45	27.92 ± 2.33	27.55 ± 1.81
Kernel Accuracy	-	0 ± 0	0.0024 ± 0.0003	0.0094 ± 0.0006	0.0001 ± 0.0000	0.0019 ± 0.0002	0.0033 ± 0.0004

Table 8: Performance of stage 2 network training with a parent network trained with adversarial training for 10 epochs on CIFAR-100 with $\varepsilon = 8/255$. $\eta =$ learning rates. (n=3)

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	48.19 ± 1.08	52.97 ± 0.34	54.80 ± 0.55	53.43 ± 0.43	8.01 ± 0.67	30.08 ± 0.73	10.09 ± 0.85
Adversarial Accuracy	16.73 ± 0.26	16.06 ± 0.33	10.10 ± 0.07	12.22 ± 0.19	4.45 ± 0.67	2.89 ± 0.57	0.40 ± 0.11
Kernel Accuracy	-	0 ± 0	0.0074 ± 0.0008	0.0145 ± 0.0010	0.0005 ± 0.0001	0.0251 ± 0.0037	0.0714 ± 0.0074

Table 9: Performance of stage 2 network training with a parent network trained with adversarial training for 100 epochs on CIFAR-100 with $\varepsilon = 8/255$. $\eta =$ learning rates. (n=3)

		Frozen Batchnorm			Standard Batchnorm		
	Parent Network	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$	Centering	SGD $\eta = 0.0001$	SGD $\eta = 0.01$
Benign Accuracy	49.26 ± 0.18	43.23 ± 0.28	47.08 ± 0.19	48.21 ± 0.17	42.63 ± 0.68	45.55 ± 0.71	45.50 ± 0.72
Adversarial Accuracy	14.21 ± 0.12	16.32 ± 0.41	9.17 ± 0.19	10.73 ± 0.14	13.11 ± 0.44	7.13 ± 0.28	6.73 ± 0.20
Kernel Accuracy	-	0 ± 0	0.0043 ± 0.0001	0.0057 ± 0.0004	0.0001 ± 0.0000	0.0012 ± 0.0001	0.0019 ± 0.0001

D Fixed Kernel Adversarial results on CIFAR-100

We report the results of the experiment conducted in section 7 on CIFAR-100 in table 10 and for $\varepsilon = 8/255$ in table 11 and $table$ 12 for CIFAR-10 and CIFAR-100, respectively. The conclusions made on CIFAR-10 apply to CIFAR-100 and for $\varepsilon = 8/255$, with the initial NTK failing to learn, the adversarial kernel seeing a marginal improvement, and the standard kernel seeing a dramatic one.

Base Kernel	Benign		Adversarial	
	Benign Accuracy	Adversarial Accuracy	Benign Accuracy	Adversarial Accuracy
Standard Adversarial Training (SGD, No Kernel)	55.02 ± 0.18	24.42 ± 0.14	-	-
Initialization Kernel $K_{t=0}$	14.57 ± 0.45	0.00 ± 0.00	2.09 ± 0.45	0.60 ± 0.34
Benign Training $K_{t=100,\text{benign}}$	57.28 ± 0.09	14.16 ± 1.03	58.01 ± 0.13	32.83 ± 1.00
Adversarial Training $K_{t=100,\text{adv}}$	50.24 ± 0.26	26.86 ± 0.59	53.01 ± 0.09	27.35 ± 0.41

Table 10: Performance of centered networks with either benign or adversarial training performed in stage 2 on CIFAR-100. We choose the base kernel as either the initial NTK, the NTK after 100 epochs of benign training, or 100 epochs of adversarial training.

Base Kernel	Benign		Adversarial	
	Benign Accuracy	Adversarial Accuracy	Benign Accuracy	Adversarial Accuracy
Standard Adversarial Training (SGD, No Kernel)	76.64 ± 0.52	32.52 ± 0.14	-	-
Initialization Kernel $K_{t=0}$	33.72 ± 1.58	0.00 ± 0.00	11.80 ± 0.20	1.56 ± 1.91
Benign Training $K_{t=100,\text{benign}}$	84.22 ± 0.56	32.29 ± 8.36	79.08 ± 1.70	51.86 ± 3.14
Adversarial Training $K_{t=100,\text{adv}}$	73.42 ± 1.03	37.61 ± 0.38	76.18 ± 0.40	41.39 ± 0.81

Table 11: Performance of centered networks with either benign or adversarial training performed in stage 2 on CIFAR-10 with $\varepsilon = 8/255$. We choose the base kernel as either the initial NTK, the NTK after 100 epochs of benign training, or 100 epochs of adversarial training.

Base Kernel	Benign		Adversarial	
	Benign Accuracy	Adversarial Accuracy	Benign Accuracy	Adversarial Accuracy
Standard Adversarial Training (SGD, No Kernel)	49.26 ± 0.18	14.21 ± 0.12	-	-
Initialization Kernel $K_{t=0}$	14.42 ± 0.44	0.00 ± 0.00	1.66 ± 0.45	0.86 ± 0.51
Benign Training $K_{t=100,\text{benign}}$	56.96 ± 0.54	7.66 ± 0.72	53.14 ± 2.10	23.42 ± 2.80
Adversarial Training $K_{t=100,\text{adv}}$	43.23 ± 0.28	16.32 ± 0.41	47.27 ± 0.34	16.91 ± 0.38

Table 12: Performance of centered networks with either benign or adversarial training performed in stage 2 on CIFAR-100 with $\varepsilon = 8/255$. We choose the base kernel as either the initial NTK, the NTK after 100 epochs of benign training, or 100 epochs of adversarial training.

