Efficient Physical Image Attacks Using Adversarial Fast Autoaugmentation Methods (Supplementary Materials)

I. Additional Experimental Results

Here, we will introduce some of our experimental results about our proposed methods.

A. *Video visualization results of physical attacks*

We conducted a video analysis by capturing footage of three printed images: the original image, the adversarial image crafted by our previous work, and the adversarial image crafted by our proposed method. These images were subjected to classification using an inception-v3 classifier. The video file ”compare.mp4” illustrates the results of this analysis, where the left side of the video displays the recorded footage, and the right side shows the classification results for each frame of the video. The classification results depict the top 10 most likely labels identified by each frame of the image, with the leftmost column representing the highest probability of recognition. Upon analysis, we observed that traditional classifiers tend to be robust in real-world environments, remaining unaffected by noise. However, when the angle of the original image is rotated, we can observe that the original category (represented by the green column, labeled ”bee eater”) struggles to maintain its highest level of probability, and in some cases, even drops out of the top 10 probabilities. For our previous work, we observed that our adversarial examples can maintain stable classification results under most environmental conditions, except for the case of cropping. However, when we replaced the image with our adversarial example, we observed that the attack effect (target label indicated by the red column, labeled ”dragonfly”) of our image was resistant to all environmental states, consistently maintaining high confidence throughout. It is important to note that our adversarial image is crafted by the combination for the AFA based Most-likely ensemble Method (AFA-MLEM) and weighted objective function-based Ensemble model Method (WOFM). This means that our adversarial image can mislead multiple classifiers at the same time. When we used WOFM, we only trained the Inception-v3 and ResNet-v2-101 classifiers, but, the image can also mislead the two classifiers of Inception-v4 and Inception-Resnet-v2. We have found that for classifiers that are not trained, our attack effect can still be maintained to a certain extent. We also did a video about the study. The vedio **”compare.mp4”** shows that our adversarial attack has a certain degree of transferability between classifiers, and we believe that this finding will also be helpful to our future research on this aspect.

It is worth emphasizing that our adversarial image is crafted using a combination of the Adversarial Fast Autoaugmentation based Most-likely Ensemble Method (AFA-MLEM) and the weighted objective function-based Ensemble Model Method (WOFM). This means that our adversarial image has the capability to deceive multiple classifiers simultaneously. Although we only trained the Inception-v3 and ResNet-v2-101 classifiers using WOFM, our adversarial image can still mislead the Inception-v4 and Inception-Resnet-v2 classifiers, which were not included in the training process. This demonstrates the transferability of our attack effect to some extent, even for classifiers that were not specifically trained for our method. To further prove this transferability, we have also produced a video titled **”ensemble-model-attack.mp4.”** This video showcases our adversarial attack and illustrates the degree of transferability between classifiers. We believe that this finding will be valuable for our future research in this field and may contribute to further advancements in adversarial attack techniques.