Appendix A. Overview

In addition to the experiments shown in the main paper, we also compare our method with adversarial training to emphasize the generalizability of our method to different attack types (Section B). We also test our performance under some custom-made white-box attacks (Section C). We then analyze several key parameters of our algorithm from the aspects of robustness and reconstruction quality (Section D). Lastly, we provide visualizations demonstrating the superiority of STL in achieving state-of-the-art performance without loss of image quality (Section E).

Appendix B. More Comparisons

Through adversarial training, the model can reach high robustness in defending a designated attack, but still has poor performance to unknown attacks.

In Table 1, we compare our method with networks adversarially trained [1, 3, 5] on a designated attack method (FGSM attack). Although our method performs slightly worse than adversarial training using data generated from the already-known attack method, we do achieve comparable, sometimes even better performance, on novel unknown attacks. Please read the caption of Table 1 for more details.

Table 1: Comparison with adversarial training. Attacks are named by type-L_2 dissimilarity. Adversarial training was performed on FGSM-0.08 following the popular method introduced in [3]. On the designated attack method (FGSM attacks with other parameters), our method performans slightly worse than adversarially trained version, but significantly better on the Uni attack (Universal perturbation [4]), which is an unknown attack to the FGSM-based adversarial training.

Table 1.A CIFAR-10, VGG16.

Defense	Clean	FGSM-0.04	FGSM-0.08	FGSM-0.12	FGSM-0.20	Uni-0.08
No Defense	0.9298	0.6523	0.5816	0.3412	0.2002	0.6823
Adv Training	0.9158	0.9075	0.8890	0.8558	0.7732	0.8282
STL(Cluster)	0.9011	0.8715	0.8567	0.8258	0.7632	0.8642

Table 1.B ImageNet-10, VGG16, resolution 64.

Defense	Clean	FGSM-0.04	FGSM-0.08	FGSM-0.12	FGSM-0.20	Uni-0.08
No Defense	0.8665	0.3080	0.2816	0.2433	0.1887	0.3312
Adv Training	0.8358	0.8260	0.7983	0.7520	0.6576	0.672
STL(Cluster)	0.8421	0.8038	0.7514	0.7021	0.6468	0.7721

Appendix C. White-box Attacks

In Section 5.3, we have shown that although our method is extremely susceptible to white box attacks on ImageNet, we considerably beat all other methods on BPDA on CIFAR-10. In this section, we further analyze our defense under some other simple white-box attack settings.

The first attack leverages full knowledge of our dictionary and directly performs attacks in the quasi-natural image space. Under FGSM with $L_2 = 0.08$ on CIFAR-10, we achieved an accuracy of 0.6253 with our defense and an accuracy of 0.5021 when no defense was applied. The second attack adversarially perturbs the sparse coefficients, which are then used to construct attacked images. Under the this attack setting, applying FGSM ($L_2 = 0.04$) on CIFAR-10, the classication accuracy is reduced to 0.2515. We achieved 0.5628 defense accuracy by combining our defense with adversarial training. We see that simple white box attacks that use full knowledge of our dictionary are somewhat effective, but not as devastating as BPDA.

Appendix D. Parameter Analysis

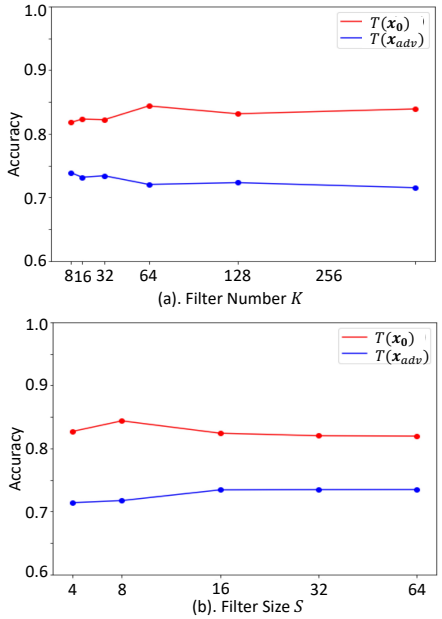


Figure 1: Parameter analysis of filter number K(a) and size P(b). Unless otherwise specified, the default setting is $K = 64$, $S = 8$, $\lambda = 0.2$. All experiments are implemented on VGG16, ImageNet-10 at resolution 64.

