DAP: A Dynamic Adversarial Patch for Evading Person Detectors

Supplementary Material

9. Limitations of GAN-based techniques

Limitation 1: Failure to converge

When attempting to reproduce the results of GAN-based approaches such as [13], we found it to be inefficient, with the code failing to converge to realistic patterns in some cases. In the experiment, we generated patches for several datasets, including CASIA, wildtrack, and INRIA, using the BigGAN generator [36] at an output resolution of 128 × 128, pre-trained on the ImageNet 1k dataset, with the class vector set to 'dog', as in [13]. The resulting patches, shown in Figure 7, were trained on different object detectors, including YOLOv2, YOLOv3, and FasterCNN. As depicted, the algorithm encountered challenges in converging towards a naturalistic patch. Despite iterative attempts, the optimization process struggled to yield a patch that seamlessly integrates with the visual context.

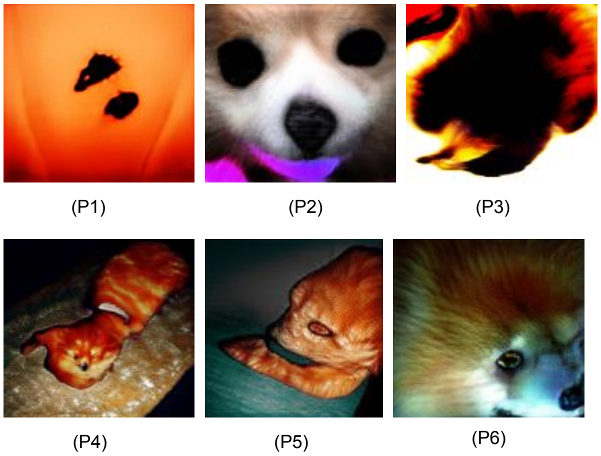


Figure 7. Failed attempts to generate GAN-based naturalistic patches: (P1) on YOLOv3, (P2) on Faster-CNN, (P3), (P4), (P5), and (P6) on YOLOv2.

Limitation 2: Too limited latent space

We run multiple experiments where we try to generate naturalistic patches with different combinations of transformations, i.e., without any transformations, with basic transformations (i.e. random noise, contrast and brightness variations), with basic transformations plus rotation, and when using all the transformations including the creases transformations that model wrinkles and creases in a person's cloth. Without loss of generality, we use the Yolov3 detector as our victim model. As illustrated in Figure 8-*Left*, for a norm threshold $\tau = 50$, the same value used in [13], with the increase in the number of considered transformation the effectiveness of the attack decreases. This could be explained by the fact that the latent space of the GAN is too limited to find a patch that is robust to all the basic, rigid and non-rigid transformations at once. To illustrate this, we run experiments by adjusting the norm threshold of the latent vector (i.e., $\tau = 1$ and $\tau = 100$), and report the convergence of the mean average precision (mAP) of the generated patch while training. As shown in Figure 8-*Middle*, corresponding to a more strict constraint, i.e., $\tau = 1$, the resultant patch converges to a higher mAP value leading to a lower attack success rate. We also test for $\tau = 100$, we notice that the generated patches converge to a lower mAP compared to the $\tau = 1$ case, which corresponds to a higher attack success rate. We present the final mAP for different experiments in Table 11. Figure 8 also shows the generated patches for different norm thresholds, where it can be noticed that the lower the threshold the more naturalistic the generated patch looks. To conclude, the norm threshold allows a trade-off between realism and attack efficiency.

Transformations	$\tau = 1$	$\tau = 50$	$\tau = 100$
No transformation	52.81%	47.57%	49.67%
Noise	52.55%	48.51%	49.60%
Noise + Rotation	54.05%	50.44%	49.99%
Noise + Rotation + Creases	55.16%	52.07%	51.76%

Table 11. mAP of GAN-based technique when training using different transformations.

