Supplementary Material

Here we present the **Supplementary Material** for the paper *Enhancing Adversarial Robustness via Test-time Transformation Ensembling*. In this document, we report comprehensive results for ablations, qualitative results, pseudo-code for the transforms we used, plots for the complete outcomes of gradient-obfuscation experiments, and results for the removal of another common transform from training.

A. Detailed Transforms Ablation

In Tables 2 and 3 we reported our best results for the transforms with which TTE is instantiated. However, Table 4 only reports results in the case of TRADES on CIFAR10. Here, for completeness, we show the analogous results for the rest of the methods reported in Table 2 and for a large set of transforms.

We report CIFAR10-only methods (HYDRA, MART and Gowal *et al.*) in Tables 9, 10 and 8. We report methods with results both on CIFAR10 and CIFAR100 (AWP, ATES, and IN-Pret) in Tables 11, 12, 13, 14, 15, and 16.

Table 8. **Adversarial robustness gains of various transforms on CIFAR10.** We test the impact in adversarial robustness of introducing various transforms to TTE on the method of Gowal *et al.* We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and Gowal *et al.* Robust accuracies larger than that of Gowal *et al.* are shown in **boldface**.

Method	Clean	Robust	Diff.
Gowal *et al.*	89.48	63.26	-
+ flip	89.41	**64.37**	+1.09
+ 1 crop	89.39	**63.52**	+0.26
+ 2 crops	89.04	63.20	-0.06
+ 3 crops	89.25	**63.77**	+0.51
+ 4 crops	89.17	63.22	-0.04
+ flip + 1 crop	89.43	**64.35**	+1.09
+ flip + 2 crops	89.16	**64.12**	+0.86
+ flip + 3 crops	89.40	**64.39**	+1.13
+ flip + 4 crops	89.18	**63.95**	+0.69
+ flip + 1 crop + 1 flipped-crop	89.49	**64.40**	+1.14
+ flip + 2 crops + 2 flipped-crops	89.05	**64.20**	+0.94
+ flip + 3 crops + 3 flipped-crops	89.41	**64.55**	+1.29
+ flip + 4 crops + 4 flipped-crops	89.21	**64.29**	+1.03

B. Detailed Algorithms for Transforms

Algorithm 2 reports the pseudo-code, detailing our implementation of the transforms used in our TTE wrapper.

C. Varying attack strength and iterations

In Section 4.6 we conducted experiments on *APGD-T* by varying the number of optimization iterations and the attack strength (ϵ) to check for signs of gradient obfuscation. We reported accuracy under attack at important milestones of each of these ablations. Here we report accuracies for *all* possible values in Figure 4. These results are consistent with the hypothesis that introducing TTE does not induce gradient obfuscation, as stated in the paper.

Table 9. **Adversarial robustness gains of various transforms on CIFAR10.** We test the impact in adversarial robustness of introducing various transforms to TTE on HYDRA. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and HYDRA. Robust accuracies larger than that of HYDRA are shown in **boldface**.

Method	Clean	Robust	Diff.
HYDRA	88.98	57.64	-
+ flip	89.1	**59.81**	+2.17
+ 1 crop	88.86	**58.36**	+0.72
+ 2 crops	88.58	**58.01**	+0.37
+ 3 crops	88.92	**58.61**	+0.97
+ 4 crops	88.59	**58.2**	+0.56
+ flip + 1 crop	89.00	**60.01**	+2.37
+ flip + 2 crops	88.87	**59.59**	+1.95
+ flip + 3 crops	88.96	**59.65**	+2.01
+ flip + 4 crops	88.64	**59.12**	+1.48
+ flip + 1 crop + 1 flipped-crop	88.89	**60.28**	+2.64
+ flip + 2 crops + 2 flipped-crops	88.81	**60.10**	+2.46
+ flip + 3 crops + 3 flipped-crops	88.82	**60.38**	+2.74
+ flip + 4 crops + 4 flipped-crops	88.70	**60.08**	+2.44

Table 10. **Adversarial robustness gains of various transforms on CIFAR10.** We test the impact in adversarial robustness of introducing various transforms to TTE on MART. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and MART. Robust accuracies larger than that of MART are shown in **boldface**.