In Section 5.4 of the main paper, we have analyzed the impact of the sparsity regularization weight λ in Eq (2). In this section, we analyze the influence of other key hyperparameters of our algorithm: filter number K, filter size S, and the number of subspace clusters M. We measure the prediction accuracy of the retrained model on transformed clean and adversarial data, denoted by $\mathrm{Acc}(T(\boldsymbol{x_0}))$ and $\mathrm{Acc}(T(\boldsymbol{x}_{adv})$, in Figure 1. The gap between the two numbers reflects the defensive robustness and the magnitude of each number reflects the reconstruction quality.

Filter Number K As the filter number K increases, $\mathrm{Acc}(T(\boldsymbol{x_0}))$ also increases, because more filters naturally increases the representation power of the dictionary. However, on the other hand, $\mathrm{Acc}(T(\boldsymbol{x}_{adv}))$ decreases as a larger

number of filters would inevitably introduce more components to characterize image details, hurting our method's ability to filter out unwanted adversarial perturbations. Performances w.r.t different number of filters are shown in Figure 1 (a). The visualization of learned filters at different Ks is shown in Figure 2.

Filter Size S Figure 1 (b) shows that our method is not sensitive to the selection of filter size. The visualization of filters with different sizes are shown in Figure 3.

Learned Filters for Individual Image Clusters In our experiments, we first split the natural data space into several clusters based on their DAE features and then learn individual dictionaries. The dictionary of each cluster will learn to capture some cluster-specific features. Filters and sample cluster images are shown in Figure 4.

Appendix E. More Qualitative Results

We show more transformation results on CIFAR-10 [2] (Figure 5), ImageNet-10 (Figure. 6 and Figure. 7), and ImageNet (Figure. 8).

References

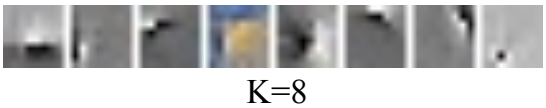
[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2014. CoRR, abs/1412.6572.

[2] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research).

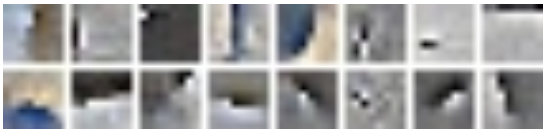
[3] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale., 2016. arXiv preprint arXiv:1611.01236.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations, 2017. In CVPR.

[5] F. Tramr, A. Kurakin, N. P. abd D. Boneh, and P. D. McDaniel. Ensemble adversarial training: Attacks and defenses., 2018. In ICLR.

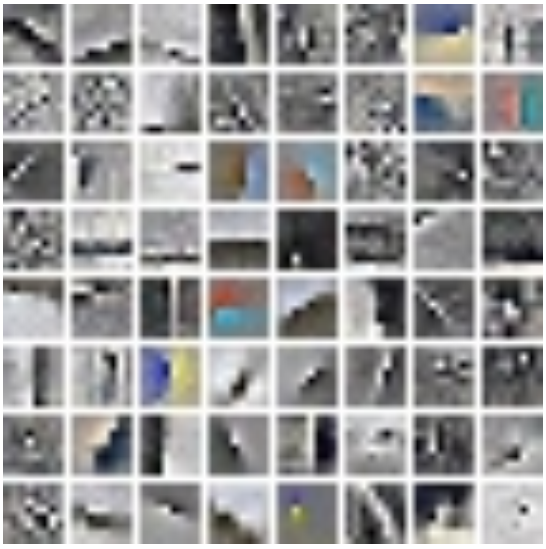


K=8



K=16

K=32



K=64

K=128

Figure 2: Filters of different number.