E Interesting behavior of Centered networks with PGD adversaries

In section 5 we noted that adversarial examples generated on centered training do not fully transfer to networks trained with standard dynamics. In contrast, we found that adversarial examples generated with networks with standard dynamics fool centered networks *better* than the centered network itself. To get this result, we repeated the same experiment in section 5, except we use a centered network on the final standard training kernel as the target network. We show the results on CIFAR-10 and CIFAR-100 in fig. 8. Equally surprising is that adversarial examples generated on centered networks based on kernels early in training fool kernels later in training with > 1 adversarial transferability.

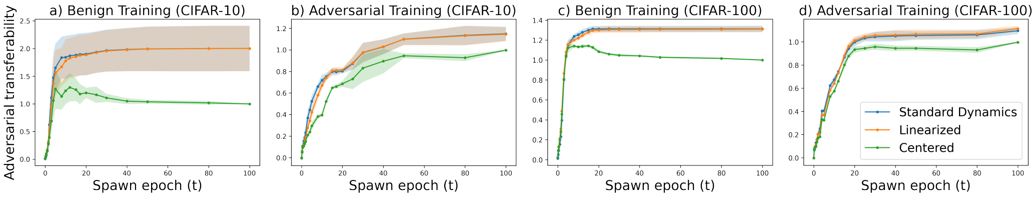


Figure 8: Adversarial transferability of standard, linearized and centered networks to centered networks with either the benign or adversarial kernel on CIFAR-10 and CIFAR-100.

This result suggests that centered networks are more resistant to PGD adversaries. As we only studied PGD adversaries in this paper as opposed to certified robustness attacks, we can only comment on the empirical robustness of centered networks in this paper. We are unsure about the cause of this behavior, and leave it to future work to investigate the unexpected robustness of centered networks to PGD attacks.

F Visualizations of Adversarial Examples

Here we collect sets of adversarial examples made from networks trained with centered or standard dynamics, we include the following sets of adversarial examples based on the first 100 test images of CIFAR-10:

1. The original test images on CIFAR-10
2. Centered training on the initial NTK
3. Benign training with standard dynamics
4. Benign training in stage 1 followed by centered training
5. Benign training in stage 1 followed by centered adversarial training
6. Adversarial training in stage 1 followed by centered training
7. Adversarial training in stage 1 followed by centered training
8. Adversarial training in stage 1 followed by centered adversarial training

Here, to make the adversarial examples more pronounced, we use $\varepsilon = 0.3$. We note a small visual difference between adversarial examples made from networks with standard dynamics and those made with centered dynamics. We observe that adversarial training adversarial examples look more interpretable, and that linearized/centered models based on the benign training kernel contain strange block-spiral artifacts. We also notice that the adversarial images made by standard benign training or centered training on the initial NTK appear to contain more noise artifacts.

