DAP: A Dynamic Adversarial Patch for Evading Person Detectors

Amira Guesmi[1], Ruitian Ding[2], Muhammad Abdullah Hanif[1], Ihsen Alouani[3,4], Muhammad Shafique[1]
[1] eBrain Lab, New York University (NYU) Abu Dhabi, UAE
[2] NYU Tandon School of Engineering, USA
[3] IEMN, CNRS-8520, INSA Hauts-de-France, France
[4] CSIT, Queen's University Belfast, UK

Abstract

Patch-based adversarial attacks were proven to compromise the robustness and reliability of computer vision systems. However, their conspicuous and easily detectable nature challenge their practicality in real-world setting. To address this, recent work has proposed using Generative Adversarial Networks (GANs) to generate naturalistic patches that may not attract human attention. However, such approaches suffer from a limited latent space making it challenging to produce a patch that is efficient, stealthy, and robust to multiple real-world transformations. This paper introduces a novel approach that produces a Dynamic Adversarial Patch (DAP) designed to overcome these limitations. DAP maintains a naturalistic appearance while optimizing attack efficiency and robustness to real-world transformations. The approach involves redefining the optimization problem and introducing a novel objective function that incorporates a similarity metric to guide the patch's creation. Unlike GAN-based techniques, the DAP directly modifies pixel values within the patch, providing increased flexibility and adaptability to multiple transformations. Furthermore, most clothing-based physical attacks assume static objects and ignore the possible transformations caused by non-rigid deformation due to changes in a person's pose. To address this limitation, a 'Creases Transformation' (CT) block is introduced, enhancing the patch's resilience to a variety of real-world distortions. Experimental results demonstrate that the proposed approach outperforms state-of-the-art attacks, achieving a success rate of up to 82.28% in the digital world when targeting the YOLOv7 detector and 65% in the physical world when targeting YOLOv3tiny detector deployed in edge-based smart cameras.

1. Introduction

Deep Neural Networks (DNNs) have demonstrated remarkable performance for various real-world applications and are now commonly used in safety-critical and security-sensitive domains such as video surveillance [5, 23, 31], self-driving cars [1], and healthcare [24].

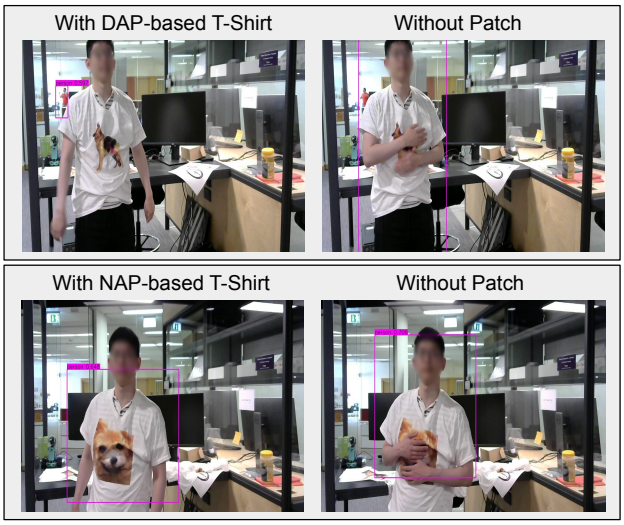

Figure 1. Illustration of different Adversarial T-shirts against Yolo detector. DAP-based t-shirt (ours) is still effective in the presence of non-rigid deformations compared to the GAN-based t-shirt (NAP) [13].

However, studies have shown that DNNs are vulnerable to adversarial perturbations [6, 18, 25, 26]. These adversarial examples can manifest physically and be deployed in real-world scenarios, presenting significant security and safety concerns.

There are two types of attack settings: digital attacks and physical attacks. In digital attacks [6, 11], attackers introduce adversarial noise to the digital input image, optimized to be undetectable by human eyes while monitoring the noise budget constraint in the generation process. In contrast, in physical attacks [9, 12, 16], the attacker designs patches that are printable in the physical world and deploys them in the scene captured by the

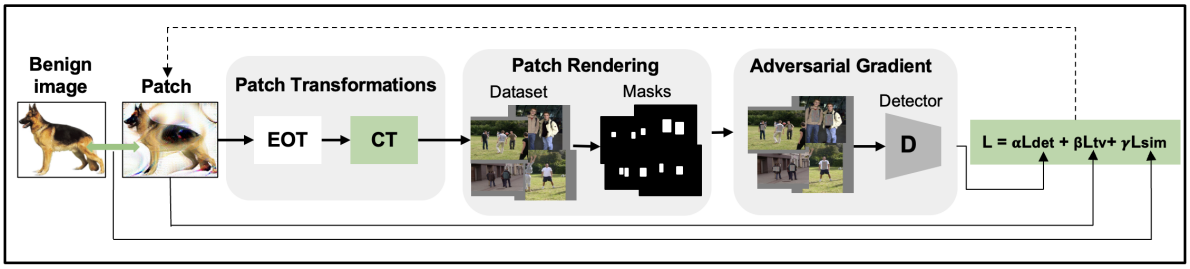

Figure 2. Overview of our dynamic adversarial patch generation framework which crafts patches that can be printed on a T-shirt and evade object detectors under different real-world conditions.