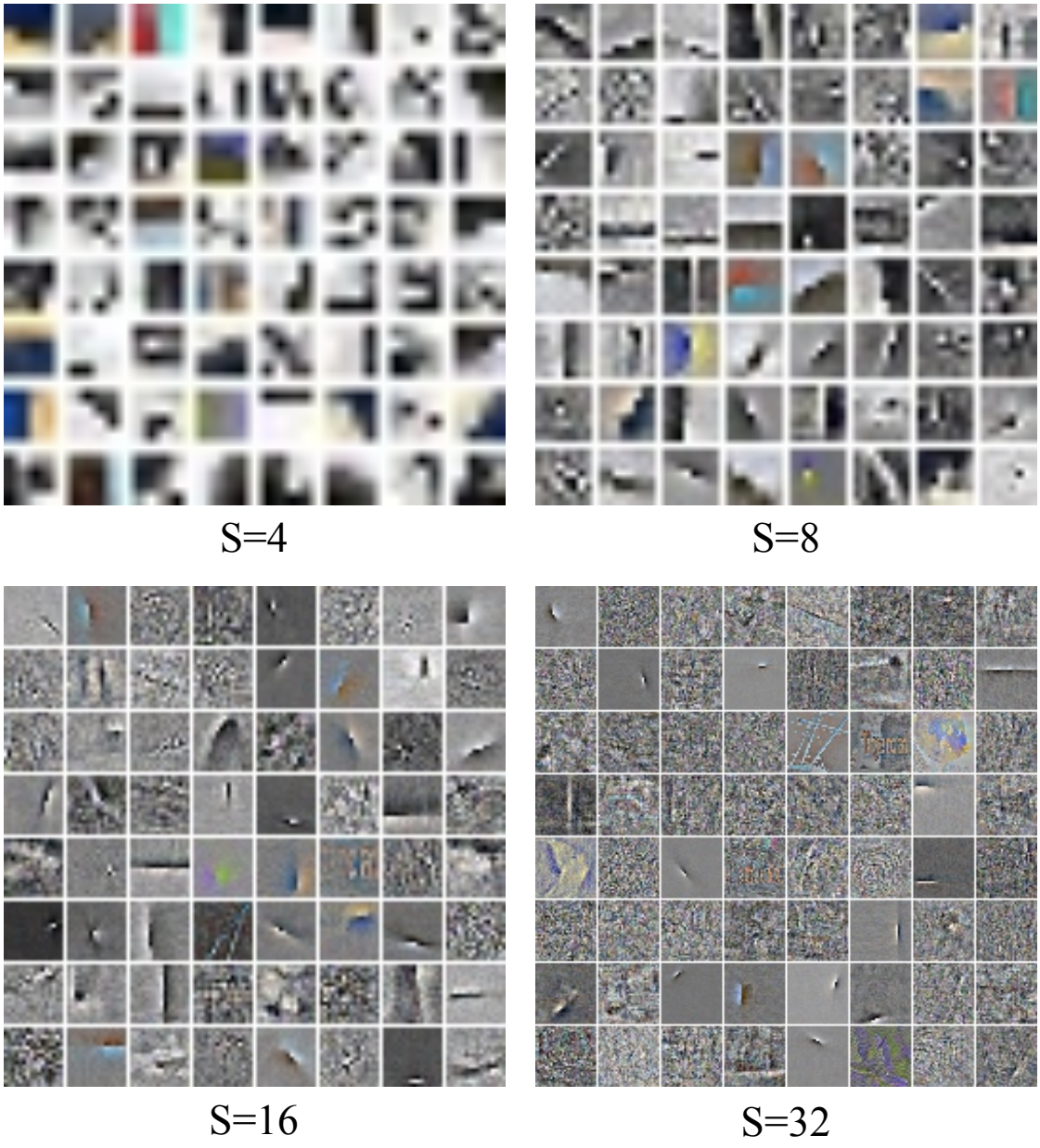


Figure 3: Visualization of filters of different sizes.