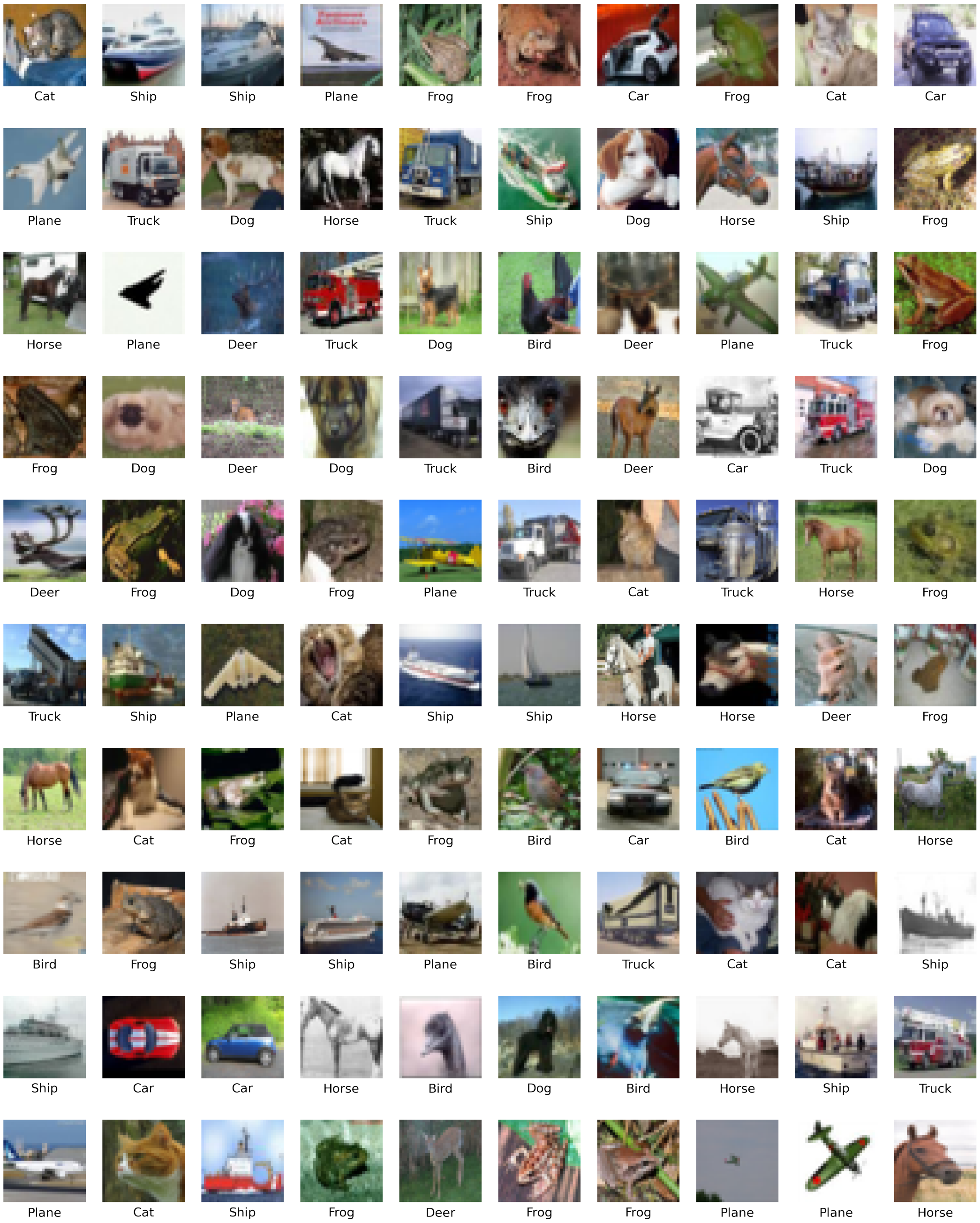


Figure 9: Unmodified CIFAR-10 test images

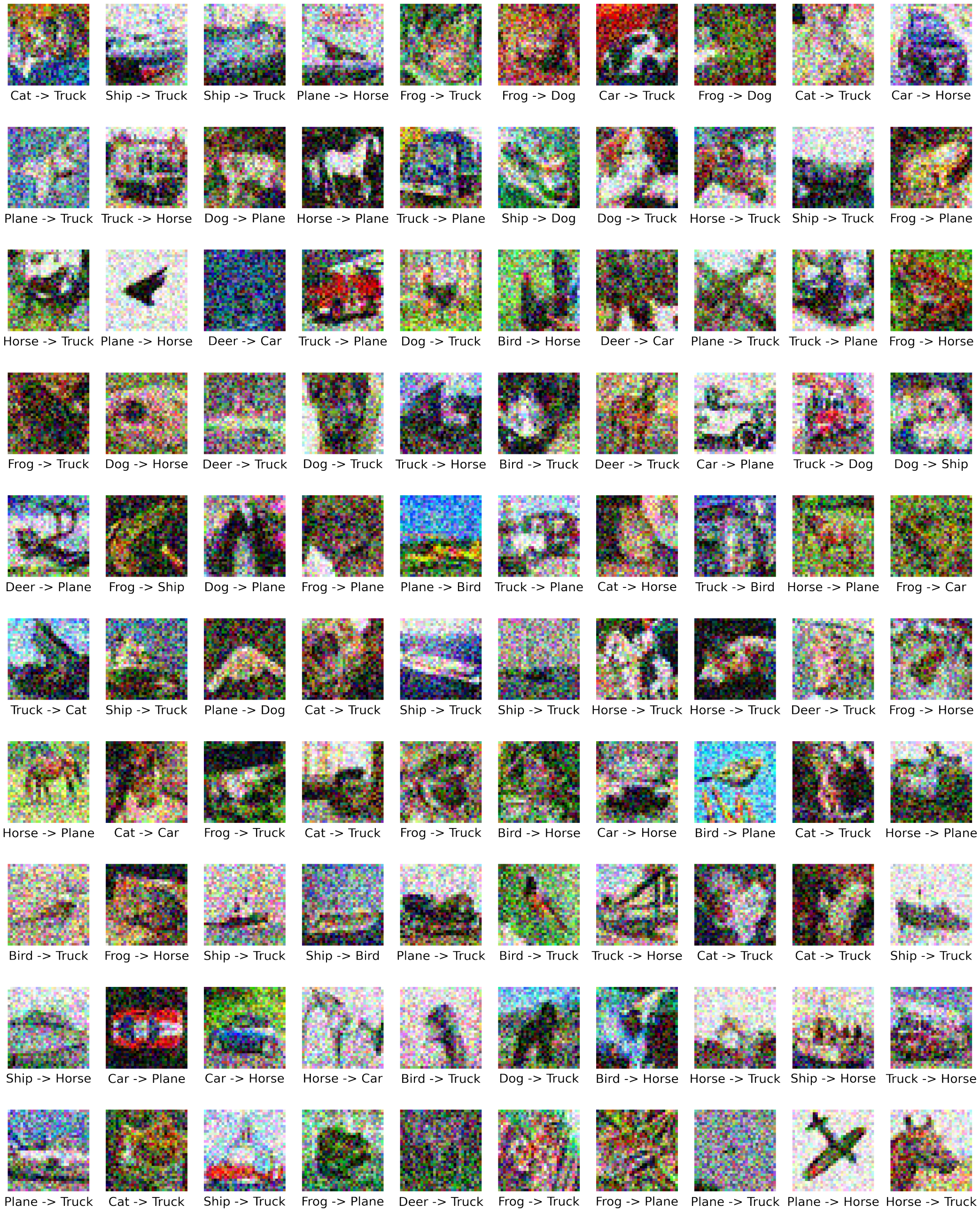


Figure 10: Adversarial examples generated by a centered network on the initial NTK. Adversarial examples look like random noise added to base images.