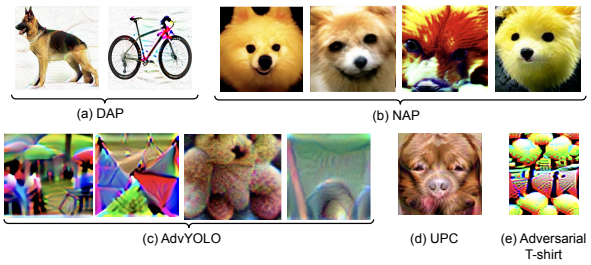

Figure 3. DAP vs State-of-the-Art patches: (a) DAP, (b) NAP (for YOLOv3tiny) [13], (c) Adversarial patch [33], (d) UPC [14], and (e) Adversarial T-shirt [37].

victim model. These patches are not generated under noise magnitude constraints but rather under location and printability constraints, making patch-based attacks more practical in real-life scenarios. However, current adversarial patches are limited in the following ways:

Conspicuous patterns. Previous research on adversarial patches for object/person detection [10, 33, 38] has largely focused on improving attack performance by increasing the strength of the adversarial noise. However, this approach often results in patches that are easily identifiable by human observers, limiting their effectiveness in real-world scenarios [4, 13, 19, 27]. To address this issue, some researchers have proposed using generative adversarial networks (GANs) to generate more natural-looking patches. While promising, these approaches can be inefficient (i.e., result in low attack success rate as depicted in Figure 1) and may not always converge to realistic/natural-looking or efficient patterns (refer to Section 2 for further details).

Assumption of static objects. Previous research assumed static objects [13, 14, 33], however, when printing an adversarial patch on a T-shirt, it is crucial to ensure that the patch is robust to additional transformations caused by changes in the fabric (e.g., deformation and orientation changes) due to a person's movements. In particular, the constantly changing wrinkles and creases in the fabric can significantly affect the effectiveness of the attack. We demonstrate that the GAN-based techniques suffer from a limited latent vector which makes incorporating multiple transformations at once very challenging (See Figure 1).

In this paper, we propose a novel framework (See Figure 2) to generate natural-looking adversarial patches for performing effective and robust adversarial attacks on object/person detection systems, by guiding the optimization process towards a target natural image and simultaneously maximizing the victim model's loss. To achieve this, we propose a novel objective function that includes a similarity loss to guide the pattern of the generated noise, resulting in natural-looking patches (see Figure 3). Additionally, we incorporate a crease transformation block to model possible deformations that could occur to the adversarial patch when used to conceal dynamic objects (See Figure 1).

Contributions – The main contributions of this paper are summarized as follows:

- We investigate the limitations of GAN-based approaches when used to generate robust naturalistic adversarial patterns. Our investigation reveals that these techniques struggle to integrate multiple transformations while sustaining optimal performance. We show that these techniques can not incorporate multiple transformations while maintaining high performance due to the limited latent space compared to the high flexibility and the larger space provided by our approach as it relies on directly manipulating the patch pixels.
- We propose a framework (See Figure 2) that generates GAN-free naturalistic patches that can resemble any predefined pattern, while maintaining high attack success rate under multiple transformations (e.g., clothing creases and wrinkles, random noise, brightness and contrast variations, re-scaling, and rotation, etc.).
- To increase robustness against non-rigid deformations experienced by a printed adversarial patch on a T-shirt and caused by pose changes of a moving person, we develop a Creases Transformation (CT) block that models these non-rigid deformations by compressing the pixels following a randomly selected direction.

- We thoroughly examine the performance/attack success rate of the proposed method in terms of mean average precision (mAP) both with and without transformations, and transferability between detectors. Our patch achieves an attack success rate of 82.28% in the digital world (INRIA dataset) when targeting the YOLOv7 detector and 65% in the physical world when attacking the YOLOv3tiny detector for edge systems.

2. Limitations of GAN-based techniques

The generator in a GAN is typically trained by sampling random vectors from a standard normal distribution, which results in a high density region centered around the origin. Thus, if the latent vector z is closer to the origin, there is a higher probability of generating realistic images, that is why a constraint τ on the norm of the latent vector z is required. In this section, we investigate the impact of incorporating multiple transformations in the naturalistic adversarial patch generation process and the impact of the norm threshold of the latent vector z on the effectiveness of GAN-based techniques for generating adversarial patches that are robust against real-world transformations.

