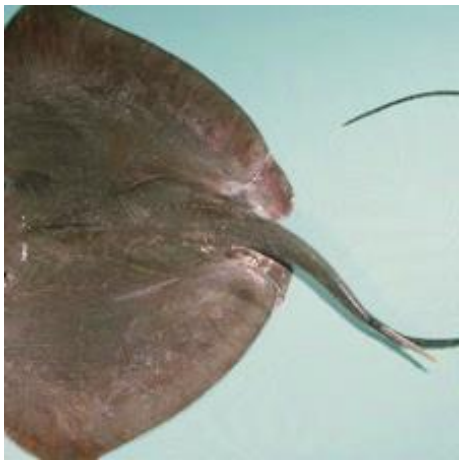
A PROTECTED IMAGENET SAMPLES

African chameleon

black grouse

electric ray



hammerhead

hen

house finch

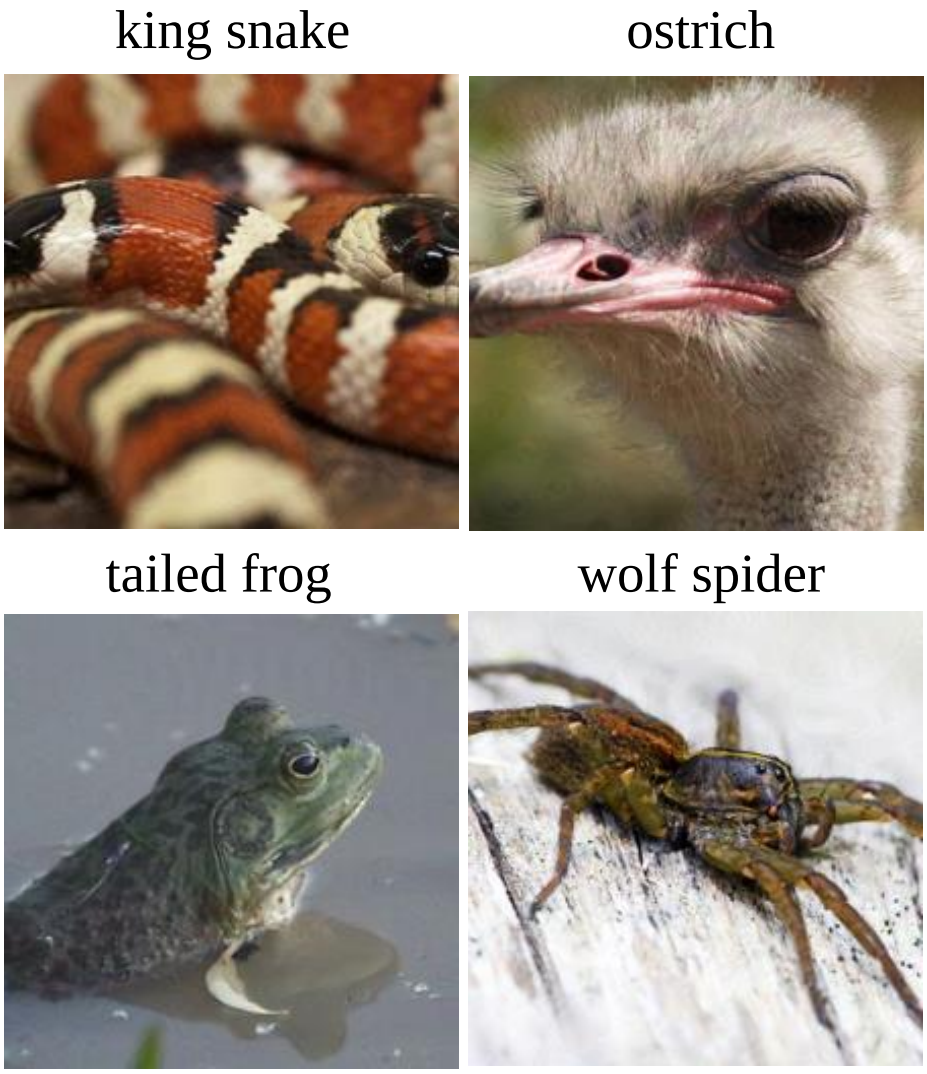


Figure 6: 10-class protected ImageNet samples by our method with $\varepsilon = 2/255$. We select images with clean backgrounds, but the perturbations are still invisible in such cases. However, it would significantly decrease the model's accuracy to below 30%.