We also used StyleGAN [15] to generate naturalistic patches, targeting YOLOv3, using two latent vector thresholds: 50 and 100. As depicted in Table 12, incorporating transformations resulted in less effective patches. This observation aligns with our findings from other GAN types. This will be included in the last version.

Threshold	**w/o**	**w**
50	44.42	47.88
100	44.87	49.14

Table 12. StyleGAN-based patch performance with and without transformation.

10. Transferability between Different Detector Architectures

In Table 13, we present the transferability results. Specifically, we utilize a patch generated for the FasterRCNN architecture to evaluate various YOLO-based

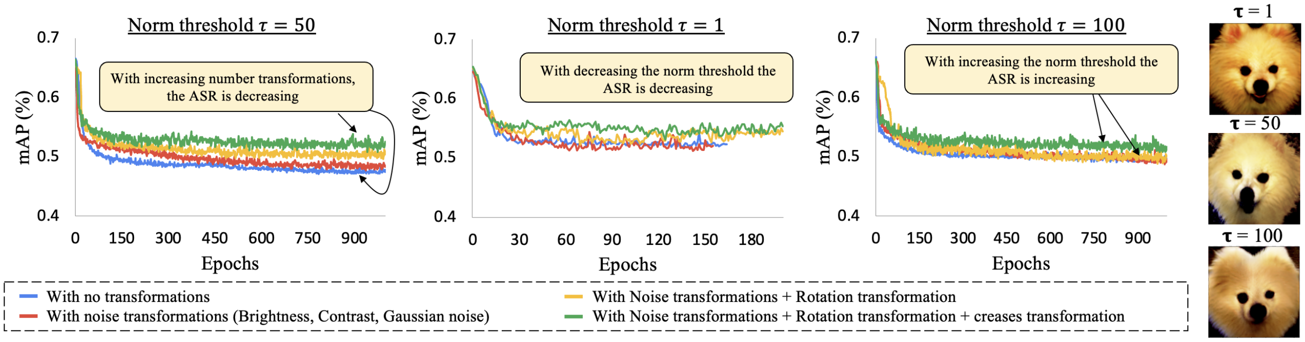


Figure 8. Mean Average Precision (mAP) convergence curves when training a GAN-based technique with and without different transformations. Illustrating the impact of adjusting the latent vector constraints on attack success rate (ASR) (*Left:* $\tau = 50$ (same used in [13]), *Middle:* $\tau = 1$, and *Right:* $\tau = 100$). Summary of these curves are presented in Table 11

detectors. Notably, our patch remains effective across these detectors.

Detectors	FasterRCNN
Yolov3	42.47%
Yolov3tiny	51.5%
Yolov7	50.09%
FasterRCNN	13.05%

Table 13. Transferability of FasterRCNN-based patches to Yolo architectures.

11. Impact of Patch Size

To assess the impact of patch size on the efficacy of our proposed adversarial patch, a series of digital experiments were carried out on the INRIA dataset. The objective was to evaluate how different patch sizes affect the effectiveness of our approach. Figure 9 visually represents the various scales employed to generate distinct patches on the targeted objects. It is worth noting that a scale of 0.5 in our experiments corresponds to a scale of 0.2 as described in [13]. By conducting these experiments and analyzing the outcomes, we aimed to gain insights into the relationship between patch size and the performance of our adversarial patch, ultimately providing valuable information for optimizing its effectiveness.