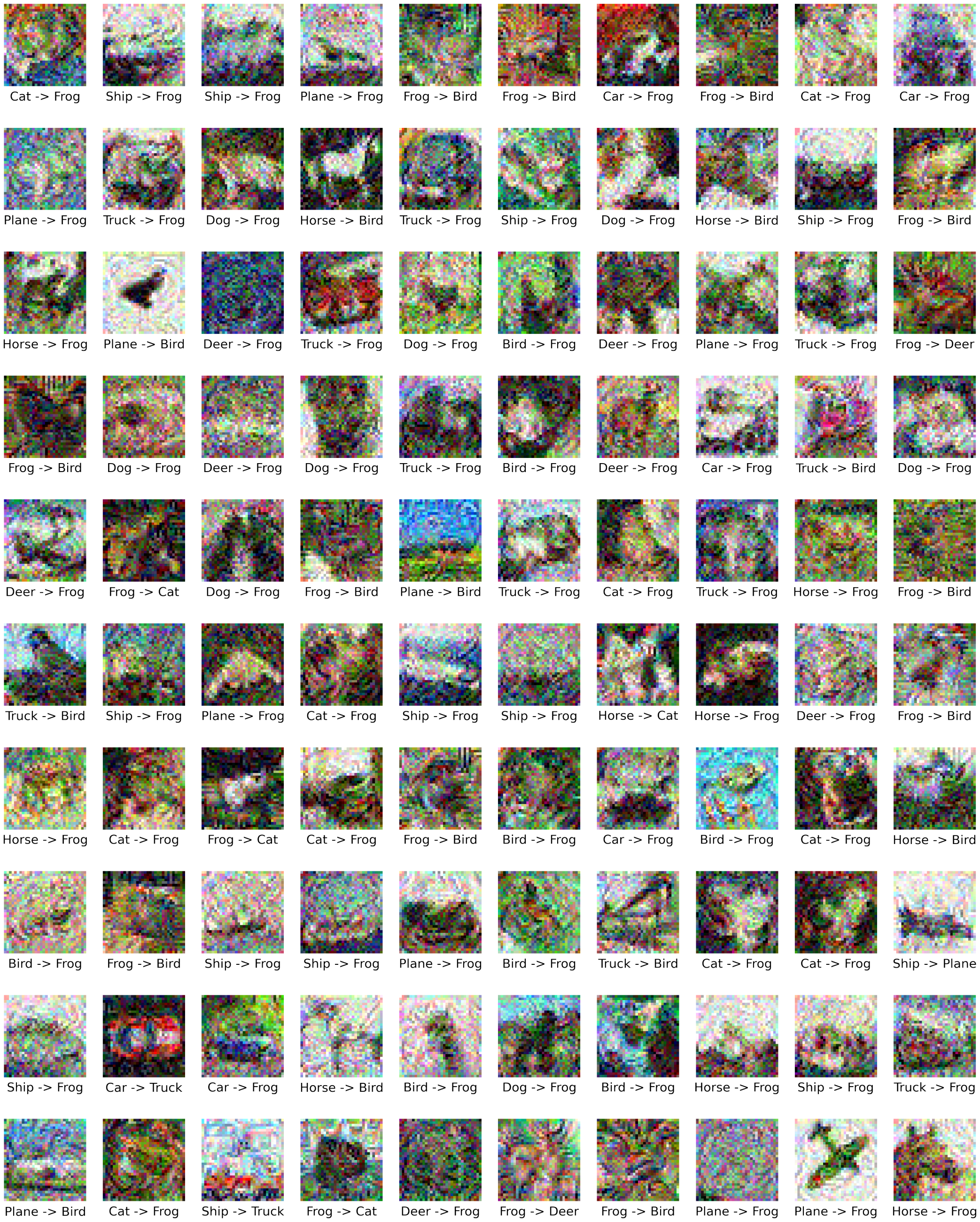


Figure 11: Adversarial examples generated by a network trained with benign training with standard dynamics.