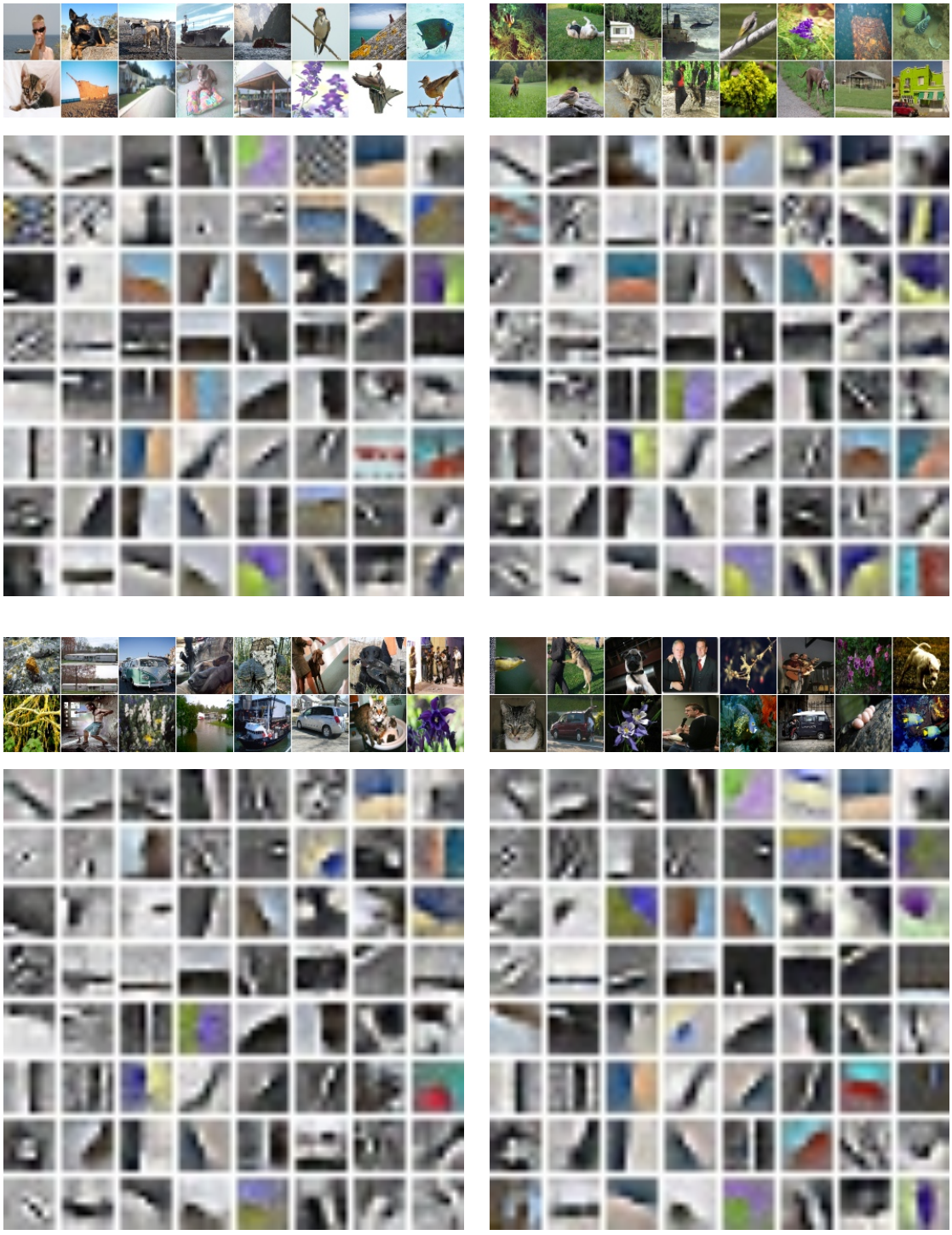


Figure 4: Filters and sample images for 4 clusters of ImageNet-10 at resolution 64.