12. Physical World Experiments

12.1. Printed Patch Performance

To evaluate the real-world effectiveness of our proposed adversarial patch, we conducted a physical attack experiment where we printed the patch and tested its performance in a real-world setting. In the experiment, we took pictures of one person holding the printed patch and tested whether it could successfully evade detection by an object detector. For the with deformations setting, we added random creases to the paper by crumpling this latter. The results of this experiment are illustrated in Figure 10

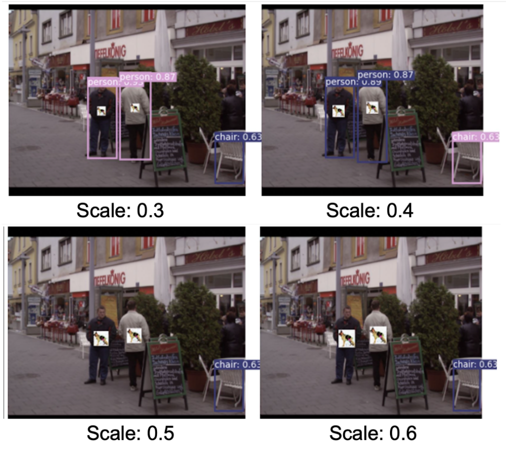


Figure 9. Illustration of different patch scales.

We carried out our real-world experiments using a set of 100 test samples, each featuring a person holding a printed patch. These samples encompassed a range of transformations, such as rotation, resizing, perspective changes, and other variations. Moreover, we report the attack success rate as an indicator of the effectiveness of our approach in the context of these real-world transformations. Our proposed patch maintained its effectiveness for different distances from the camera, different angles, different scales and for different lighting conditions. Table 14 reports the patch success rate in the physical world and when using our DAP, only 35% of the time a person is detected.

As shown in Figure 10, even after being rotated, our patch was still effective in hiding the person from detection. In contrast, when we applied the same deformations to the naturalistic patches from [13] (NAP), we found that the

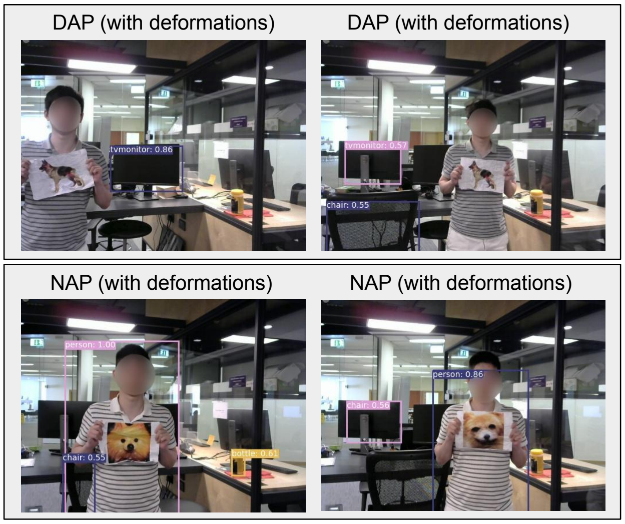


Figure 10. Detection results when applying random creases and rotations to the DAP patches (*upper*) for Yolov3tiny and NAP patches (*lower*) for Yolov3 and Yolov3tiny.

patch was no longer effective in evading detection. This highlights the strength of our proposed patch, which is more robust to physical deformations and can maintain its effectiveness even when the patch is crumpled and rotated. As shown in Table 14, the detection recall dropped to 45% even with the presence of multiple creases in the patch in addition to applying rotations. In a physical setting, when using yolov7, a person is detected approximately 48% of the time when using our proposed patch and 82% of the time when using NAP (See Table 15).

	Benign	**NAP**	**DAP**
Without creases	100%	30%	75%
With creases	100%	20%	65%

Table 14. Attack Success Rate (ASR) in Benign scenarios and when using DAP and NAP [13] when attacking Yolov3tiny with and without creases.

Detector	**Benign**	**NAP**	**DAP**
Yolov7	100%	18%	52%

Table 15. Attack Success Rate (ASR) in Benign scenarios and when using DAP and NAP [13] when attacking Yolov7.

