Adversarial Pixel Masking: Supplementary Materials

Ping-Han Chiang
bhchiang@datalab.cs.nthu.edu.tw
National Tsing Hua University
Hsinchu, Taiwan R.O.C.

Chi-Shen Chan
csch@datalab.cs.nthu.edu.tw
National Tsing Hua University
Hsinchu, Taiwan R.O.C.

Shan-Hung Wu
shwu@cs.nthu.edu.tw
National Tsing Hua University
Hsinchu, Taiwan R.O.C.

ACM Reference Format:
Ping-Han Chiang, Chi-Shen Chan, and Shan-Hung Wu. 2021. Adversarial Pixel Masking: Supplementary Materials. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3474085.3475338

In this document, we provide more details about the settings of our experiments. We also conduct more experiments to verify the effectiveness of APM.

1 ATTACK MODEL

APM can be used to defend against most existing physical attacks provided that the attacks have been considered by the threat model $\mathcal{T}$ in Algorithm 1 in the main paper. However, generating patch pixels by replaying all known attacks could slow down adversarial training significantly. Since most existing physical attacks aim to manipulate the objectiveness scores and/or classification scores of candidate objects, we can define an unified objective for generating patch content. For ease of presentation, we consider YOLO [1, 8, 9] as the pre-trained object detector here. The objective can be easily adapted to other types of detectors. A YOLO detector divides the input space into grids $\mathcal{G}$, and in every grid $g \in \mathcal{G}$ there is a set $\mathcal{A}$ of pre-defined anchors. Given a masked input image $\boldsymbol{x} \odot \boldsymbol{m}$, the detector (prepended by MaskNet) outputs three elements for every anchor $a \in \mathcal{A}$: the coordinates of a bounding box (relative to a), the corresponding objectiveness scores $o(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\theta})_{g,a} \in \mathbb{R}$, and the corresponding classification scores $\boldsymbol{c}(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\theta})_{g,a} \in \mathbb{R}^{|C|}$ for all candidate classes C. The unified objective is then defined as

$$\arg\min_{\boldsymbol{p} \in \mathcal{T}} \sum_{g \in \mathcal{G}, a \in \mathcal{A}, c \in C} o(\boldsymbol{x}'; \xi, \boldsymbol{\theta})_{g,a} \cdot c(\boldsymbol{x}'; \xi, \boldsymbol{\theta})_{g,a,c} + \lambda \Omega(\boldsymbol{p}), \quad (1)$$

where $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{p}$ is a perturbed image and $\Omega(\boldsymbol{p})$ is a regularization term that encourages, for example, pixel smoothness. Eq. (1) is a realization of the inner max problem of Eq. (2) in the main paper. By considering both the objectiveness and classification scores, the generated adversarial examples can guide the MaskNet to defend against the ignorance attacks [4, 14, 16, 19, 20] which aim to make some important objects disappear, false-positive attacks [2, 14] which aim to create non-existing objects, and classification attacks [2, 12] which aim to mislead object labels.

Table 1: Performance of LGS with different grid sizes on INRIA dataset.

	LGS (15x15)		LGS (40x40)	
	High	Low	High	Low
Clean	**87.2**		86.7	
Noise	**86.4**		86.0	
ATK(10)	76.9	76.7	**78.1**	**78.1**
ATK(30)	72.8	64.8	**75.4**	**69.4**
ATK(50)	72.5	58.7	**75.0**	**63.1**
ATK(100)	69.4	49.7	**73.3**	**58.5**

(a) (b)

Figure 1: Example masks given by the MaskNet after the (a) first and (b) second stages of adversarial training.

2 DETAILED SETTINGS OF EXPERIMENTS

Local Gradients Smoothing (LGS). The performance of LGS is largely influenced by two hyper-parameters: threshold and grid size. As mentioned in Section 4 of the main paper, we followed the original paper [6] to set the threshold. On the other hand, we use more fine-grained grids by setting the grid size to 40×40 instead of the original 15×15. As Table 1 shows, this leads to better robustness.

Role of Spatial Concept (ROC). During training, the ROC [11] adds an additional regularization term that encourages a model to focus only on the features within the bounding box of each candidate object. This is done by maximizing the saliency maps within the bounding box, where each saliency map is the gradients of the object's confidence score with regard to a feature map at a deep layer. We use the last layer of the feature extractor of the object detector to calculate the saliency maps.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475338

Table 2: Performance of different defenses with pre-trained RetinaNet on INRIA dataset. The RetinaNet weights are not fixed in APM.