To assess GAN-based techniques for generating naturalistic patches, we conducted experiments targeting the Yolov3tiny detector. Our emphasis was on comparing attack effectiveness under different conditions: no transformations vs. all transformations (including Noise, Rotation, and Creases). We also investigated various norm thresholds. This systematic exploration allowed us to assess how different transformation combinations influenced patch effectiveness in deceiving the Yolov3tiny detector, with consistent trends also noted in supplementary material for the Yolov4tiny detector. Specifically, as we incorporated more transformations into the generation process, we noticed a decrease in the effectiveness of the generated patch (See Figure 4 and Table 1). This observation suggests that there exists a trade-off between the number of transformations applied to the patch and its ability to deceive the detector effectively. While transformations can enhance the patch's camouflage and resilience against real-world transformations, an excessive number of transformations may introduce distortions that hinder its effectiveness.

Furthermore, we explored the impact of adjusting the norm threshold for the latent vector z on the performance and appearance of the generated patches. Through fine-tuning these thresholds, we observed enhancements in the effectiveness of the generated patches. Higher norm thresholds, such as $\tau = 100$, allowed for larger space, increasing the patch's effectiveness as shown in Table 1. However, it's essential to highlight that as we pushed the norm thresholds higher, the generated patches began to manifest unrealistic and visually conspicuous characteristics (as depicted in Figure 4). These unrealistic attributes could potentially compromise the patch's ability to remain inconspicuous and might raise suspicions, limiting its practicality in real-world scenarios. However,

Transformations	$\tau = 1$	$\tau = 100$
No transformation	23.43%	13.21%
Noise + Rotation + Creases	29.82%	21.53%

Table 1. mAP of GAN-based technique when training using different transformations.

with lower thresholds, the generated patches appear more naturalistic but are less effective at deceiving the target detector. This also results in a narrower latent space, reducing the flexibility to incorporate multiple transformations. It is worth noting that the tested GANs were not specifically trained on images augmented with these transformations, such as creases deformations. It was shown that training GANs with multiple data augmentation techniques could introduce additional challenges and distortions in the generated samples, potentially making them less naturalistic [34]. Further limitations of the existing GAN-based approaches, such as failure to converge, are discussed in the supplementary material.

3. Proposed Approach

In this paper, we propose an attack that simultaneously satisfies the three key requirements needed for the adoption of adversarial attacks in the real world. These requirements include: ***Effectiveness*** in degrading the person detector's performance. ***Stealthiness*** against human visual inspection (i.e., being unrecognizable by the observer). ***Robustness*** in maintaining attack ability in a dynamic environment including robustness to physical constraints.

3.1. Attack Effectiveness

Figure 2 represents an overview of the proposed framework. Our goal is to generate physical adversarial patches that are naturalistic while still maintaining their attack performance in real-world scenarios. We iteratively perform gradient updates on the adversarial patch (P) in the pixel space that optimizes our objective function, as defined below:

$$L_{total} = \alpha L_{det} + \beta L_{sim} + \gamma L_{tv} \quad (1)$$

L_{det} is the adversarial detection loss. L_{sim} is the similarity loss (See Section 3.2). L_{tv} is the total variation loss on the generated image to encourage smoothness (See Section 3.3). α, β, and γ are hyper-parameters used to scale the three losses. For our experiments we set $\alpha = 1$, $\beta = 4$, and $\gamma = 0.5$. We optimize the total loss using Adam [15] optimizer. We try to minimize the object function

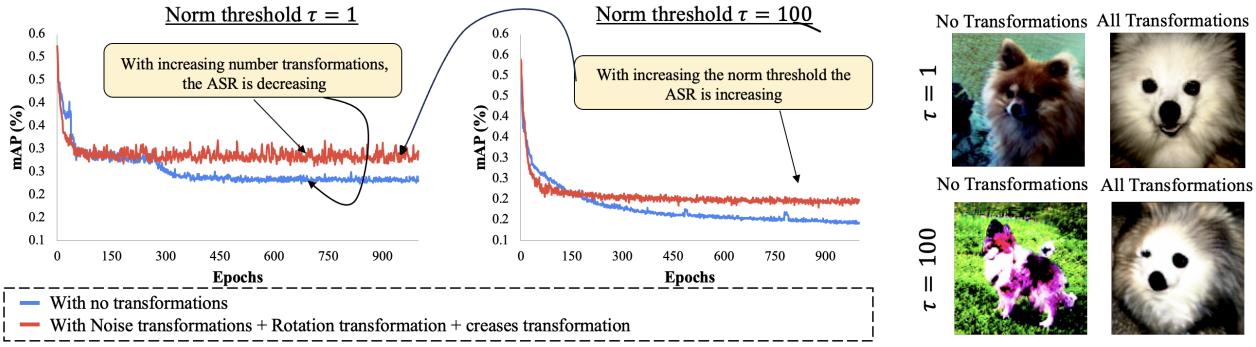

Figure 4. Mean Average Precision (mAP) convergence curves when training a GAN-based technique with and without different transformations. Illustrating the impact of adjusting the latent vector constraints on attack success rate (ASR) (*Left:* $\tau = 1$ and *Right:* $\tau = 100$) and the corresponding generated patch.