13. Adversarial T-shirt Performance

To thoroughly evaluate the robustness of our adversarial patch, we subjected the T-shirt, along with our patch, to aggressive transformations. Notably, even under these challenging conditions, our patch consistently outperformed the GAN based patch. Despite the aggressive transformations applied to the T-shirt, our patch demonstrated remarkable resilience and maintained its effectiveness in evading detection (See Figure 11). In contrast, the NAP patch struggled to retain its deceptive properties under similar transformations (See Figure 12). These findings provide compelling evidence of the superior performance and robustness of our adversarial patch in the face of extreme alterations. Our patch's ability to withstand aggressive transformations further underscores its potential for reliable and effective evasion of detection systems, solidifying its superiority over the GAN-based patch.

The key metrics used to evaluate the performance of the detection system are as follows:

- True Positive Rate (TPR): This represents the percentage of total results that are correctly identified as positive, specifically when a patch is present and the person is not detected ($P = 1 \ \& \ D = 0$). It is computed as: $\mathrm{TPR} = \dfrac{\mathrm{TP}}{\mathrm{P}} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$.
- False Positive Rate (FPR): This indicates the percentage of total test results that are incorrectly identified as positive, occurring when a patch is present and a person is detected ($P = 1 \ \& \ D = 1$). It is calculated as: $\mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{N}} = \dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$.
- True Negative Rate (TNR): This corresponds to the percentage of cases ($P = 0 \ \& \ D = 1$) where the patch is not present, and the person is correctly detected. It is calculated as: $\mathrm{TNR} = \dfrac{\mathrm{TN}}{\mathrm{N}} = \dfrac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$.
- False Negative Rate (FNR): This represents the percentage of cases where the detector fails to detect a person even though the patch is not present ($P = 0 \ \& \ D = 0$). It is calculated as: $\mathrm{FNR} = \dfrac{\mathrm{FN}}{\mathrm{P}} = \dfrac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}$.

These evaluation metrics allow us to assess the performance of our adversarial patches and provide quantitative insights into the detection results obtained during our experiments.

14. Adversarial Patches in Different Classes

In these experiments, we selected alternative target natural images (i.e., Bicycle and Cat), and tested the performance of the generated adversarial patches. To ensure practical applicability, the patches were printed on A4 papers. Figure 13 provides a visual representation of our patch. Across various scales, rotation angles, lighting conditions, and distances from the camera, our proposed patches consistently exhibited high efficacy in concealing the presence of a person and deceiving the detector.

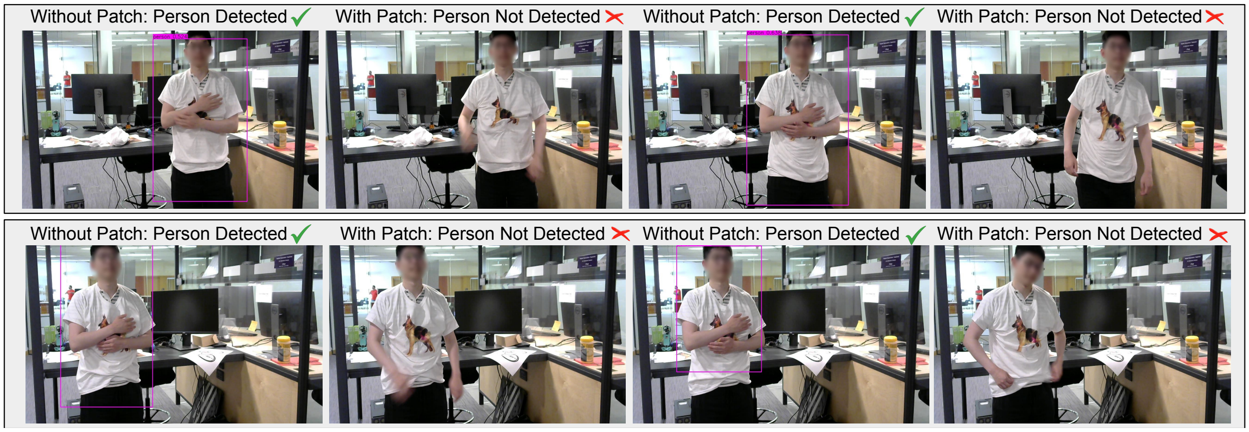

Without Patch: Person Detected ✓
With Patch: Person Not Detected ✗
Without Patch: Person Detected ✓
With Patch: Person Not Detected ✗
Without Patch: Person Detected ✓
With Patch: Person Not Detected ✗
Without Patch: Person Detected ✓
With Patch: Person Not Detected ✗

Figure 11. DAP-based T-shirts performance while applying different deformations.