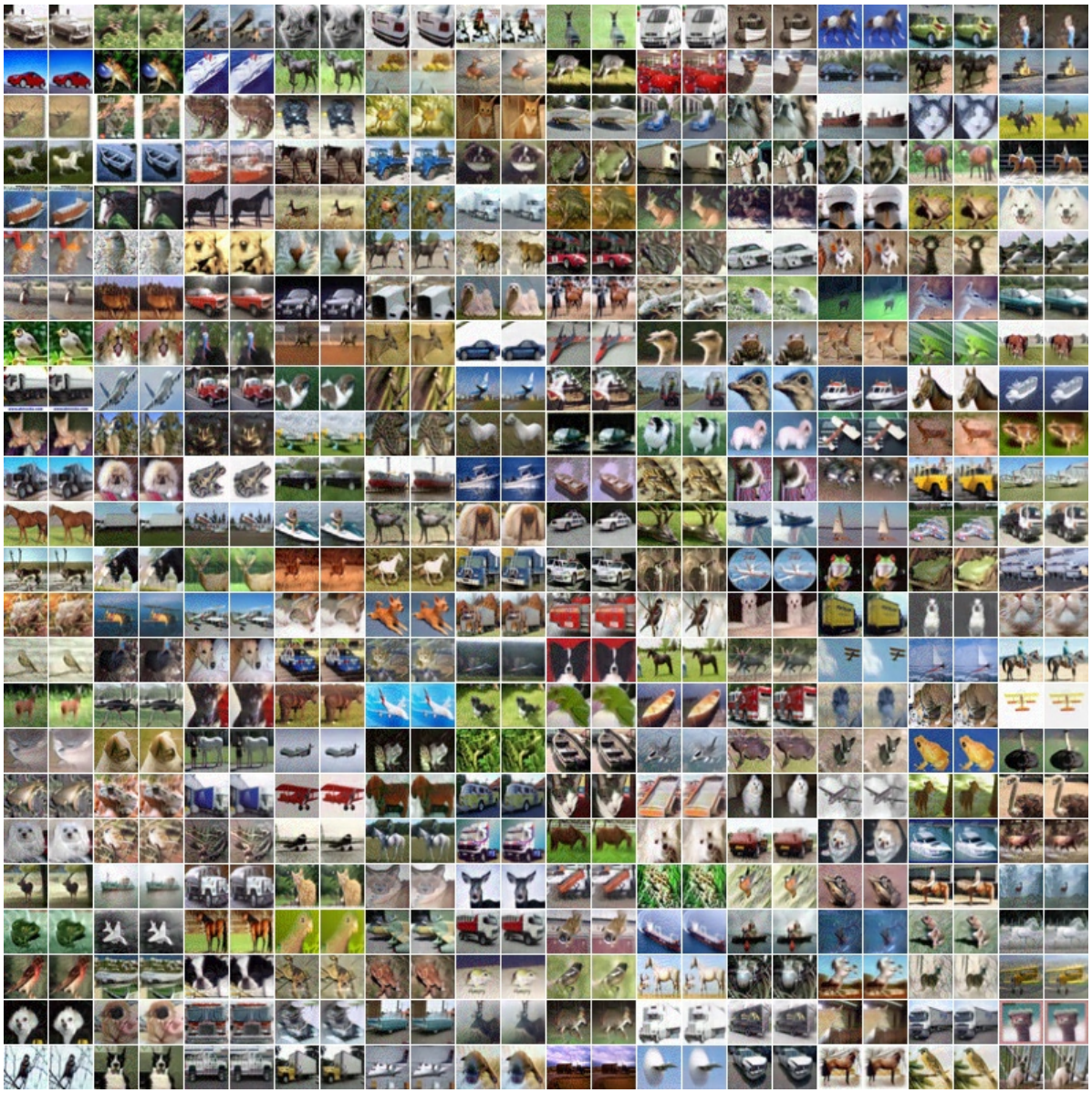


Figure 5: Transformation results for CIFAR-10. For every pair of images, the left is the input adversarial image and the right is the transformed image.