L_{total} and optimize the adversarial patch. We freeze all parameters of the object detector, and update only the pixel values of the adversarial patch starting from a random initialization. Object detectors, such as YOLO, output an arbitrary number of anchor boxes. For each detection j, the goal is to attack both the objectness score D^j_{obj} and its class probability D^j_{cls}. Minimizing the objectness score D^j_{obj} causes the jth object not to get detected. Minimizing D^j_{cls} causes the jth object to be misclassified. In this paper, we focus on targeting the person class, e.g., considering the ML-based smart surveillance use cases. Thus, we minimize both the objectness D^j_{obj} and class probabilities D^j_{cls} pertaining to the person class. Our adversarial detection loss is defined by:

$$L_{det} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\#objects}\sum_{j=1}^{\#objects}[D^j_{obj}(I'_i)D^j_{cls}(I'_i)]\right) \quad (2)$$

3.2. Attack Stealthiness

The key idea is to generate a patch similar to an existing image, and this is accomplished by defining a novel loss term that represents the distance of the patch to the original target image. The proposed similarity loss is designed to maximize the cosine similarity between the target benign image and the adversarial patch P. The similarity loss is defined as:

$$L_{sim} = -\left(\frac{\sum_{i,j} P_{i,j}N_{i,j}}{\sqrt{\sum_{i,j} P^2_{i,j}}\sqrt{\sum_{i,j} N^2_{i,j}}}\right)^2 \quad (3)$$

During the optimization process, we observed discrepancies in the convergence rates of various loss terms. Specifically, the similarity loss converges notably faster than the detection loss, resulting in a stagnation of the total loss. To address this imbalance, we introduce a non-linear adjustment aimed at slowing down the rate of decrease of the similarity loss. Our approach involves squaring the difference between the benign image and the adversarial patch. This modification is to mitigate the dominance of the similarity loss during optimization, leveraging the characteristic slower rates of increase exhibited by quadratic functions compared to linear functions. Effectively, this strategy downweights the impact of the squared term, striking a balance between the two loss terms and preventing one from unduly dominating the optimization process. This similarity loss provides a higher flexibility compared to the GAN-based technique and the limited latent space.

3.3. Attack Robustness

Introducing the digital attack into the physical world poses an additional challenge, as the perturbation must be strong enough to withstand real-world distortions arising from variations in viewing distances and angles, lighting conditions, camera limitations, and dynamic objects. Previous studies have revealed that adversarial examples generated via conventional techniques frequently cease to be adversarial after undergoing slight transformations [20, 21]. In order to ensure patch robustness we use two preprocessing blocks (i.e., EOT and CT blocks) to generate a large variation of transformed patches to be used in the patch training process.

3.3.1 Expectation Over Transformation (EOT)

EOT is a general framework for improving the adversarial robustness of physical attack on a given transformation distribution T [3]. Essentially, EOT takes potential transformation in the real world into account during the optimization, resulting in higher efficiency. EOT is used to add random distortions in the optimization to make the perturbation more robust. The transformation distribution is presented in Table 2.

Transformations	Parameters	Remark
Rotation	$\pm 20°$	Camera Simulation
Affine	0.7	Perspective
Scale	$[0.25, 1.25]$	Distance/Resize
Random Noise	± 0.1	Noise
Brightness	± 0.1	Illumination
Contrast	$[0.8, 1.2]$	Camera Parameters

Table 2. Transformation distribution.

3.3.2 Creases Transformation (CT)

Another possible transformation when printing the generated patch on a t-shirt is constantly changing creases in the clothes resulting from a person's movements. To overcome this challenge we propose to perform the following transformations: Each crease is modeled by randomly selecting a point on the patch, along with a 2D vector representing the crease's direction, which is chosen within a range of 5 degrees to simulate the alignment on the clothing surface. The selected pixels within the patch are displaced in the direction defined by the vector, with varying degrees of movement based on their proximity to the vector's line. This displacement emulates the natural variation in the crease intensity along their length. The displacement of a point (x, y) resulting from the chosen crease point (x_0, y_0) and the associated 2D vector is calculated as Displacement = Vector x Multiplier, where the multiplier captures the dynamic nature of creases. This process is performed for each incorporated crease. The movement of a point (x, y) when the chosen point is (x_0, y_0) is the vector multiplied by a multiplier, which follows the equation:

$$mult(x, y) = 1 - \frac{\sin^2 \theta[(x - x_0)^2 + (y - y_0)^2]}{width^2 + height^2} \quad (4)$$

Where θ is the angle between the direction of the (x, y) from (x_0, y_0) and the chosen vector, and *width* and *height* are the dimensions of the patch. The resulting transformations are illustrated in Figure 5.