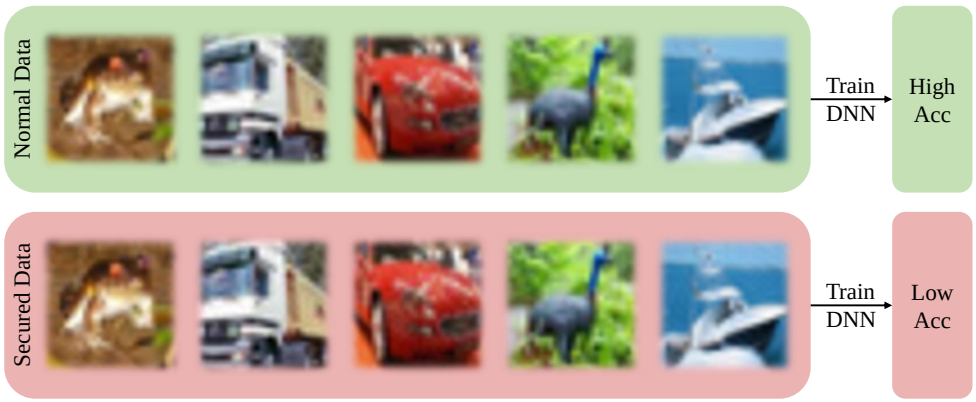


Figure 7: Our protected CIFAR-10 images with $\ell_\infty = 2/255$.

B GRADIENT ANALYSIS OF CIFAR-100 AND IMAGENET CHECKPOINTS

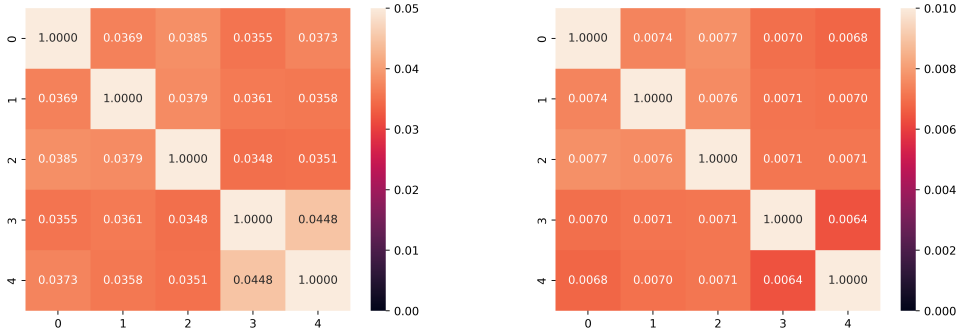


Figure 8: The average absolute cosine value of cross-model gradients for CIFAR-100 (left) and ImageNet (right) checkpoints calculated by 1K samples. The gradients stay nearly orthogonal, indicating good diversity of sub-models in self-ensemble.