Method	Clean	Robust	Diff.
MART	87.5	56.75	-
+ flip	87.74	**58.38**	+1.63
+ 1 crop	87.55	**57.27**	+0.52
+ 2 crops	87.11	**57.16**	+0.41
+ 3 crops	87.45	**57.66**	+0.91
+ 4 crops	87.31	**57.4**	+0.65
+ flip + 1 crop	87.66	**58.53**	+1.78
+ flip + 2 crops	87.54	**58.16**	+1.41
+ flip + 3 crops	87.69	**58.42**	+1.67
+ flip + 4 crops	87.58	**58.11**	+1.36
+ flip + 1 crop + 1 flipped-crop	87.76	**58.83**	+2.08
+ flip + 2 crops + 2 flipped-crops	87.61	**58.87**	+2.12
+ flip + 3 crops + 3 flipped-crops	87.79	**58.94**	+2.19
+ flip + 4 crops + 4 flipped-crops	87.61	**58.92**	+2.17

D. Removing flip transform from training

Following a similar spirit to Section 4.7, we *remove* the usual flipping transform from the official TRADES training routine, and refer to this model as TRADES^{nf}. We record the adversarial robustness of TRADES^{nf} both when tested

Table 11. **Adversarial robustness gains of various transforms on CIFAR10.** We test the impact in adversarial robustness of introducing various transforms to TTE on AWP. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and AWP. Robust accuracies larger than that of AWP are shown in **boldface**.

Method	Clean	Robust	Diff.
AWP	88.25	60.53	-
+ flip	88.2	**61.54**	+1.01
+ 1 crop	88.08	**60.82**	+0.29
+ 2 crops	87.76	60.42	-0.11
+ 3 crops	88.04	**60.99**	+0.46
+ 4 crops	87.87	**60.79**	+0.26
+ flip + 1 crop	88.28	**61.71**	+1.18
+ flip + 2 crops	87.92	**61.35**	+0.82
+ flip + 3 crops	88.08	**61.6**	+1.07
+ flip + 4 crops	88.03	**61.29**	+0.76
+ flip + 1 crop + 1 flipped-crop	88.23	**61.68**	+1.15
+ flip + 2 crops + 2 flipped-crops	88.06	**61.54**	+1.01
+ flip + 3 crops + 3 flipped-crops	88.07	**61.99**	+1.46
+ flip + 4 crops + 4 flipped-crops	87.98	**61.62**	+1.09

Table 12. **Adversarial robustness gains of various transforms on CIFAR100.** We test the impact in adversarial robustness of introducing various transforms to TTE on AWP. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and AWP. Robust accuracies larger than that of AWP are shown in **boldface**.

Method	Clean	Robust	Diff.
AWP	60.38	28.86	-
+flip	60.27	**29.66**	+0.80
+1 crop	60.48	**29.31**	+0.45
+2 crops	59.96	**29.29**	+0.43
+3 crops	60.41	**29.53**	+0.67
+4 crops	60.3	**29.41**	+0.55
+flip + 1 crop	60.36	**29.79**	+0.93
+flip + 2 crop	60.26	**29.70**	+0.84
+flip + 3 crop	60.43	**29.80**	+0.94
+flip + 4 crop	60.44	**29.74**	+0.88
+flip + 1 crop + 1 flipped-crop	60.28	**29.86**	+1.00
+flip + 2 crop + 2 flipped-crop	60.22	**29.94**	+1.08
+flip + 3 crop + 3 flipped-crop	60.39	**30.01**	+1.15
+flip + 4 crop + 4 flipped-crop	60.13	**29.78**	+0.92

Table 13. **Adversarial robustness gains of various transforms on CIFAR10.** We test the impact in adversarial robustness of introducing various transforms to TTE on ATES. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and ATES. Robust accuracies larger than that of ATES are shown in **boldface**.

Method	Clean	Robust	Diff.
ATES	86.84	51.46	-
+ flip	86.96	**53.11**	+1.65
+ 1 crop	86.86	**52.08**	+0.62
+ 2 crops	86.68	**52.37**	+0.91
+ 3 crops	86.86	**52.59**	+1.13
+ 4 crops	86.62	**52.31**	+0.85
+ flip + 1 crop	86.96	**53.26**	+1.80
+ flip + 2 crops	86.89	**53.54**	+2.08
+ flip + 3 crops	87.03	**53.46**	+2.00
+ flip + 4 crops	86.82	**53.25**	+1.79
+ flip + 1 crop + 1 flipped-crop	87.08	**53.71**	+2.25
+ flip + 2 crops + 2 flipped-crops	86.95	**53.94**	+2.48
+ flip + 3 crops + 3 flipped-crops	87.03	**54.05**	+2.59
+ flip + 4 crops + 4 flipped-crops	86.86	**54.17**	+2.71

Table 14. **Adversarial robustness gains of various transforms on CIFAR100.** We test the impact in adversarial robustness of introducing various transforms to TTE on ATES. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and ATES. Robust accuracies larger than that of ATES are shown in **boldface**.