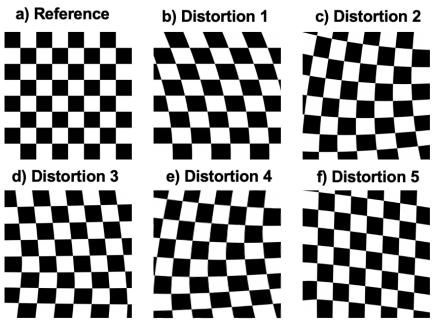

Figure 5. Visual examples of random crease distortion.

3.3.3 Total Variation Norm (TV loss)

The characteristics of natural images include smooth and consistent patches with gradual color changes within each patch [22]. Therefore, to increase the plausibility of physical attacks, smooth and consistent perturbations are preferred. Additionally, extreme differences between adjacent pixels in the perturbation may not be accurately captured by cameras due to sampling noise and non-smooth perturbations may not be physically realizable [32]. To address these issues, the total variation (TV) [22] loss is introduced to maintain the smoothness of the perturbation. For a perturbation P, TV loss is defined as:

$$L_{tv} = \sum_{i,j} \sqrt{(P_{i+1,j} - P_{i,j})^2 + (P_{i,j+1} - P_{i,j})^2} \quad (5)$$

Where i and j refer to the pixel coordinate of the patch P.

4. Evaluation of Attack Performance

4.1. Experimental Setup

As victim object detectors, we used Yolov2 [28], Yolov3 [29], Yolov3tiny, Yolov4 [2], Yolov4tiny, Yolov7 [35] and FasterRCNN [30] with an input image resolution of 416×416. For the optimization, we use Adam optimizer with a learning rate of 0.001, $\beta 1 = 0.9$ and $\beta 2 = 0.999$. In our evaluations, we use the images of the INRIA [7]. The dataset is popular in benchmarking Pedestrian Detection applications. It consists of 614 person detections for training and 288 for testing.

4.2. DAP Transferability

Our evaluation metric is the mean average precision (mAP), which is a commonly used performance measure in object detection tasks. To calculate mAP, we adopt the approach used in prior work [13, 33]: we take the detection boxes generated by each detector on the clean dataset as the ground truth boxes (assuming no adversarial patch is present), and then report the mAP when the same detectors are used to detect objects with the added adversarial patches. Table 3 presents the evaluation results on the INRIA dataset, where we trained our adversarial patches using four different detectors. The victim detectors used during testing are shown in the horizontal rows, while the detectors used during training are shown in the vertical columns. As shown in the table, our approach achieves lower mAP scores across various detector combinations, demonstrating its effectiveness and transferability. Table 3 shows that when using the same victim detector for training and testing, our attack achieves high performance (93.46% attack success rate when attacking Yolov3tiny). Furthermore, our technique exhibits strong transferability across different victim detectors (64.07% attack success

Detector	Yolov3	Yolov3tiny	Yolov4	Yolov4tiny
Yolov3	**32.63**%	37.13%	44.31%	38.08%
Yolov3tiny	35.93%	**6.54**%	43.96%	35.21%
Yolov4	50.21%	51.33%	**24.65**%	48.8%
Yolov4tiny	41.36%	26.47%	45.18%	**16.98**%

Table 3. Attack performance in terms of mAP of DAP on INRIA dataset using different detectors.

rate when attacking Yolov3 using a patch trained using Yolov3tiny). Further transferability results across different architectures are presented in the supplementary material.

4.3. DAP vs. State-of-the-Art Techniques

To assess the performance of our proposed adversarial patch, we compare it against four state-of-the-art techniques: Naturalistic Patches [13], Adversarial T-shirt [37], Adversarial YOLO [33], and Universal Physical Camouflage (UPC) [14]. Table 4 summarizes the mean average precision results of these methods on the INRIA dataset. Our proposed approach achieves competitive attack performance compared to the state-of-the-art techniques. Notably, Adversarial YOLO [33] delivers the highest attack performance for most of the detectors, but the generated patches lack realism and can be easily detected due to their noticeable appearance. In contrast, our approach generates adversarial patches that blend more naturally into the surrounding while maintaining high attack performance. Overall, our approach delivers comparable performance to that of state-of-the-art techniques, while offering a more natural and subtle visual appearance.

