Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors

Zuxuan Wu[1,2], Ser-Nam Lim[2], Larry S. Davis[1], and Tom Goldstein[1,2]

[1]University of Maryland, College Park [2]Facebook AI

More baselines . We present results of using cropped patches from natural images (*e.g.*, cats and dogs) and adversarial patches generated for classification tasks [1,2] as baselines.

	R50-C4	R50-FPN
R50_C4 Patch	24.5	31.4
R50_FPN Patch	20.9	23.5
Cat	50.8	49.9
Dog	50.9	48.2
Seurat	47.9	51.6
Grey	45.9	50
Clean	78.7	82.2
Brown *et al.* [1]	51.9	54.0
Moosavi-Dezfooli *et al.* [2]	61.5	61.3

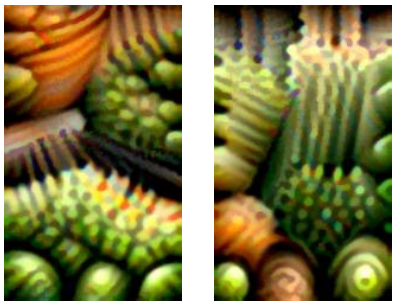

Fig. 1: **Patches learned with the TV norm (left) and without the TV norm**, using the R50-C4 as the backbone of Faster-RCNN.

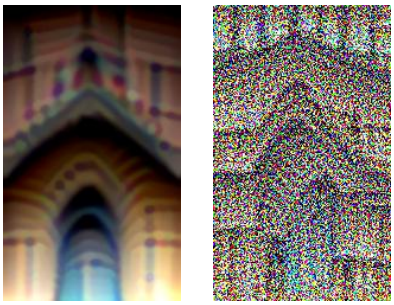

Fig. 2: **Patches learned for the "bus" and "horse" classes, respectively**, using the R50-C4 as the backbone of Faster-RCNN.

Effectiveness of the TV norm. Figure 1 presents patches learned with and without the TV norm. We can see that using the norm leads to smoothed patterns where all pixels are involved. Quantitatively, without the TV norm, the patch degrades the AP of the person detector from 78.7% to 26.9%, and it is worse than the performance of its counterpart with the TV norm (24.5%). We observe similar trends with other backbones.

Patches of "bus" and "horse" . We visualize learned "bus" and patches in Fig. 2.

A Gallery of Clothes. We demonstrate selected images of adversarial clothes in Figure 3 and show they can evade YOLOv2 with adversarially learned patterns.

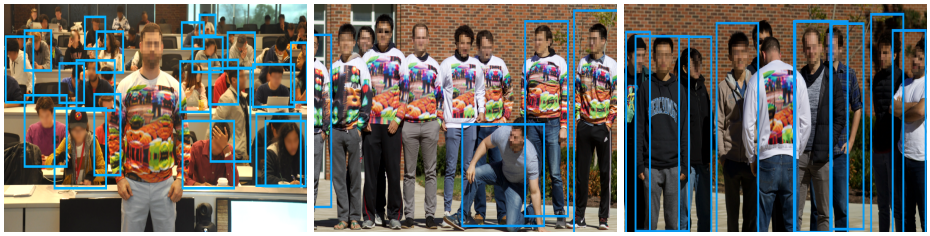

Fig. 3: **Selected images of adversarial clothes.**

References

1. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017) 1
2. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017) 1