Method	Clean	Robust	Diff.
ATES	62.82	24.96	-
+flip	63.11	**26.27**	+1.31
+1 crop	62.88	**25.77**	+0.81
+2 crops	62.70	**26.14**	+1.18
+3 crops	63.12	**26.07**	+1.11
+4 crops	62.88	**25.75**	+0.79
+flip + 1 crop	63.27	**26.45**	+1.49
+flip + 2 crop	62.70	**26.14**	+1.18
+flip + 3 crop	63.20	**26.61**	+1.65
+flip + 4 crop	63.21	**26.43**	+1.47
+flip + 1 crop + 1 flipped-crop	63.17	**26.72**	+1.76
+flip + 2 crop + 2 flipped-crop	62.97	**27.04**	+2.08
+flip + 3 crop + 3 flipped-crop	63.47	**26.79**	+1.83
+flip + 4 crop + 4 flipped-crop	63.24	**27.09**	+2.13

on *(i)* clean images and *(ii)* on both the original image and its flipped version. Table 17 reports our results. From this table we note that *(i)* the training-time flip transform is essential for TRADES: clean and robust accuracies drop, approximately, by 3% and 7%, respectively; *(ii)* even with TRADES^{nf}, adding a flipped version of the image is beneficial for adversarial robustness: clean and robust accuracies increase, approximately, by 1% and 4%, respectively. These results suggest that, for a TRADES model, there is little distribution shift between the original images and their flipped versions, as expected. Thus, adding a flipped version of the image at test-time does not induce vulnerabilities into the defense.

E. Visualization

To exemplify what the model receives as input when equipped with the TTE wrapper, we illustrate the four most complete cases from our ablations in Section A. That is, when the model is fed with flipped, cropped and flipped-cropped versions of the image, in addition to the original image. These cases are depicted in Figure X.

We also display some adversarial examples that fool the FD model (on ImageNet) with and without the wrapper en-

Table 15. **Adversarial robustness gains of various transforms on CIFAR10.** We test the impact in adversarial robustness of introducing various transforms to TTE on IN-Pret. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and IN-Pret. Robust accuracies larger than that of IN-Pret are shown in **boldface**.

Method	Clean	Robust	Diff.
IN-Pret	87.11	55.31	-
+ flip	87.06	**55.66**	+0.35
+ 1 crop	87.23	**56.20**	+0.89
+ 2 crops	86.74	**55.34**	+0.03
+ 3 crops	86.96	**55.67**	+0.36
+ 4 crops	86.71	55.06	-0.25
+ flip + 1 crop	87.22	**56.45**	+1.14
+ flip + 2 crops	86.85	**56.11**	+0.80
+ flip + 3 crops	87.06	**56.24**	+0.93
+ flip + 4 crops	86.84	**55.87**	+0.56
+ flip + 1 crop + 1 flipped-crop	87.13	**56.43**	+1.12
+ flip + 2 crops + 2 flipped-crops	86.93	**56.49**	+1.18
+ flip + 3 crops + 3 flipped-crops	87.17	**56.50**	+1.19
+ flip + 4 crops + 4 flipped-crops	86.85	**56.41**	+1.10

Table 16. **Adversarial robustness gains of various transforms on CIFAR100.** We test the impact in adversarial robustness of introducing various transforms to TTE on IN-Pret. We report clean and robust accuracies, and the difference in robustness between each TTE-enhanced model and IN-Pret. Robust accuracies larger than that of IN-Pret are shown in **boldface**.