Detector	DAP	NAP	Adv. YOLO	UPC
Yolov2	19.51%	12.06%	2.13%	48.62%
Yolov3	32.63%	34.93%	22.51%	54.40%
Yolov3tiny	6.54%	10.02%	8.74%	63.82%
Yolov4	24.65%	22.63%	12.89%	64.21%
Yolov4tiny	16.98%	8.67%	3.25%	57.93%
Yolov7	17.72%	60.78%	N/A	N/A

Table 4. DAP vs State-of-the-Art adversarial patches **without transformations**.

4.4. Impact of Adding Non-Rigid Deformations on Patch Performance

To evaluate the robustness of different adversarial patch techniques to variations in clothing appearance, we applied random clothes deformation transformations to the INRIA dataset and measured the attack success rates. We compared the performance of our proposed DAP patch against that of the state-of-the-art Naturalistic Patches [13]. Table 5 summarizes the results. We found that the effectiveness of the NAP was drastically degraded after applying the clothes deformation transformations. For example, the success rate of the Yolov3tiny-based NAP dropped from 89.98% to 30.8%. In contrast, our DAP patch maintained its effectiveness against the deformed clothing, with only a minor drop of 8.9% in the success rate. These results suggest that our proposed DAP patch is more robust and less affected by variations in clothing appearance modeled by the creases in the patch compared to the state-of-the-art Naturalistic Patches. Our proposed patch continues to demonstrate remarkable performance on the FasterRCNN model, surpassing the capabilities of the state-of-the-art NAP method. In fact, in the presence of non-rigid transformations, our DAP approach maintains a mean average precision (mAP) of 30.60%, outperforming NAP, which experiences a decrease in performance through an increase in the mAP from 42.47% to 75.1%.

	NAP		DAP	
Detector	**w/o**	**w/**	**w/o**	**w/**
Yolov3	34.93%	77.37%	32.63%	37.70%
Yolov3tiny	10.02%	69.20%	6.54%	15.44%
Yolov4	22.63%	70.9%	24.65%	34.45%
Yolov4tiny	8.67%	62.32%	16.98%	27.21%
FasterRCNN	42.47%	75.1%	19.19%	30.60%
Yolov7	60.78%	72.46%	17.72%	36.67%

Table 5. Attack performance (mAP) of different Patches **with and without rigid and non-rigid transformations**.

4.5. Naturalness Evaluation

We conduct a user study for the naturalness evaluation of our adversarial patch. This study included 20 participants of diverse backgrounds, both male and female, aged 19 to 59. Participants were asked to provide numerical scores (ranging from 0 to 100%) to assess naturalness and absence of conspicuous patterns in presented images. Results showed that the proposed patch received the highest score, indicating superior performance in terms of naturalness. This will be included in the last version.

Patches	DAP	NAP	UPC	AdvYOLO
Scores	93.12%	47.75%	18.15%	15%

Table 6. Subjective test for the naturalness evaluation of our adversarial patch with other baselines.

5. Physical Attack Evaluations

For physical attack evaluation and to compare the effectiveness of our adversarial patch (DAP) and the naturalistic patch (NAP), both generated using Yolov3tiny,

we printed them on T-shirts in 20.5cm x 21.5cm format. We filmed two videos to showcase the effectiveness of our DAP-based T-shirt and that of the GAN-based T-shirt in fooling the Yolov3tiny object detector (The two demo videos are provided in the supplementary material). The participant is approximately three meters away from the camera. We asked the participant to move in different directions: back and forth as well as side-to-side within the duration of the videos. The participant hid the patch using his hand forming the baseline performance of the detector and also deliberately created aggressive wrinkles to the patches to assess their robustness to different transformations.

The detection results were recorded in these videos. Subsequently, we extracted frames from the videos and performed annotations using the following procedure:

- *Patch Presence (P) Annotation*: If a patch is detected in a frame, we assign a value of $P = 1$. If no patch is present, we set $P = 0$.
- *Person Detection (D) Annotation*: We determine whether a person is detected in the frame or not. If a person is detected, we assign $D = 1$. If no person is detected, we set $D = 0$.

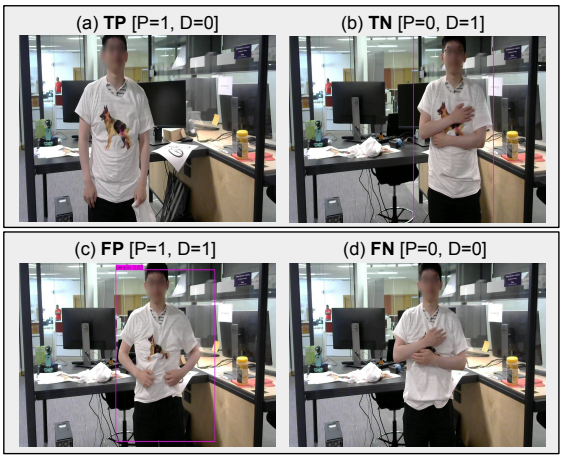

Figure 6. Example of annotated samples.