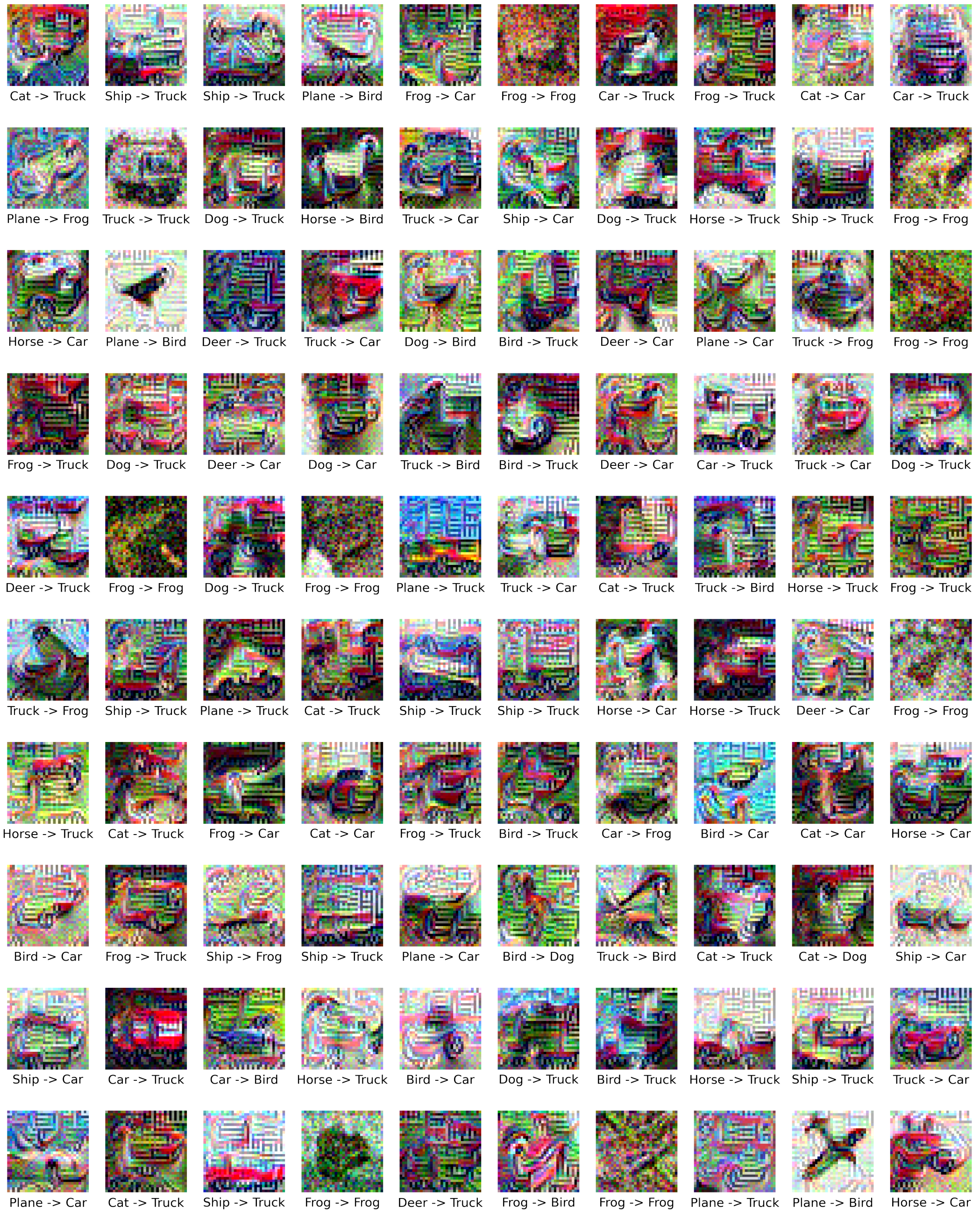
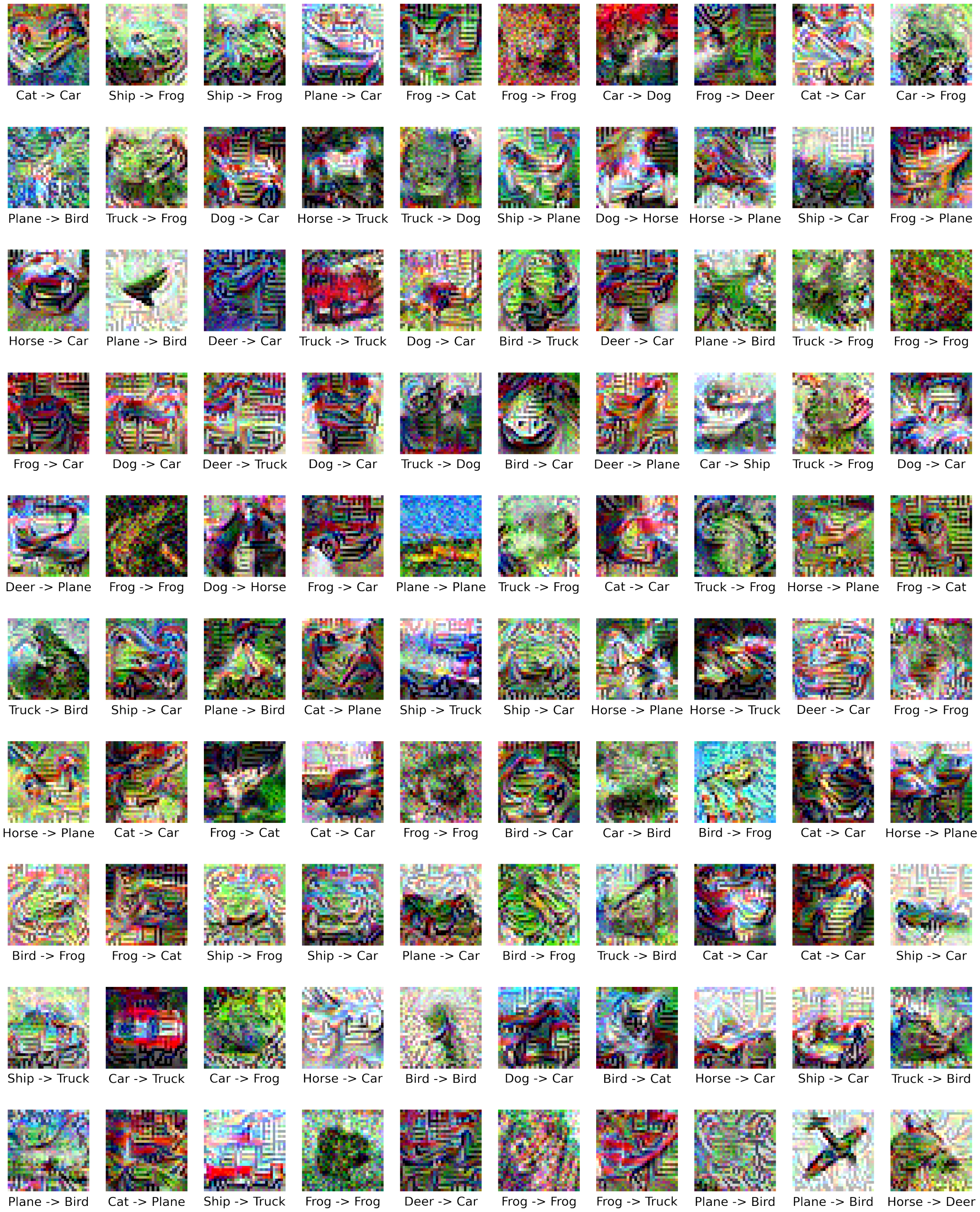