Method	Clean	Robust	Diff.
IN-Pret	59.37	28.96	-
+flip	59.52	**29.40**	+0.44
+1 crop	58.96	**29.02**	+0.06
+2 crops	58.64	28.67	-0.29
+3 crops	59.15	**29.10**	+0.14
+4 crops	58.61	28.67	-0.29
+flip + 1 crop	59.39	**29.46**	+0.50
+flip + 2 crop	58.94	**29.24**	+0.28
+flip + 3 crop	59.10	**29.39**	+0.43
+flip + 4 crop	58.87	**29.12**	+0.16
+flip + 1 crop + 1 flipped-crop	59.38	**29.50**	+0.54
+flip + 2 crop + 2 flipped-crop	58.75	**29.32**	+0.36
+flip + 3 crop + 3 flipped-crop	59.16	**29.61**	+0.65
+flip + 4 crop + 4 flipped-crop	58.93	**29.68**	+0.72

Table 17. **Robustness of TRADES trained *without* the flipping transformation.** We train a TRADES model without the flipping transformation (TRADES^{nf}). We test the model's adversarial robustness when tested on *(i)* clean images and *(ii)* on both the clean image and its flipped version. Results show that, even when the model was not trained on flipped images, introducing a flipped version of the image is still beneficial for adversarial robustness.

Method	Clean	Robust	Diff.
TRADES	84.92	53.11	-
TRADES^{nf}	81.89	46.29	-6.82
TRADES^{nf} + flip	82.82	50.19	-2.92

Algorithm 2 Differentiable transforms pseudocode in PyTorch style.

```
class PadCrop:
   def __init__(self, o_x, o_y, crop_size, pad_size):
      self.pad_size = pad_size
      # starting points
      self.o_x = o_x
      self.o_y = o_y
      # ending points
      self.e_x = o_x + crop_size
      self.e_y = o_y + crop_size

   def forward(self, x):
      # pad input
      x = pad(x, pad=self.pad_size)
      # crop
      x = x[:,:,self.o_x:self.e_x,self.o_y:self.e_y]
      return x

class Flip:
   def forward(self, x):
      return x.flip(3) # the left-right dimension

class FlipPadCrop:
   def __init__(self, o_x, o_y, crop_size, pad_size):
      self.flip = Flip()
      self.pad_crop = PadCrop(o_x, o_y, crop_size,
          pad_size)

   def forward(self, x):
      return self.flip(self.pad_crop(x))
```

`pad`: zero padding.

hancement. We display the adversarial examples and the noise introduced into the original image. For the baseline model, we extract the 224×224 center crop from the 256×256 crop. For this reason, there is a random-like padding as the initial perturbation is initialized from random noise. The adversarial examples are shown in Figures 6 through 25.

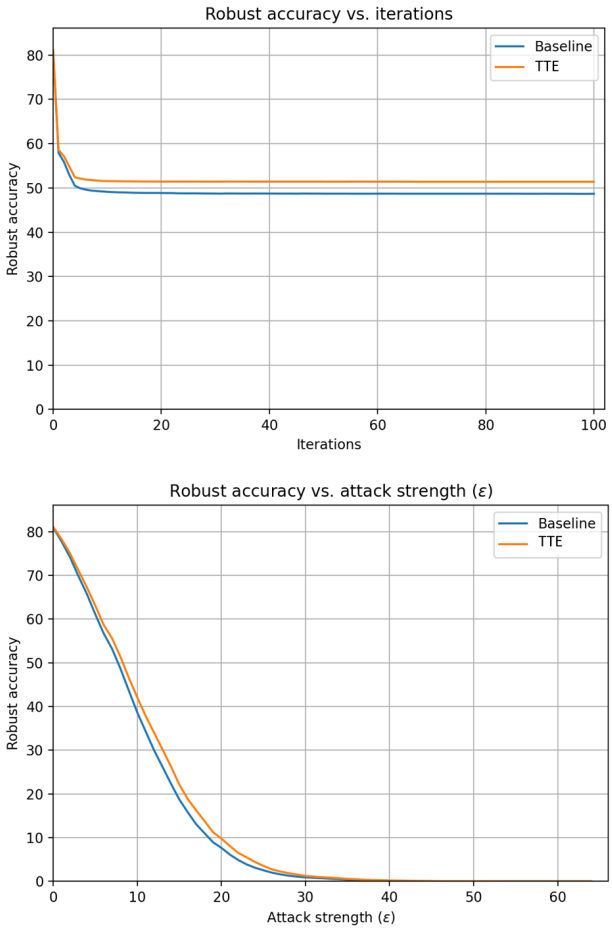

Figure 4. **Accuracy under APGD-T attacks *vs.* optimizations iterations and attack strength.** We report accuracy plots for the entire set of values considerd in the experiments from Section 4.6.