To illustrate this annotation process, Figure 6 showcases four different cases. The key metrics used to evaluate the performance of the detection system are as follows: The True Positive Rate (TPR), the False Positive Rate (FPR), the True Negative Rate (TNR), and the False Negative Rate (FNR) (Further details about these metrics and the way they were computed is provided in the supplementary material). 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.

Table 7 provides a comprehensive overview of the evaluation metrics for our adversarial patch, DAP, as well as the naturalistic Patch (NAP). These metrics highlight the performance of the patches in terms of true positive rate (TPR) and false positive rate (FPR). For instance, our DAP patch demonstrates an impressive TPR of 92.01%, indicating that in 92.01% of cases where the person was not detected was because of our patch and 7.99% because of the detector not working well. On the other hand, the NAP patch achieved a considerably lower success rate of only 57.22% of the cases the miss-detection of the person was because of the patch. Furthermore, our patch achieves the lowest false positive rate (FPR) of 8.33%, where the person was detected while the patch was visible. This contrasts with the 42.77% occurrence where the patch was visible, but it failed to escape detection. Further details and discussion can be found in the supplementary material.

Patch	TPR	TNR	FPR	FNR
DAP	**92.01**%	**59.66**%	**8.33**%	**40.33**%
NAP	57.22%	43.32%	42.77%	56.67%

Table 7. TPR, TNR, FPR, and FNR for DAP and NAP-based adversarial T-shirts

6. Discussion

6.1. Ablation Study

We conducted an ablation study for each term in our loss function.

Similarity Loss Ablation: When the similarity loss function is removed, the generated patch essentially turns into noisy, conspicuous patterns but more efficient patch achieving a mAP of 2.54% for Yolov3tiny. This underscores the vital role of the similarity loss in guiding the patch to resemble a benign image and enhancing its inconspicuousness.

Detection Loss Ablation: In the absence of the detection loss, the generated patch closely resembles the target image. As expected, this change doesn't seem to noticeably impact the detector's performance. This suggests that the detection loss plays a critical role in maintaining the patch's adversarial nature and its ability to evade detection.

Total Variation Loss Ablation: Removing the total variation loss function results in sharper color changes within the patch. This indicates that the total variation loss significantly contributes to the smoothness and natural appearance of the generated patch.

6.2. Impact of Patch Scale

We evaluate the patch performance with respect to the size of the target object (person). We use scales of $0.6, 0.5, 0.4$, and 0.3 representing the size of the patch with respect to the size of the bounding box (0.5 scale corresponds to 0.2 in [13]). Our results in Table 8 confirm that larger patches generally result in stronger attack performance, as expected.

This is due to the fact that a larger patch covers more of the target object, making it more difficult for object detectors to identify the person. Overall, our findings suggest that patch size is an important factor to consider when designing effective adversarial patches for object detection.

Scale	0.3	0.4	0.5	0.6
mAP	58.98%	37.20%	15.44%	6.54%

Table 8. Attack performance against YOLOv3tiny with adversarial patches in different scales with respect to the target object size for the INRIA dataset.

6.3. Adversarial Patches in Different Classes

Our proposed adversarial patch is not limited to targeting only the "Dog" class. In fact, we were able to successfully generate effective patches for other object classes as well, such as the "Cat" and "Bike" classes. This demonstrates the versatility and potential of our approach, which can be applied to a wide range of target patterns of patch appearance. The patches performance is summarized in Table 9. For instance, when using Bike as target pattern, and when applying creases to the generated patch we get 75.95% ASR.

	Bike Class		Cat Class	
Detector	w/o	w/	w/o	w/
Yolov3	45.46%	47.58%	38.87%	47.25%
Yolov3tiny	18.43%	24.05%	42.67%	50.05%

Table 9. DAP performance with and without transformations for 'Bike' and 'Cat' classes for Yolov3 and Yolov3tiny.

6.4. Limitations

Enhancing the interpretability and explainability of our approach is crucial for gaining deeper insights into the underlying mechanisms that contribute to its success. By understanding the specific features or characteristics that make certain target classes result in more effective patches, we can refine our approach to improve its performance across a wider range of classes. Exploring techniques such as feature visualization, attribution methods, or saliency analysis can help us identify the discriminative patterns that our approach leverages to deceive the detector. This knowledge can guide the development of more effective and generalizable adversarial patches.