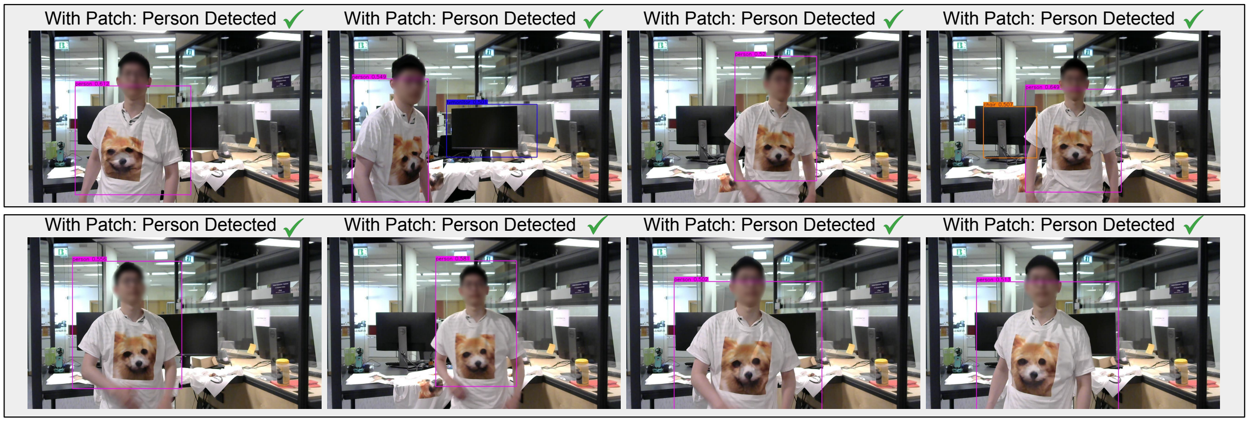
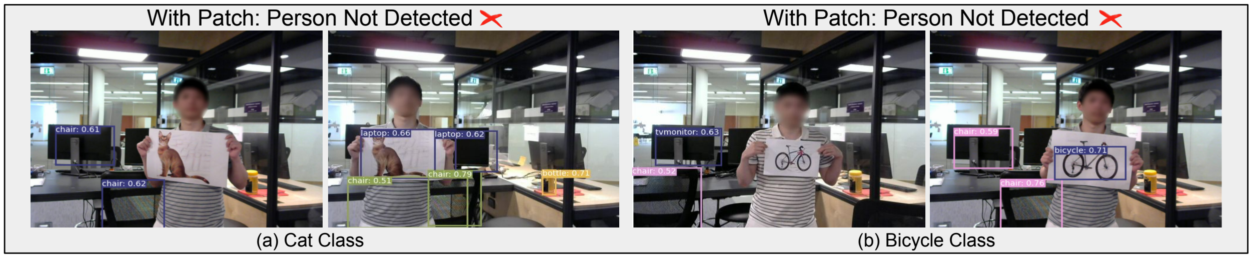

With Patch: Person Detected ✓
With Patch: Person Detected ✓
With Patch: Person Detected ✓
With Patch: Person Detected ✓
With Patch: Person Detected ✓
With Patch: Person Detected ✓
With Patch: Person Detected ✓
With Patch: Person Detected ✓

Figure 12. NAP-based T-shirts performance while applying different deformations.



(a) Cat Class

(b) Bicycle Class

Figure 13. Detection results for different angles, distances from the camera, scales, and lighting conditions for Cat and Bicycle-based DAP.