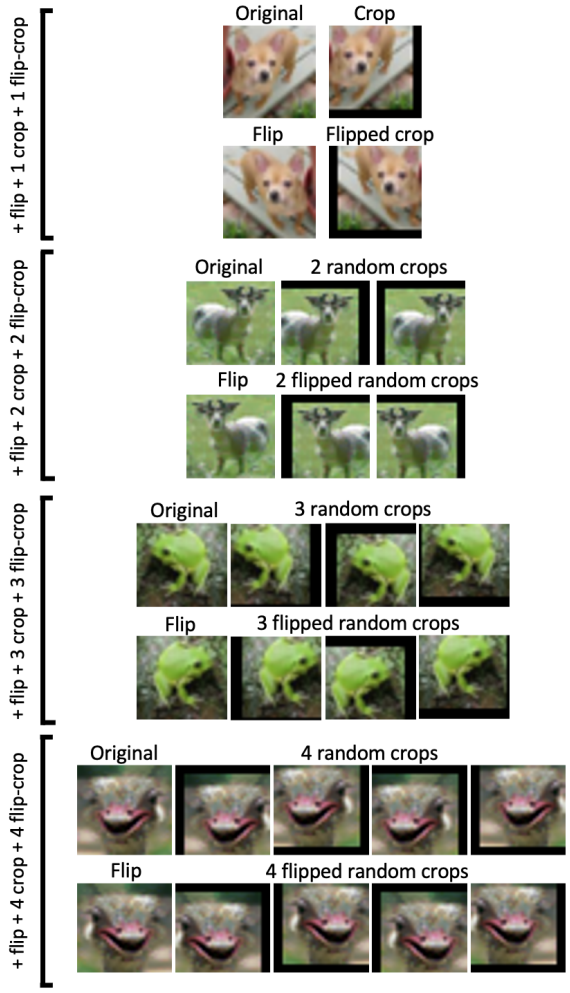

Figure 5. **Visualization of what an TTE-enhanced model receives as input.** We show what an input looks like to a model that has been enhanced with TTE. Here we exemplify what several sets of transforms result in.

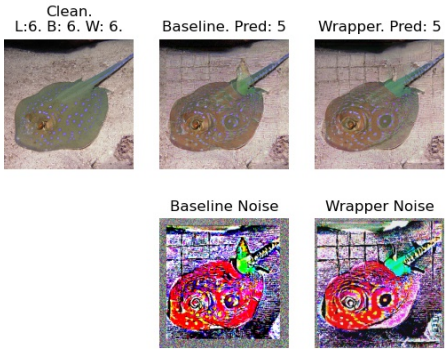

Figure 6. Original Image is labeled as *Stingray*. The adversaries are predicted as *Electric Rays*. Some electric rays are characteristic of having circular patterns like the ones induced by the adversarial noise.

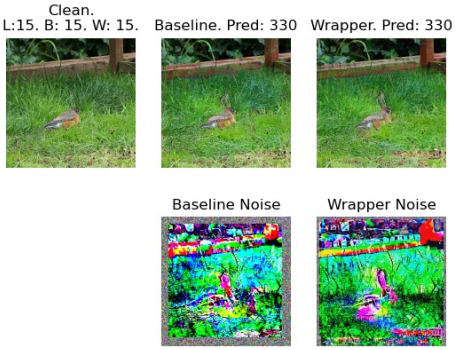

Figure 7. Original Image is labeled as *Robin*. The adversaries are predicted as *Wood Rabbits*. The adversarial example's noise from the wrapper-enhanced model clearly visualizes the ear from a bunny while the FD model does not exhibit this pattern.

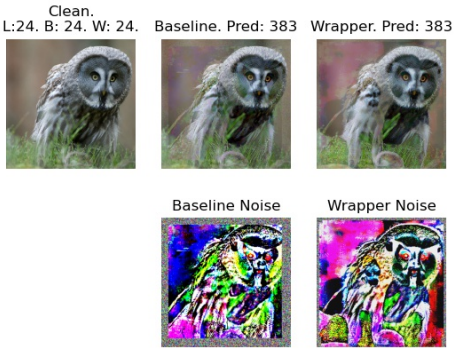

Figure 8. Original Image is labeled as *Great Grey Owl*. The adversaries are predicted as *Madagascar Cat*. The adversarial example's noise from the wrapper-enhanced model clearly visualizes the face from this animal while the FD model does not exhibit this pattern.