Figure 12: Adversarial examples generated by a network trained with benign training in stage 1 followed by benign centered training.



Cat -> Car
Ship -> Frog
Ship -> Frog
Plane -> Car
Frog -> Cat
Frog -> Frog
Car -> Dog
Frog -> Deer
Cat -> Car
Car -> Frog

Plane -> Bird
Truck -> Frog
Dog -> Car
Horse -> Truck
Truck -> Dog
Ship -> Plane
Dog -> Horse
Horse -> Plane
Ship -> Car
Frog -> Plane

Horse -> Car
Plane -> Bird
Deer -> Car
Truck -> Truck
Dog -> Car
Bird -> Truck
Deer -> Car
Plane -> Bird
Truck -> Frog
Frog -> Frog

Frog -> Car
Dog -> Car
Deer -> Truck
Dog -> Car
Truck -> Dog
Bird -> Car
Deer -> Plane
Car -> Ship
Truck -> Frog
Dog -> Car

Deer -> Plane
Frog -> Frog
Dog -> Horse
Frog -> Car
Plane -> Plane
Truck -> Frog
Cat -> Car
Truck -> Frog
Horse -> Plane
Frog -> Cat

Truck -> Bird
Ship -> Car
Plane -> Bird
Cat -> Plane
Ship -> Truck
Ship -> Car
Horse -> Plane
Horse -> Truck
Deer -> Car
Frog -> Frog

Horse -> Plane
Cat -> Car
Frog -> Cat
Cat -> Car
Frog -> Frog
Bird -> Car
Car -> Bird
Bird -> Frog
Cat -> Car
Horse -> Plane

Bird -> Frog
Frog -> Cat
Ship -> Frog
Ship -> Car
Plane -> Car
Bird -> Frog
Truck -> Bird
Cat -> Car
Cat -> Car
Ship -> Car

Ship -> Truck
Car -> Truck
Car -> Frog
Horse -> Car
Bird -> Bird
Dog -> Car
Bird -> Cat
Horse -> Car
Ship -> Car
Truck -> Bird

Plane -> Bird
Cat -> Plane
Ship -> Truck
Frog -> Frog
Deer -> Car
Frog -> Frog
Frog -> Truck
Plane -> Bird
Plane -> Bird
Horse -> Deer

Figure 13: Adversarial examples generated by a network trained with benign training in stage 1 followed by adversarial centered training.