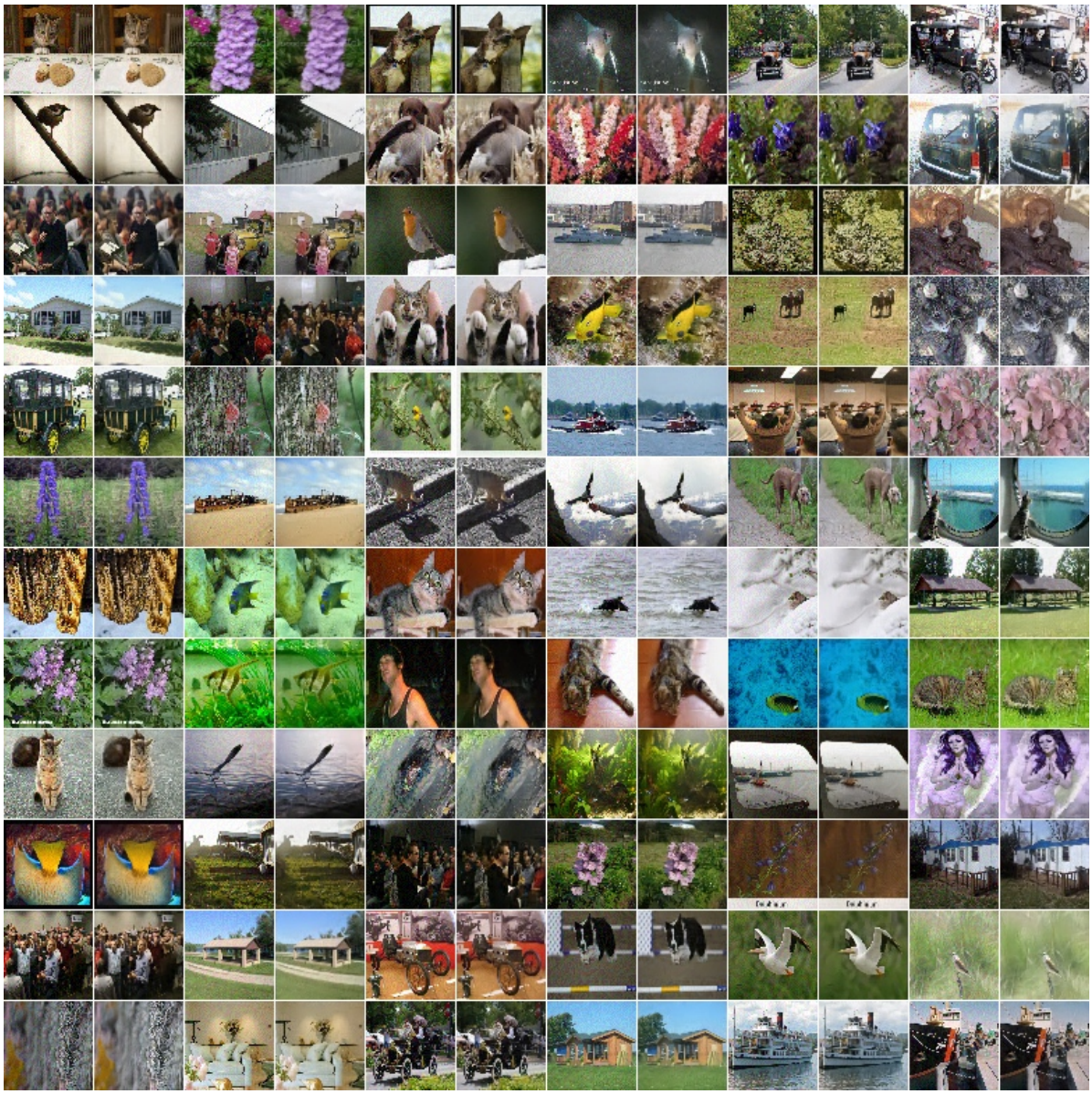


Figure 6: Transformation results for ImageNet-10 at resolution 64. For every pair of images, the left is the input adversarial image and the right is the transformed image.