	RetinaNet	APM
Clean	**71.1**	66.4
Noise	57.9	**80.8**
ATK(10)	6.5	**63.0**
ATK(30)	1.5	**41.7**
ATK(50)	0.7	**32.1**
ATK(100)	0.7	**19.7**

MaskNet architecture. We experimented with several architectures for the MaskNet in APM, and found that the U-Net architect [10] works the best. The U-Net consists of fully-convolutional layers, each is structured by a contracting path with an expansive path. So, it maintains the spatial information in feature space and increase the resolution of output. We also found that APM with U-Net converges faster than with other CNNs during adversarial training.

Transfer learning. As discussed in the main paper, we let the weights of the pre-trained object detector fine-tunable for a transfer learning task. To implement this, we employ a two-stage training process. In the first stage, we fix the weights of the pre-trained object detector and only train the MaskNet. Once the weights of MaskNet converge, we then fine-tune the weights of MaskNet and object detector jointly. These two-stage training stages both follow Algorithm 1 described in the main paper. Empirically, it prevents the random initial weights of MaskNet to ruin the object detector and results in more clear masks, as shown in Figure 1.

3 ADVERSARIAL ROBUSTNESS OF RETINANET

Here, we show that APM can also improve the robustness of a pre-trained object detector based on RetinaNet [5]. We consider the ignorance attack following the main paper. However, since the output of RetinaNet is different from that of YOLO, we use a slightly different objective for generating adversarial examples during adversarial training. Specifically, RetinaNet uses only the classification score as the confidence of a candidate object. In order to lower the confidence score and make the candidate object get suppressed by the non-maximum suppression procedure, we use the objective:

$$\arg\min_{\boldsymbol{p}\in\mathcal{T}} \sum_{g\in\mathcal{G},a\in\mathcal{A},c\in C} c(\boldsymbol{x}';\xi,\boldsymbol{\theta})_{g,a,c} + \lambda\Omega(\boldsymbol{p}).$$

We test the adversarial robustness of APM on INRIA with the same settings described in Section 4.2 of the main paper. The results are shown in Table 2. We can see that APM improves the adversarial robustness of RetinaNet. However, the improvement is not as significant as we have seen in the main paper, and the clean performance drops. We believe this problem is rooted from the pre-trained RetinaNet itself. The pre-trained RetinaNet we used, which is directly downloadable from the TensorFlow repository,[1] only give an mAP of 71.1 for clean images. This is lower than the 87.4 given by the pre-trained YOLOv3 used in the main paper. Consider Eq. (2) in the main paper, the object detector needs to be able to correctly identify objects when the MaskNet (ξ) successfully masks an adversarial patch. Without a properly pre-trained object detector, the MaskNet cannot learn to "fix" the vulnerability of the detector at pixel space because the loss L is largely resulted from $\boldsymbol{\theta}$ rather than $\boldsymbol{p}$. Empirically, we found that the MaskNet tends to output all-pass masks in failed cases, as shown in Figure 2.

4 APM IN THE REAL WORLD

Here, we show more results of the experiment described in Section 4.5 of the main paper. To see whether APM can defend physical attacks in the real world, we created an universal adversarial patch [14, 17, 18] by regularizing the perturbations $\boldsymbol{p}$ (via the $\Omega(\boldsymbol{p})$ term in Eq. (1)) such that $\boldsymbol{p}$ 1) is "universal" in the sense that it can be applied to different input images, and 2) takes into account some common post-print distortions such as white noise, rotation, brightness shift, etc. Subsequently, we print the universal patch out and see if it can mute objects when held by real people. Figures 5 in the main paper and 3 here show that APM can successfully defend against the universal patch in the real world. Interestingly, Figure 3 also shows that APM can defend against the universal patch even in the indoor scenes, which is very different from those in the INRIA training dataset. We also find that, APM can successfully defend against the attack by only masking some critical portions of the patch. An universal real-world patch seems to work only when all its pixels (or critical portions) are visible to the object detector. As our future work, we will conduct larger-scale experiments to further verify the effectiveness of APM in different real-world applications.

5 UNSEEN ATTACKS

To study the generalizability of APM, we at test time modify the threat model $\mathcal{T}$ such that it generates different adversarial patches than the ones used during adversarial training. Recall from Section 4 of the main paper that, on the INRIA dataset, APM was trained by the adversarial patches of 1) square shape, 2) size equal to 0.8 of the width of the corresponding objects, and 3) horizontal placement at the middle of the corresponding objects. Table 3(a) and Figure 4 shows the performance of APM when the size of adversarial patches varies. We can see that APM performs well even if the size of adversarial patches varies at test time. This is because the "human" objects in the INRIA dataset are of different sizes, so APM was trained to generalize. We further modify the attack by changing the horizontal placement of adversarial patches to the bottom of the corresponding objects. Table 3(b) and Figure 5 shows the results. We also test APM using the adversarial patches of different aspect ratios, and the results are shown in Table 3(c) and Figure 6. APM can generalize to rectangular adversarial patches, despite they have never be seen by APM. We leave the study of more test-time variety as our future work.