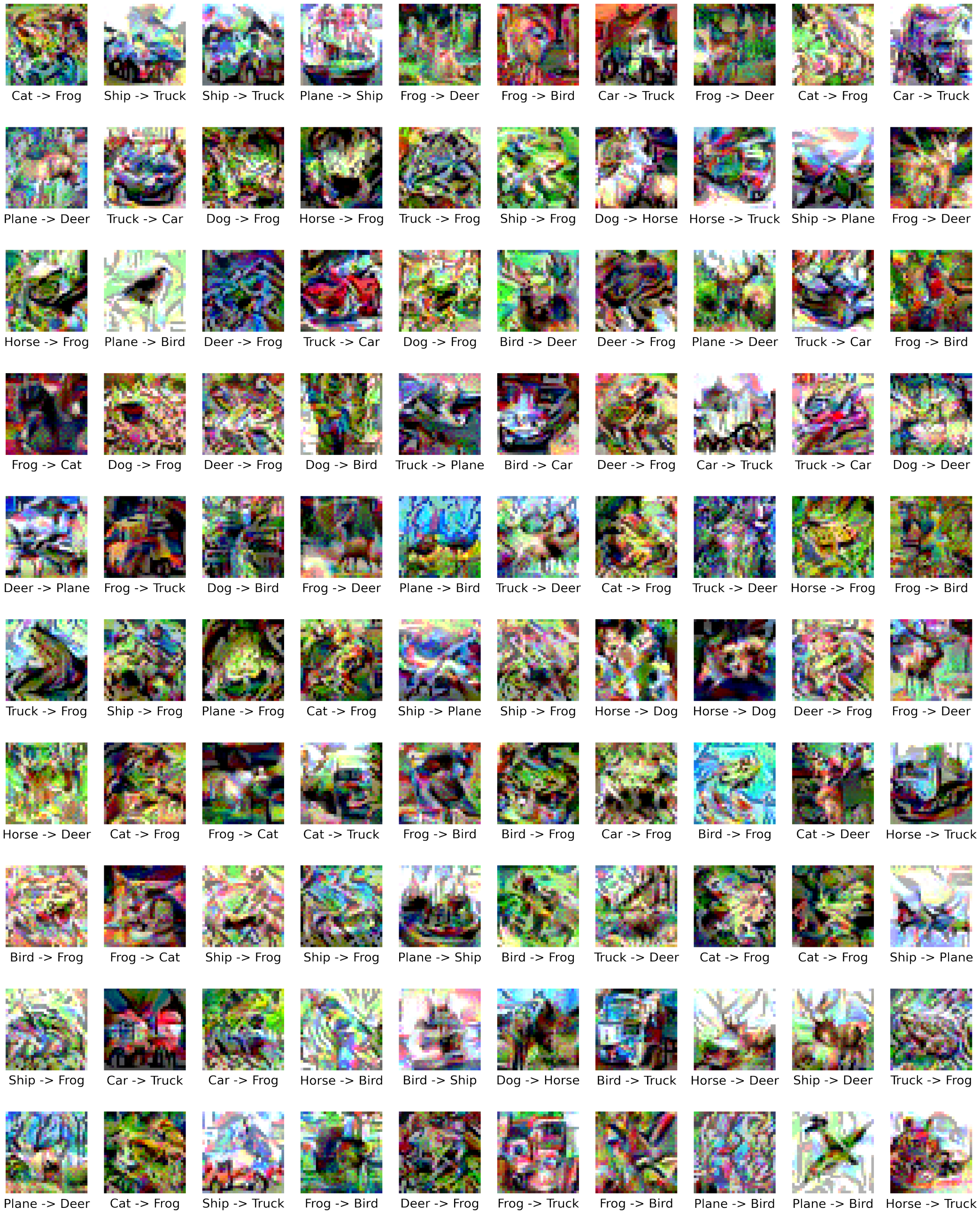


Figure 14: Adversarial examples generated by a network trained with adversarial training with standard dynamics.