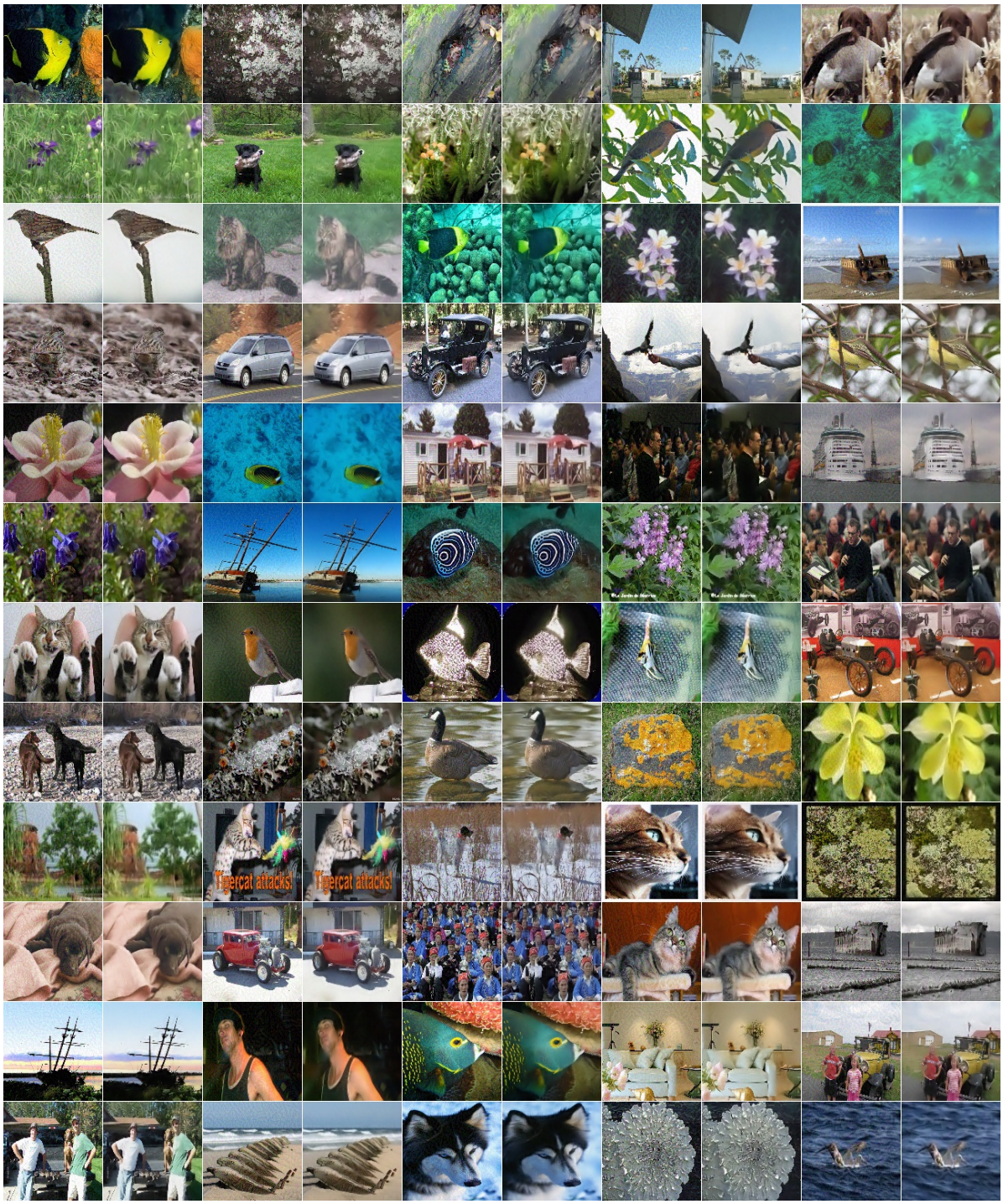


Figure 7: Transformation results for ImageNet-10 at resolution 128. For every pair of images, the left is the input adversarial image and the right is the transformed image.