[1] See https://github.com/tensorflow/models/tree/master/research/object_detection.

Figure 2: When paired up with a pre-trained object detector having a high error rate, the MaskNet tends to output all-pass masks to help the detector see as mush information as possible to make correct predictions.

Table 3: Generalizability of APM on INRIA dataset under the attacks with adversarial patches having (a) different sizes, (b) different locations, and (c) different aspect ratios ($M = 50$).

Size (Ratio to Object Width)				
0.9	0.7	0.6	0.5	0.4
82.3	85.4	84.9	84.7	82.5

(a)

Clean	Size (at Bottom)			
	0.8	0.7	0.6	0.5
81.8	81.7	82.5	83.7	62.3

(b)

Aspect Ratio	
(2.5:1) at top	(2:1) at middle
79.2	81.0

(c)

Figure 3: APM against an universal physical attack in the real world.

6 BLACK-BOX ATTACKS

We also consider *black-box attacks*, where the weights δ and ξ in Eq. 2 in the main paper are *not* accessible to an adversary. The black-box attacks have shown to be possible in digital domains and/or for classification tasks [3, 7, 13, 15]. Nevertheless, to the best of our knowledge, there is no existing study that reports the existence of black-box attacks for object detection tasks in physical domains. To implement an black-box attack, we generate a physical attack using RetinaNet and then apply it to YOLO. Table 4 shows the results. As we can see, APM consistently improves the robustness.

Table 4: Robustness of APM under a black-box attack on INRIA dataset, where adversarial patches are generated using RetinaNet and then applied to YOLO.

	ATK(10)	ATK(10)	ATK(30)	ATK(50)
YOLO	81.6	82.8	83.0	83.5
APM	**91.3**	**91.0**	**91.3**	**91.2**

However, the black-box physical attack does not seem to transfer well across object detectors as the vanilla YOLO already has high robustness against the attack. This controverts the seemly-universal transferability of digital attacks in classification tasks [3, 7, 13, 15] and motivates further investigation, which we will leave as our future work.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018.

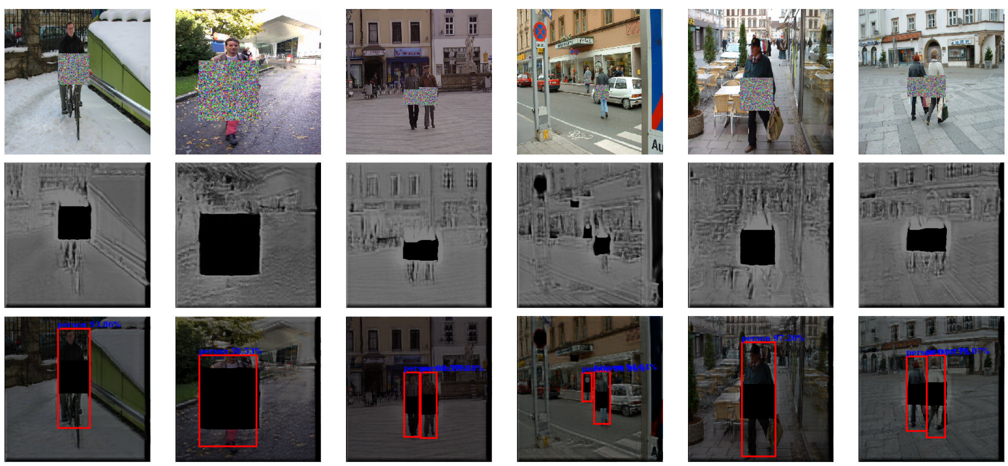
[3] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018.

[4] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*, 2019.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[6] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307. IEEE, 2019.

(a) Size: 0.4 of object width



(b) Size: 0.9 of object width

Figure 4: Generalizability of APM in defending unseen adversarial patches of different sizes.

[7] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[9] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. Role of spatial context in adversarial robustness for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 784–785, 2020.

[12] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[13] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[14] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[15] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[16] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020.

[17] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Evading real-time person detectors by adversarial t-shirt. *arXiv preprint arXiv:1910.11099*, 3, 2019.

[18] Darren Yu Yang, Jay Xiong, Xincheng Li, Xu Yan, John Raiti, Yuntao Wang, HuaQiang Wu, and Zhenyu Zhong. Building towards" invisible cloak": Robust physical adversarial attack on yolo object detector. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 368–374. IEEE, 2018.

[19] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2018.

Figure 5: Generalizability of APM in defending unseen adversarial patches at different locations.

Figure 6: Generalizability of APM in defending unseen adversarial patches of aspect ratios.

[20] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1989–2004, 2019.