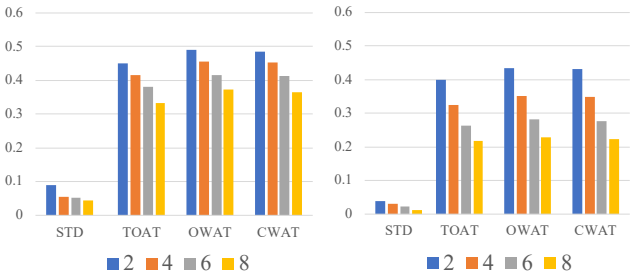
Supplemental Materials for Class-Aware Robust Adversarial Training for Object Detection

. A. Attack under Different Number of PGD Steps and Different Budgets

To evaluate the performance of the proposed adversarial training for object detection and compare with previous methods, we first attack the models using the adversarial examples generated with different number of PGD steps. As shown in Figure 3, the proposed OWAT and CWAT both can enhance the robustness for these settings. With the proposed CWAT, the performance can be significantly enhanced as compared with our implemented MTD-fast where MTD [3] is the recent state-of-the-art adversarial training method for object detection. In addition, by taking both training time and the training settings of other related works into consideration, we choose PGD-10 to generate the adversarial examples for training. Moreover, we also evaluate each model under the adversarial attacks with different budgets as shown in Figure 1 and Figure 2.



(a) A_{cls} PGD-10 attack (b) A_{reg} PGD-10 attack

Figure 1. The robustness of each model under PGD-10 attacks from different budgets in PASCAL VOC 2007 test set.

. B. The Impact of Fast Adversarial Training

It can be 7 to 30 times faster than the corresponding PGD-based adversarial training as mentioned in [1]. Moreover, the original PGD-10 adversarial training in [3] needs 23 back-propagations (each task costs 10 to generate task-oriented adversarial example, 2 to determine which example is used for final training, and 1 back-propagation for the final model update) per-iteration. On the other hand, our proposed methods only uses 2 back-propagation (1 CWT, and 1 for the update). Therefore, the original approach MTD would take additional 21 back-propagations. For the experiments, our fast CWAT is 3.19x faster than MTD with 4 2080Ti GPUs and batch size 14 for each GPU.

. C. More Details for Training

For the proposed adversarial training, we select all the positive anchors after each anchor has predicted. The positive anchors in the SSD are those that their IOUs between the ground truth are greater than 0.5. When we calculate the loss, we use all positive anchors and choose a certain percentage of negative anchors. Then we utilize this loss to calculate the attack gradient. Note that this procedure does not include non-maximum suppression (NMS). The same as the original SSD training, we do not use NMS when training, and the NMS is used in inference and test. The proposed method will attack all positive anchors rather than the single anchor that has the maximum IOU.

. D. The Results under Different Kinds of Attacks

The visualization of the detection results of an image under different attack are shown in Figure 3. These detection examples show the adversarial examples generated by the proposed method can more evenly attack all the objects occurred in the image than 3b and 3c which use total losses to generate the adversarial attack.

. E. More Qualitative Results for the Proposed CWAT Detector

Figure 4 illustrates the visualization results of object detection for the standard and the proposed CWAT models under different adversarial attacks for object detection. The first column is the detection results of the standard model (STD) upon clean images. The second column is the detection results of the standard model under the proposed class-wise attack (CWA). As the figure shown, all the objects in the images are detected incorrectly. The CWA is effective to fool the object detection model as demonstrated in the main paper. Furthermore, the third and fourth columns are the detection results of the CWAT model to defend against CWA and DAG attacks [2]. As the figures illustrated, the detection results using the proposed CWAT trained detector

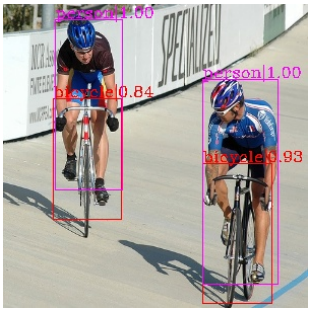
(a) $\epsilon = 0$ (b) $\epsilon = 2$ (c) $\epsilon = 4$ (d) $\epsilon = 6$ (e) $\epsilon = 8$

Figure 2. The detection results of the class-wise adversarial attack with PGD-10 in different ϵs, using the clean SSD as the targeted model. White label, yellow label, magenta label, blue label, green label, and black label represent classes of the dining table, chair, potted plant, sofa, tv-monitor, and respectively.

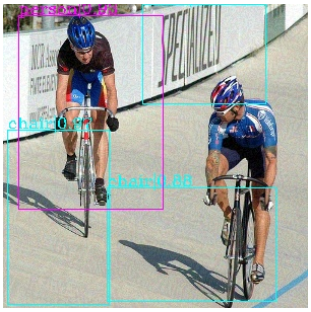
(a) Clean Image Result (b) Vanilla Adversarial Attack (c) Multi-task domain attack (d) Object-wise attack (e) Class-wise Attack

Figure 3. Detection results after attacked by different adversarial examples to the vanilla SSD model. (a) the detection result of a clean image, (b) the detection result after attacked by the adversarial example crafted through the 20-step PGD optimization with the budget $\epsilon = 16$ on the multi-task loss as described in equation 1, (c) the detection result after the multi-task domain attack which we follow [3] to implement it, (d) the detection result after the proposed object-wise attack, (e) the detection result after the proposed class-wise attack.