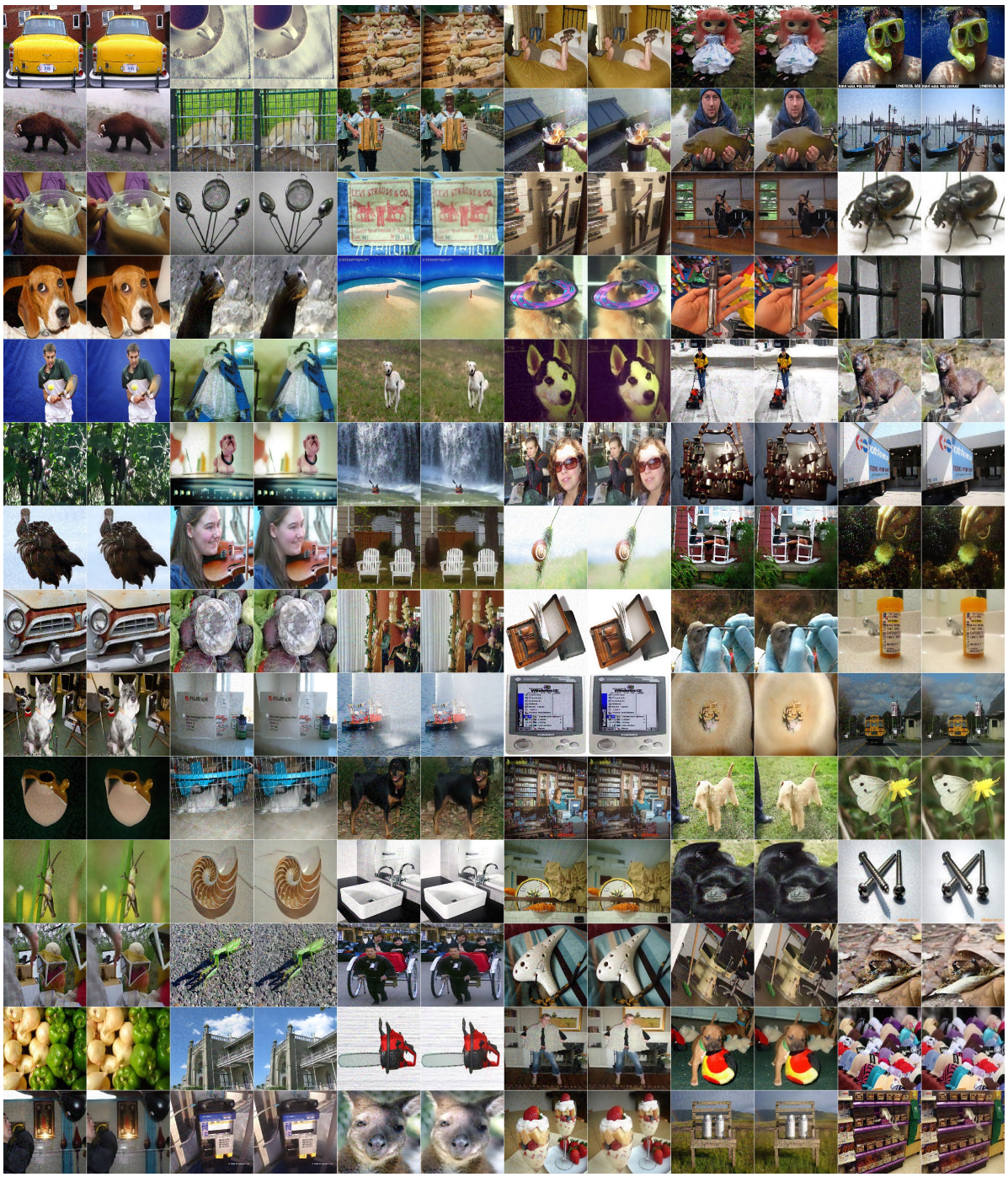


Figure 8: Transformation results for ImageNet-10 at resolution 224. For every pair of images, the left is the input adversarial image and the right is the transformed image.