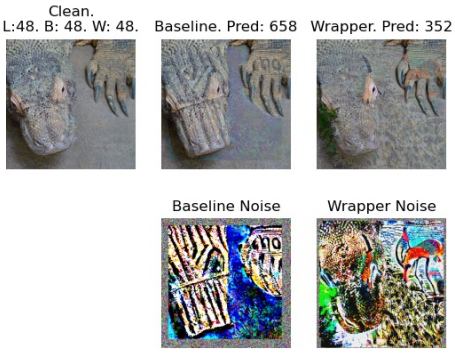

Figure 9. Original Image is labeled as *Komodo Dragon*. The baseline model classifies its adversary as a *Mitten*. The adversary of the wrapper, classified as *Impala*, needed to modify all the image to fool the network.

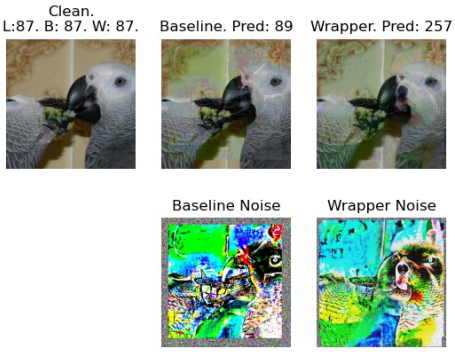

Figure 10. Original Image is labeled as *African Grey Parrot*. The baseline model classifies its adversary as *Sulphur-Crested Cockatoo*. The wrapper-enhanced version classifies its adversary as a *Great Pyrenees*. The noise clearly displays the nose of this dog breed.

Figure 11. Original Image is labeled as *Italian Greyhound*. Both models classify their adversarial examples as *Polecat*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.

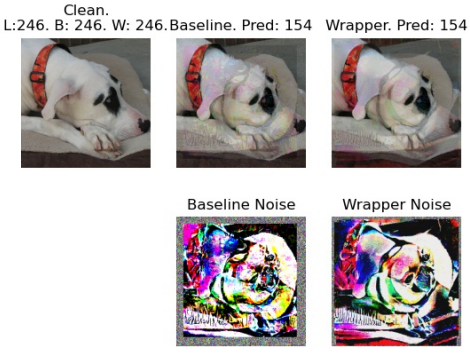

Figure 12. Original Image is labeled as *Great Dane*. Both models classify their adversarial examples as *Pekinese*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.

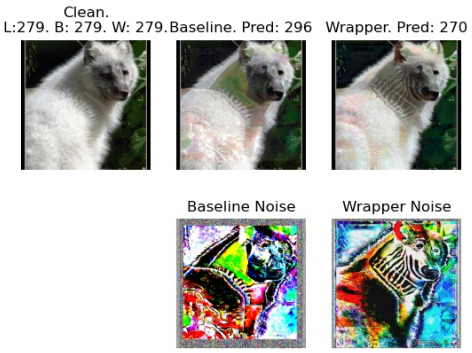

Figure 13. Original Image is labeled as *Artic Fox*. The baseline model classify its adversarial example as *Ice Bear* while the wrapper-enhanced version classifies his example as *White Wolf*.

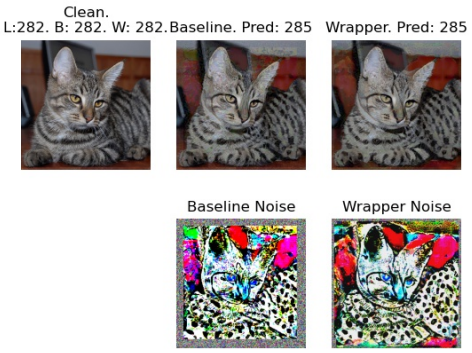

Figure 14. Original Image is labeled as *Tiger Cat*. Both models classify their adversarial examples as *Egyptian Cat*. The Egyptian cat is characterized by its dot marks.

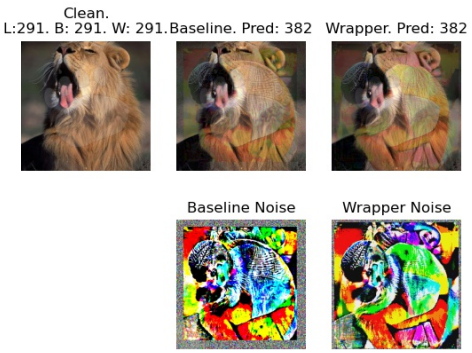

Figure 15. Original Image is labeled as *Lion*. Both models classify their adversarial examples as *Squirrel Monkey*. The shape of the monkey is clearly seen on the noise of both adversaries.