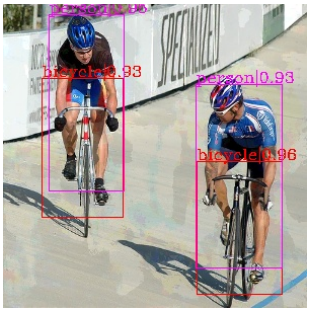
are almost the same as the ones using the clean model upon clean images. This further confirms the effectiveness of the proposed CWAT method.



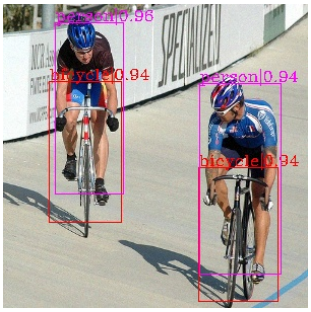
(a) No attack; Model: **STD**



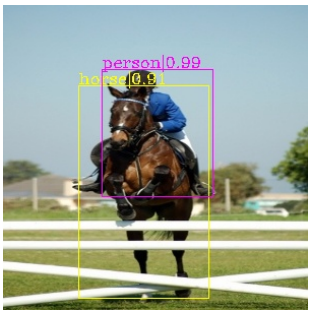
(b) Attack: **CWA**; Model: **STD**



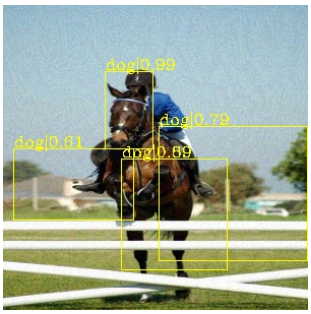
(c) Attack: **CWA**; Model: **CWAT**



(d) Attack: **DAG**; Model: **CWAT**

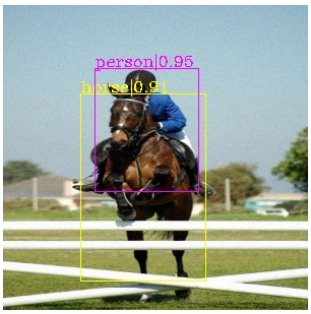


(e) No attack; Model: **STD**

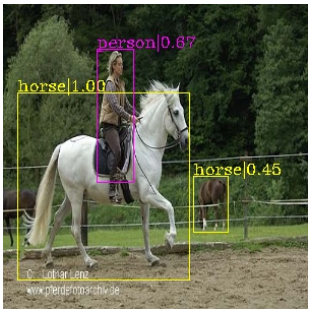


(f) Attack: **CWA**; Model: **STD**

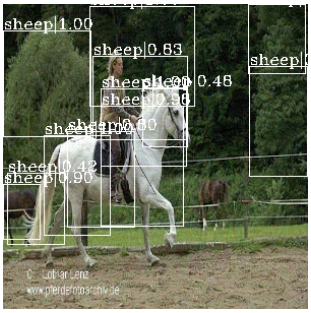
(g) Attack: **CWA**; Model: **CWAT**



(h) Attack: **DAG**; Model: **CWAT**



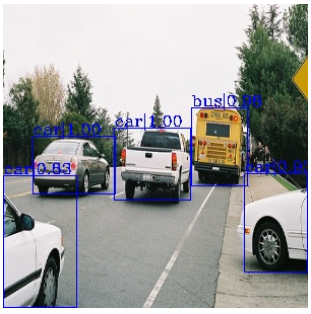
(i) No attack; Model: **STD**



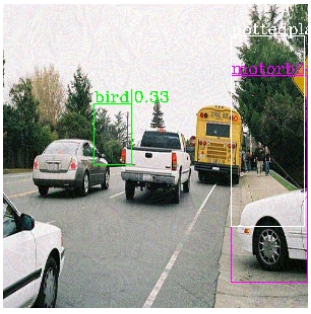
(j) Attack: **CWA**; Model: **STD**

(k) Attack: **CWA**; Model: **CWAT**

(l) Attack: **DAG**; Model: **CWAT**



(m) No attack; Model: **STD**



(n) Attack: **CWA**; Model: **STD**

(o) Attack: **CWA**; Model: **CWAT**

(p) Attack: **DAG**; Model: **CWAT**

Figure 4. Visualization results (attack budget = 8/255). The first column is STD model with no attack. The second column is STD model under class-wise attacks. The third column is CWAT defense model against class-wise attacks. The fourth column is CWAT defense model against DAG[2] attack.

References

[1] Rey Wiyatno and Anqi Xu. Maximal jacobian-based saliency map attack. *CoRR*, abs/1808.07945, 2018. 1

[2] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017. 1, 3

[3] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 421–430, 2019. 1, 2