

Supplementary Materials: Adversarial Example Quality Assessment: A Large-scale Dataset and Strong Baseline

Anonymous Authors

In this supplementary file, we present more details and experimental results that are not included in our main paper due to the 8-page limitation. The outline is as follows:

- Sec. A: Implementation details of our approach.
- Sec. B: More analysis on the correlations between attack intensity and JS-Distance.
- Sec. C: More qualitative comparison of our AdvDSS and existing IQA methods for evaluating adversarial examples.
- Sec. D: More details and results of the experiments on generating adversarial examples using AEQA.

A. IMPLEMENTATION DETAILS

Our method is implemented with Python on an NVIDIA RTX 3060Ti. In order to measure the attack intensity of adversarial examples, we use the backbone model i.e., Inception v3 network pre-trained on ImageNet, which produces the adversarial examples for AdvDB to generate the logits distribution of adversarial examples and clean images. Note that the AEQA cannot assist in the evaluation of a single adversarial example whose source model is unknown but in the evaluation of an adversarial attack technique.

B. THE CORRELATIONS BETWEEN ATTACK INTENSITY AND JS-DISTANCE

In AEQA, we consider perceptual quality and attack intensity as two crucial factors for evaluating the overall quality of adversarial examples and adopt AdvDSS and JS-Distance to quantify these two factors. We have proven the positive correlations between AdvDSS and the perceptual quality of adversarial examples in the main paper, here we further prove the efficacy of JS-Distance in evaluating the attack intensity of adversarial examples. Specifically, we compute the average JS-Distance between the logits distribution of benign images and adversarial examples with different attack settings and present the results in Figure 1. We can see that as the perturbation gets stronger, i.e., larger ϵ and iterative steps, can lead to an increase of JS-Distance between the generated adversarial examples and benign images, which indicates that a larger JS-Distance between the adversarial example and its corresponding benign image denotes a stronger attack intensity.

C. MORE VISUALIZATION COMPARISON OF COMPARED METHODS

We present more visualizations of the scatter plots of MOS versus different IQA methods in Figure 2, where we can see that the points obtained by ILNIQE, MUSIQ, CW-SSIM and PI resides in a certain region, indicating that these three metrics cannot accurately capture the adversarial distortions. The points of LPIPS, LPIPS-VGG, FISM are distributed better, but still do not well fit the curve. Among all the methods, the points obtained by our proposed AdvDSS are more

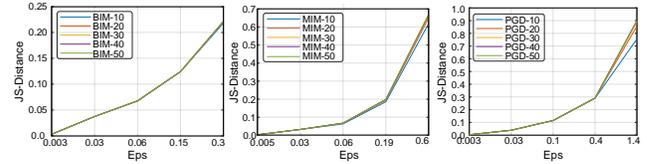


Figure 1: The JS-Distance between the logits distribution of benign images and adversarial examples with different attack settings.

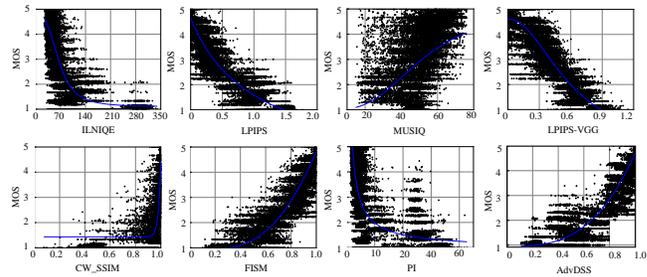


Figure 2: Scatter plots of different IQA methods versus MOS for all the adversarial examples from our AdvDB.

tightly distributed on the fitted curve, indicating a better correlation with MOS.

D. GENERATING ADVERSARIAL EXAMPLES USING AEQA

D.1 Generation

To optimize an adversarial example with AEQA, we set the AEQA as the objective and use L-BFGS algorithm to directly generate an adversarial example from a clean image with high AEQA score. We also use the Inc V3 pre-trained on ImageNet as the target model and also the backbone to compute JS-Distance. We set the iteration step as 20 during optimization.

D.2 Qualitative comparison

The generated adversarial examples are shown in Figure 3. The first column denotes the clean images and the second to sixth columns denote the adversarial examples produced by different methods. As we can see, directly optimizing AEQA can have a visual-pleasant result compared to the adversarial examples produced by other methods. As mentioned in our paper, the superior results contribute to the optimization of both attack intensity and perceptual quality, which also indicate that the proposed AEQA can well evaluate the adversarial example quality.

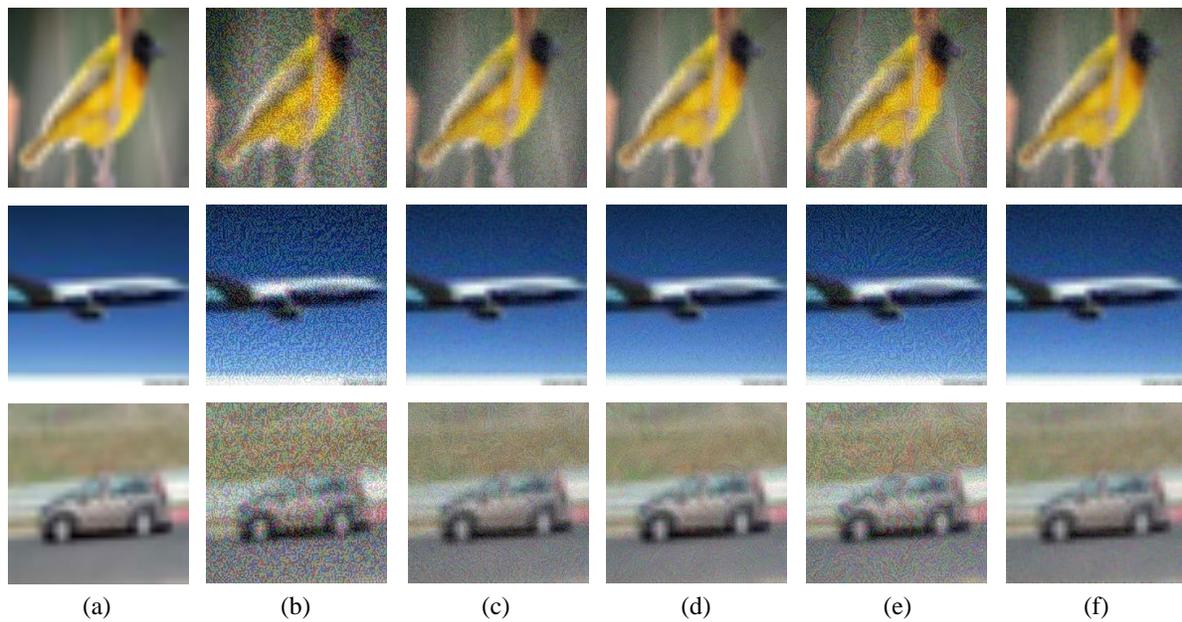


Figure 3: Adversarial example generation using different methods. (a) The benign images. (b)-(f) Adversarial examples produced by FGSM, PGD, BIM, MIM, and AEQA, respectively.