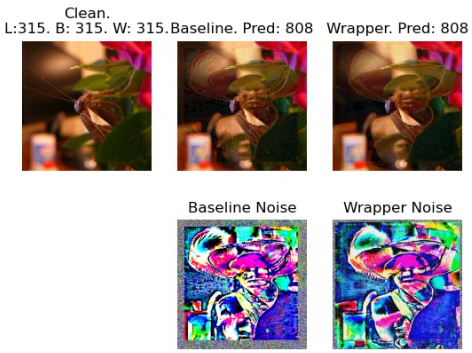

Figure 16. Original Image is labeled as *Mantis*. Both models classify their adversarial examples as *Sombrero*. Both adversaries display face-like attributes.

Figure 17. Original Image is labeled as *Marmot*. Both models classify their adversarial examples as *Macaque*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.

Figure 18. Original Image is labeled as *Arabian Camel*. Both models classify their adversarial examples as *Impala*. The wrapper-enhanced version exhibits clearer horns on the noise.

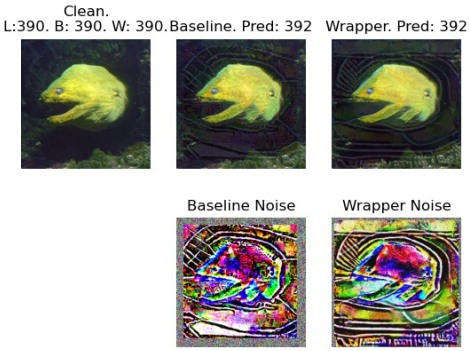

Figure 19. Original Image is labeled as *Eel*. Both models classify their adversarial examples as *Rock Beauty*. The wrapper-enhanced adversarial version exhibits clearer patterns.

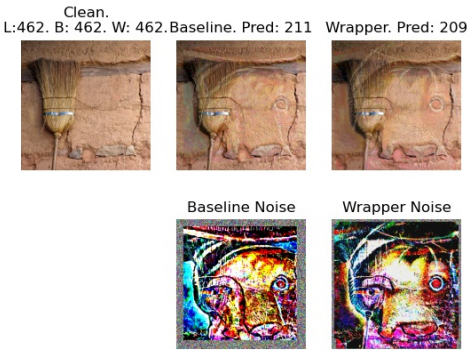

Figure 20. Original Image is labeled as *Broom*. Both models classify their adversarial examples as dog breeds: *Vizsla* for the baseline and *Chesapeake Bay Retriever* for the wrapper version. The noise from both adversaries expose different shapes of the nose.

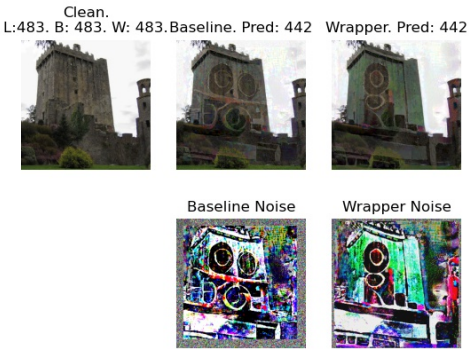

Figure 21. Original Image is labeled as *Castle*. Both models classify their adversarial examples as *Bell Cote*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.

Figure 22. Original Image is labeled as *Container Ship*. Both models classify their adversarial examples as *Pirate*.

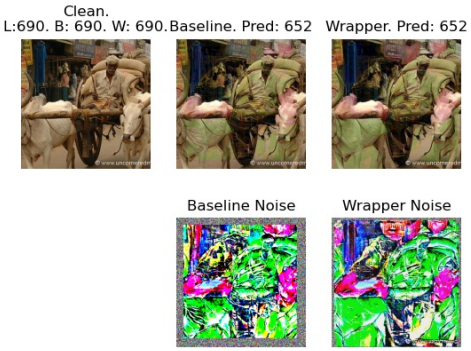

Figure 23. Original Image is labeled as *Oxcart*. Both models classify their adversarial examples as *Military Uniform*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.

Figure 24. Original Image is labeled as *Slide Rule*. Both models classify their adversarial examples as *Magnetic Compass*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.

Figure 25. Original Image is labeled as *Wardrobe*. Both models classify their adversarial examples as *Shower Curtain*. The noise from the wrapper-enhanced version clearly displays less blurry patterns.