Supplementary material for the paper titled: Combining Input Transformation and Noisy Training

In this document, we provide a visual comparison of the images used by our proposed defense algorithm. In the implementation, we trained five models using different qualities of images. Figure 1 gives an example, to show the training and test flow of one single model in our model set. To generate training examples, we firstly employ our proposed compression algorithm to compress the image at different levels of qualities, and then the corresponding compressed images are injected Gaussian noise (with $\mu = 0, \delta = 2$ and clipping from -4 to 4). Figures 2 and 3 list the resultant images after compression transformation and noise injection. The leftmost column presents the images compressed at five levels of quality, i.e., original image, level of 90, level of 70, level of 50, level of 30. The second column shows the results of compressed images after injecting with Gaussian noise. For comparison, the rightmost column illustrates the adversarial example generated by the attack algorithm "FGSM " on the compressed images in the first column with $\epsilon = 4$. As we can see, our Gaussian noised images have very similar visual results with the adversarial examples.

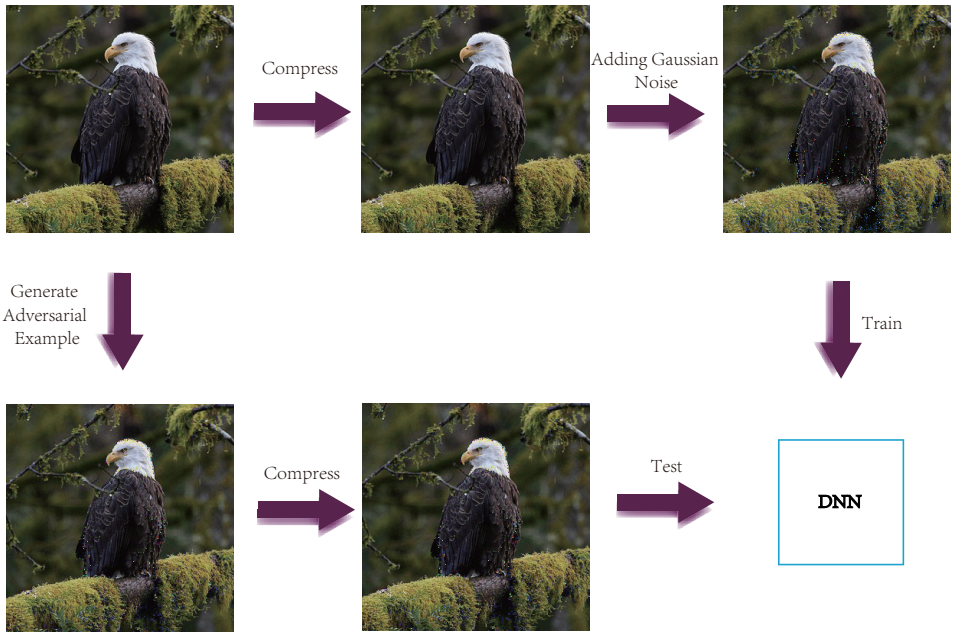

Figure 1: Training and testing flow of a single model.