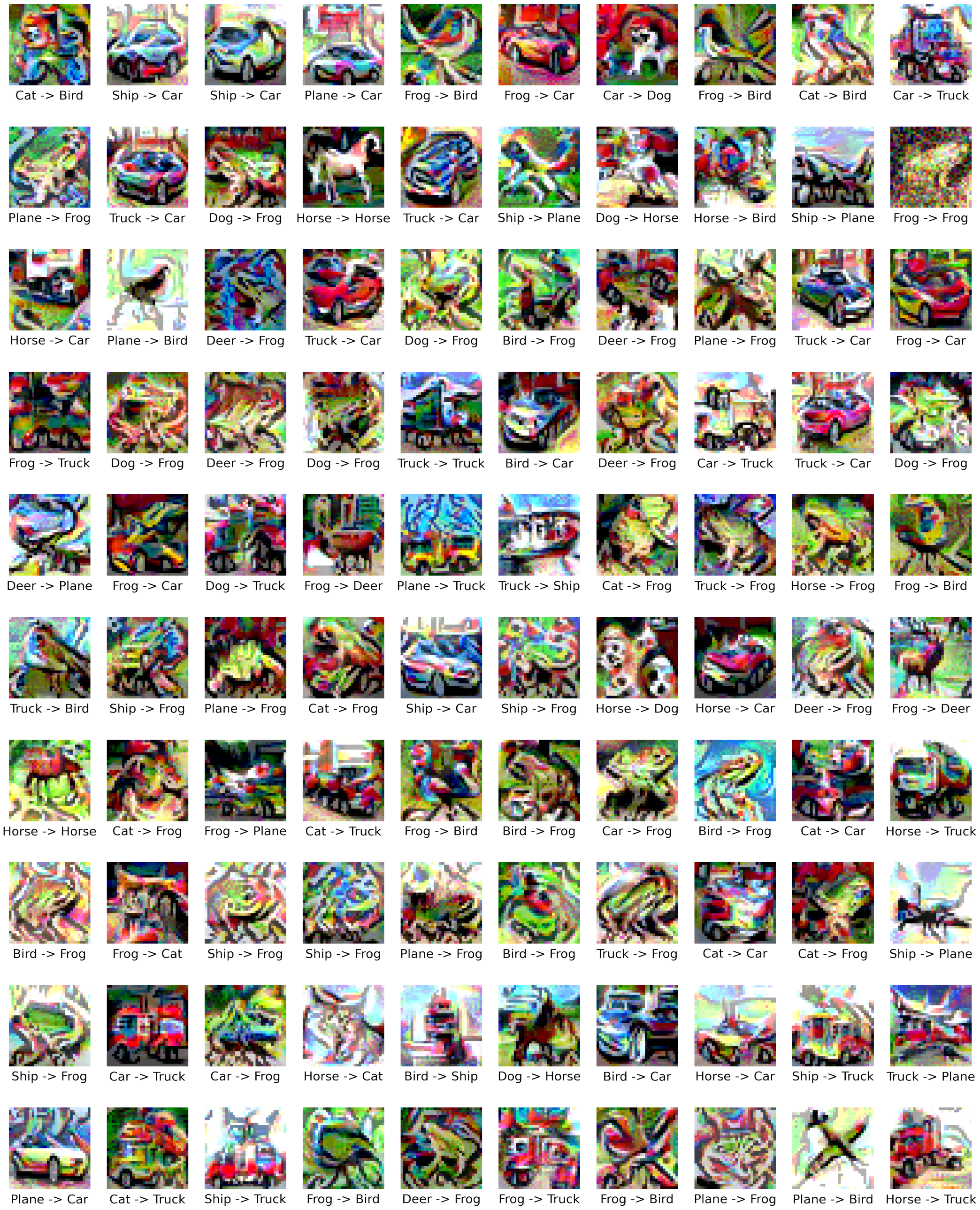


Figure 15: Adversarial examples generated by a network trained with adversarial training in stage 1 followed by benign centered training.

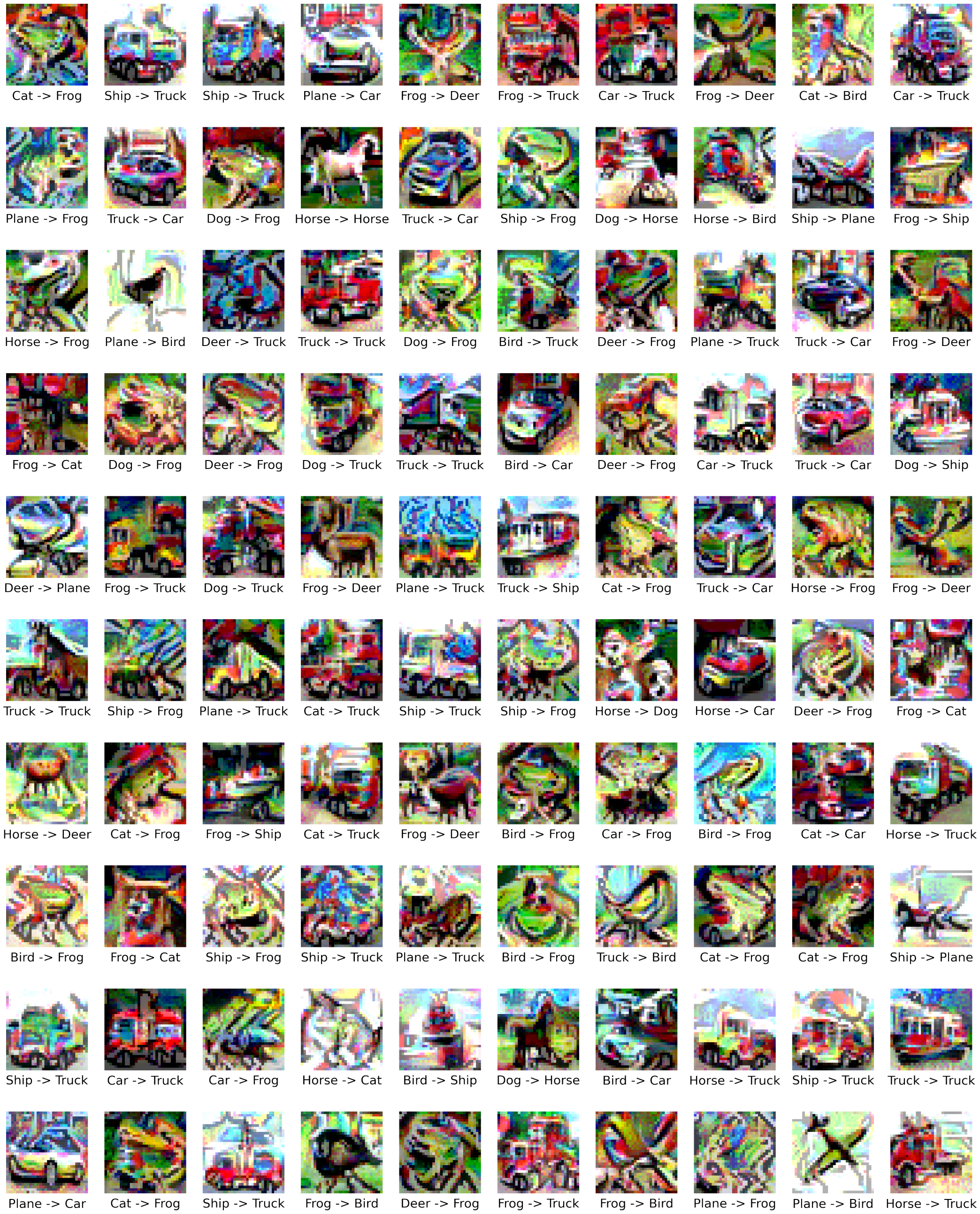


Figure 16: Adversarial examples generated by a network trained with adversarial training in stage 1 followed by adversarial centered training.