C CONFUSION MATRIX OF THE APPROPRIATOR DNN

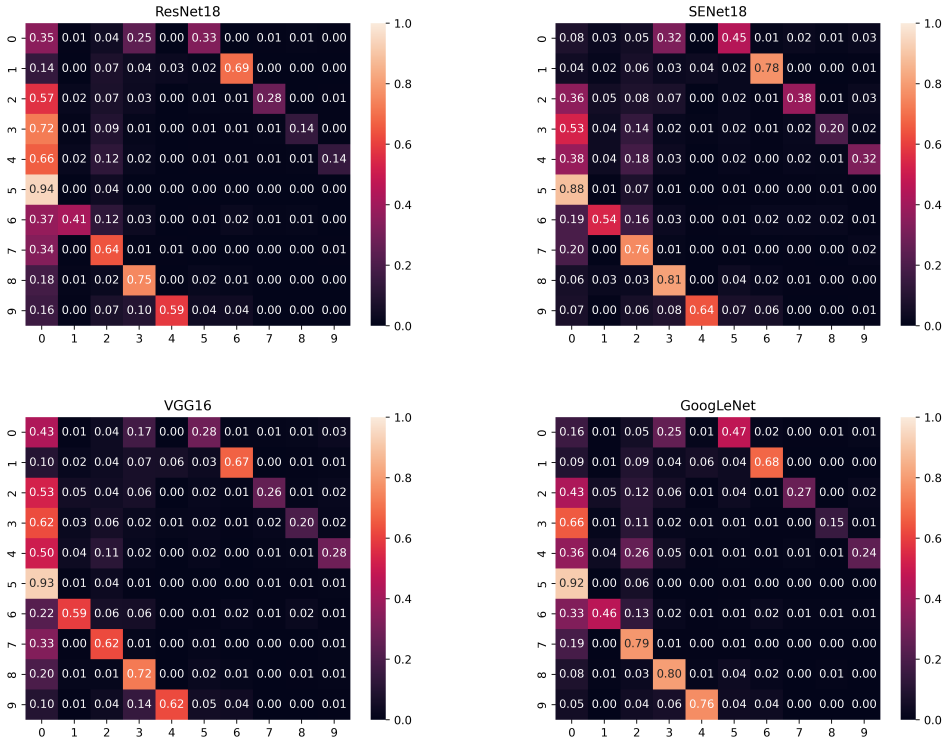


Figure 9: The confusion matrices of DNNs trained on $\varepsilon = 8/255$ CIFAR-10 perturbations by our method. As also shown in (Fowl et al., 2021b), the model tends to predict clean samples to the target class $g(y) = (y + 5)\%10$. We also discover that DNNs prefer to classify them to a fixed class, which is class 0 here, but it may vary during the training. We normalize by the number of samples here.

D OTHER BASELINES AND DEFENSES

We present comparisons with weaker baselines and the overall experimental setups here. Random Noise uses a variance of 8/255. We use the default settings of the baseline papers. TensorClog is with regularization strength of 0.01, an attack optimization rate of 1, and maximum attack iterations of 100. Gradient Alignment is with restarts 8 and optimization steps 240. DeepConfuse is with No. trials of 500, a learning rate of the classification model of 0.01, a batch size of 64, and a learning rate for the noise generator of 1e-4. Unlearnable Example is with PGD step 20, step size 8/2550, stop condition error rate 0.01, and attack iterations 10. Robust Unlearnable Example is with adversarial perturbation radius of REM noise 4/255, sampling number for expectation estimation 5. Adversarial Poison is with PGD step 250, and the step size is 1/255. Autoregressive Poison is crafted by padding the image to 36 with a crop of 4 and using 10 autoregressive processes.

Table 5: ResNet18 testing accuracy trained on CIFAR-10 secured by different methods.

Method ($\epsilon = 8/255$)	Accuracy (↓)
None	94.56
Random Noise	90.52
TensorClog (Shen et al., 2019)	84.24
Gradient Alignment (Fowl et al., 2021a)	53.67
DeepConfuse (Feng et al., 2019)	31.10
Ours	**4.73**

Besides adversarial training, we also investigate whether other training strategies can hurt our data protection method. Strategies include image processing (Gaussian smoothing with kernel size 5, adding random noise with variance $8/255$), data augmentations (mixup, cutmix, and cutout with an alpha 1.0), and privacy protection optimization (DPSGD with a clipping parameter 1.0 and noise parameter 0.005). As in Table 6, the above strategies cannot train a good DNN from protective examples, and we also surpass AdvPoison in this setting.

Table 6: Effectiveness of data protection methods under difference training strategies.

Training Strategy ($\epsilon = 8/255$)	AdvPoison	Ours
None	6.25	**4.73** (-1.52)
Gaussian Smoothing	11.94	**9.95** (-1.99)
Adding Random Noise	6.55	**2.89** (-3.66)
Mixup (Zhang et al., 2018)	15.86	**10.06** (-5.80)
Cutmix (Yun et al., 2019)	10.09	**4.47** (-5.62)
Cutout (DeVries & Taylor, 2017)	8.11	**5.51** (-2.60)
DPSGD (Abadi et al., 2016)	24.61	**4.60** (-20.01)