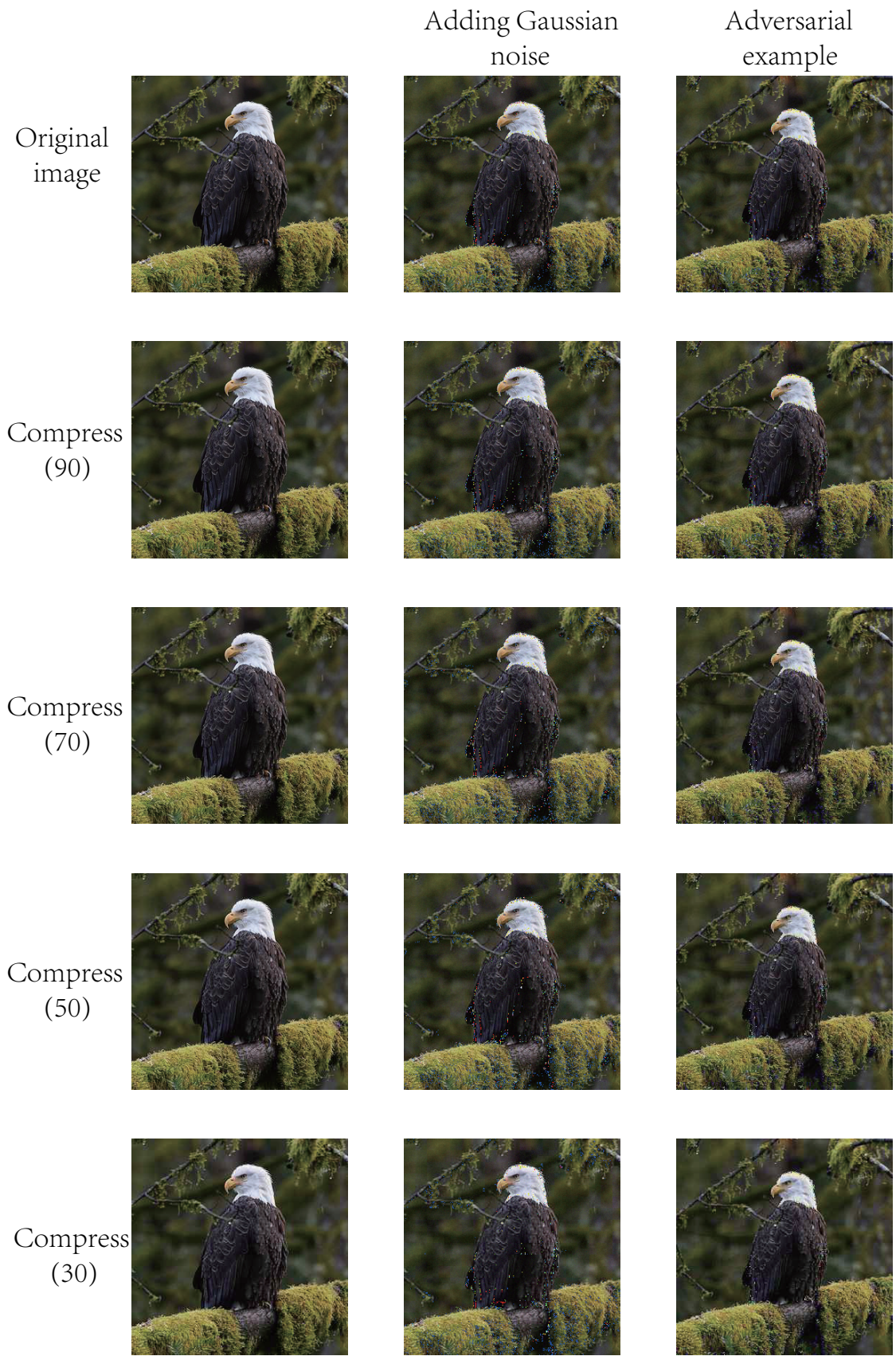

Figure 2: Visual results of images generated during our algorithm. Images in each row are compressed at certain quality. Our model set are retrained using images in mid column. The right column are adversarial images.

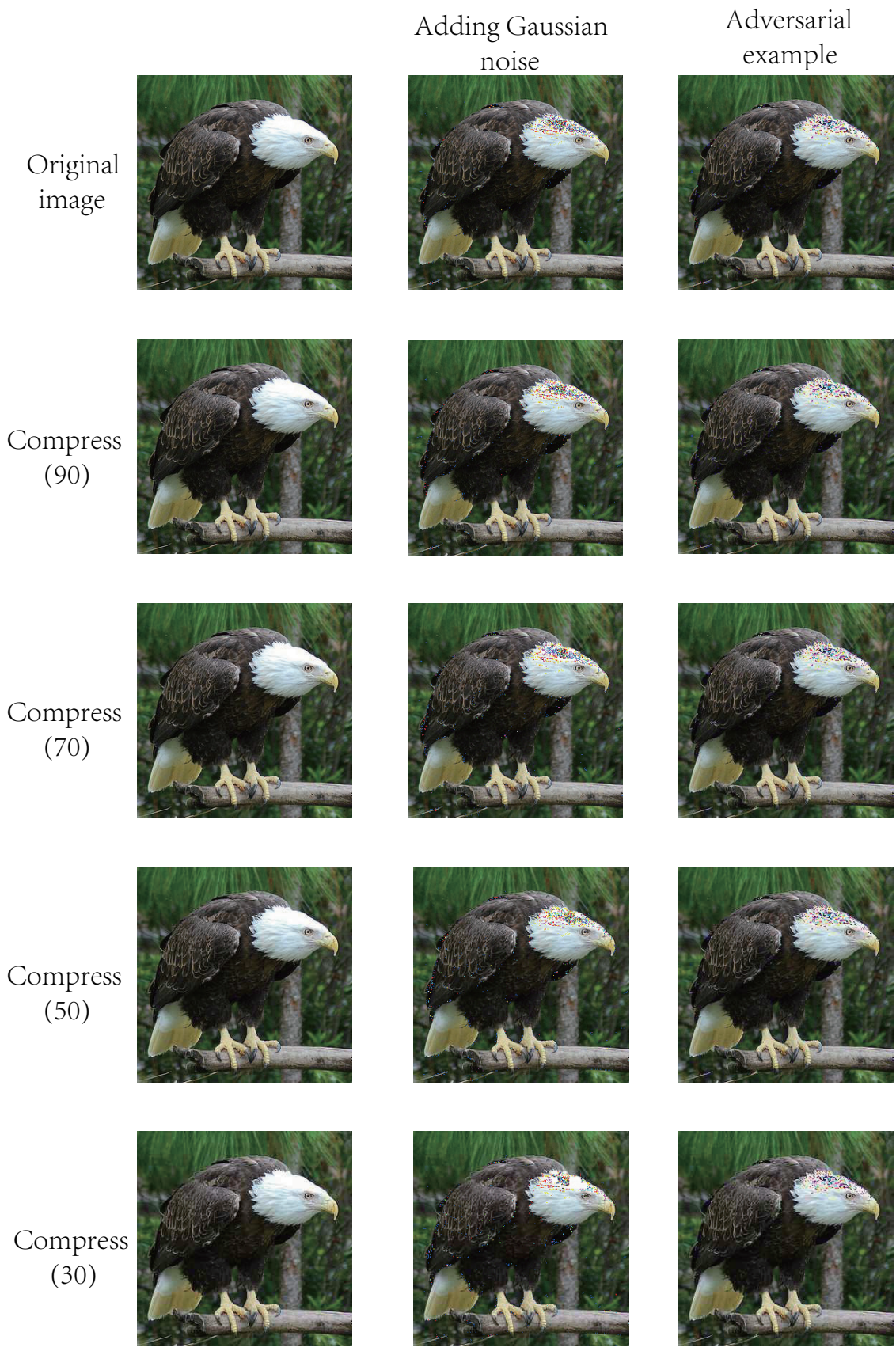

Figure 3: Visual results of images generated during our algorithm. Images in each row are compressed at certain quality. Our model set are retrained using images in mid column. The right column are adversarial images.