Supplementary Material for "ADC: Adversarial attacks against object Detection that evade Context consistency checks"

In this supplementary material, we provide: (a) the reconstruction error distribution plots of the context profiles for each object category, (b) gray-box attack performance on the MS COCO dataset, and (c) ablation study on the MS COCO dataset.

1. Reconstruction error distribution

As shown in Fig3.(a) in the main paper, the reconstruction error distribution of the context profiles from ADC perturbed images is very similar to that from benign images, and that is why ADC attack can evade the context-inconsistency based defense. In other words, the proposed ADC method achieves context-aware attacks to fool both the object detector and the attack detector.

Note that in the defense system, there is one auto-encoder for one category. What is plotted in Fig3.(a) is the reconstruction error distribution for all the categories on the PASCAL VOC dataset. We in this section present the reconstruction error distribution per object category. Fig. 1 shows the results. As we can see, our previous observation holds for each category (subplot), i.e., the distribution of the ADC generated images mimics that of the benign images for each object category. This further proves that ADC can generate context-aware attacks that are able to bypass context-consistency checks.

2. Gray-box attack performance on MS COCO dataset

We present in the main paper the gray-box attack performance on the PASCAL VOC dataset. For completeness, in this section, we show the gray-box attack performance on the MS COCO dataset. The surrogate system is train on *coco14valminusminival*. As shown in Table 1, gray-box attack the MS COCO dataset achieves very similar results compared to white-box attack, which is aligned with what we observe on the PASCAL VOC dataset.

3. Ablation study on MS COCO dataset

We show in the main paper that by constraining the perturbed area to the target object region, the attack performance is worse, which implies that context-aware attack need to perturb not only the target region, but also other regions to achieve context consistency. In this section, we present a same ablation study on the MS COCO dataset. The results are shown in Table 2. Similar to the results on the PASCAL VOC dataset, hiding attack is not affected by much, implying hiding attacks do not need to perturbations over other regions; however, mis-categorization and appearing attack performance is lower when the other regions are not perturbed.

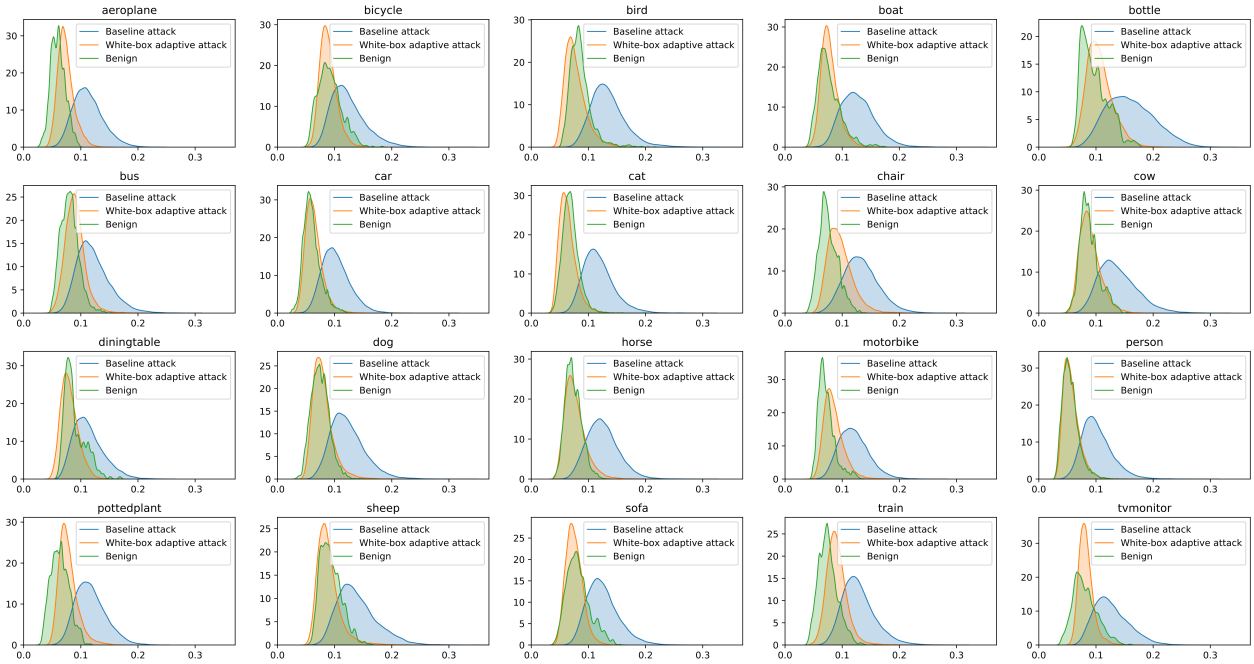

Figure 1: Reconstruction error distribution plot for each category.

Table 1: Gray-box attack performance for the three attack goals on the the MS COCO dataset.

Threat Model	Fooling Rate			Bypass Rate		
	Mis-categorization	Hiding	Appearing	Mis-categorization	Hiding	Appearing
White-Box	86.90%	90.82%	66.36%	83.67%	75.52%	86.08%
Gray-Box	76.25%	90.82%	60.51%	83.86%	75.89%	85.98%

Table 2: Attack performance when only attacking the target object region for the three attack goals on the MS COCO dataset.

Attack Region	Fooling Rate			Bypass Rate		
	Mis-categorization	Hiding	Appearing	Mis-categorization	Hiding	Appearing
Whole image	86.90%	90.82%	66.36%	83.67%	75.52%	86.08%
Target object	67.22%	90.82%	43.14%	88.04%	80.93%	89.48%