7. Related Work

Initially, physical attacks aimed at fooling person detectors were generated without considering patch stealthiness, but rather focused on performance and producing effective attacks. For example, [33] proposed printable adversarial patches attached to a cardboard, [17] proposed patches trained for random placement on the scene, and [36] proposed an invisibility cloak. However, these patches had conspicuous and easily identifiable patterns. To overcome this limitation, some works proposed leveraging the learned image manifold of pre-trained GANs upon real-world images to create naturalistic patches [8, 13, 27]. Authors in [14] proposed a universal camouflage pattern that is visually similar to natural images and for stealthiness added a L_∞ norm constraint to control the adversarial noise. These attacks were aimed to be printed on a T-shirt, but they ignored non-rigid deformations caused by a moving person. Authors in [37] attempt to model these deformations using a thin plate spline (TPS) based transformer. Nevertheless, this method ignores the stealthiness of the patch and results in conspicious patterns. To address all these limitations and solve the trilemma of efficiency, stealthiness, and robustness, we propose an attack that generates a patch, DAP, that maintains naturalistic patterns, is robust to multiple transformations, and can be printed on a T-shirt while being stealthy. A comparison of DAP with state-of-the-art attacks is provided in Table 10.

Attack	Robustness Techniques	Stealthiness Techniques	Object	Space
[33]	EOT, TV, NPS	N/A	Static	2D
[37]	EOT, TPS	N/A	Dynamic	2D
[14]	EOT, TV	L_∞ norm	Static	3D
[13]	TV	GAN	Static	2D
Ours	EOT, TV, CT	L_{sim}	Dynamic	2D

Table 10. DAP vs State-of-the-Art adversarial patches in terms of robustness and stealthiness techniques, targeted objects, and the Space (EOT: Expectation Over Transformations, TV: Total Variation, NPS: Non-Printability Score, TPS: Thin Plate Spline, CT: Creases Transformation).

8. Conclusion

This paper presents a novel approach to generate naturalistic physical adversarial patches for object detectors. Our proposed framework overcomes the limitations of GAN-based approaches including the limited latent space, producing stealthy patches that achieve competitive attack performance. Furthermore, we introduce a creases transformation blocks to model non-rigid deformations aimed for dynamic objects. Our approach results in an effective, stealthy, and robust adversarial patch.

Acknowledgment

This work was supported in parts by the NYUAD Center for Cyber Security (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104.

References

[1] M. Al-Qizwini, I. Barjasteh, H. Al-Qassab, and H. Radha. Deep learning algorithm for autonomous driving using googlenet. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 89–96. IEEE, 2017. 1

[2] Hong-Yuan Mark Liao Alexey Bochkovskiy, Chien-Yao Wang. Yolov4: Yolov4: Optimal speed and accuracy of object detection. *arXiv*, 2020. 5

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 4

[4] Tao Bai, Jinqi Luo, and Jun Zhao. Inconspicuous adversarial patches for fooling image-recognition systems on mobile devices. *IEEE Internet of Things Journal*, 9(12):9515–9524, 2022. 2

[5] Kamel Boudjit and Naeem Ramzan. Human detection based on deep learning yolo-v2 for real-time uav applications. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(3):527–544, 2022. 1

[6] N. Carlini and D. A. Wagner. Towards evaluating the robustness of neural networks. *CoRR*, abs/1608.04644, 2016. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893 vol. 1, 2005. 5

[8] Bao Gia Doan, Minhui Xue, Shiqing Ma, Ehsan Abbasnejad, and Damith C Ranasinghe. Tnt attacks! universal naturalistic adversarial patches against deep neural network systems. *IEEE Transactions on Information Forensics and Security*, 17:3816–3830, 2022. 8

[9] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on machine learning models. *CoRR*, abs/1707.08945, 2017. 1

[10] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. In *Proceedings of the 12th USENIX Conference on Offensive Technologies*, page 1, USA, 2018. USENIX Association. 2

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 1

[12] Amira Guesmi, Muhammad Abdullah Hanif, and Muhammad Shafique. Advrain: Adversarial raindrops to attack camera-based smart vision systems. *arXiv preprint arXiv:2303.01338*, 2023. 1

[13] Yu-Chih-Tuan Hu, Jun-Cheng Chen, Bo-Han Kung, Kai-Lung Hua, and Daniel Stanley Tan. Naturalistic physical adversarial patch for object detectors. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7828–7837, 2021. 1, 2, 5, 6, 7, 8

[14] Lifeng Huang, Chengying Gao, Yuyin Zhou, Changqing Zou, Cihang Xie, Alan L. Yuille, and Ning Liu. UPC: learning universal physical camouflage attacks on object detectors. *CoRR*, abs/1909.04326, 2019. 2, 6, 8

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 3

[16] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. 1

[17] Mark Lee and J. Zico Kolter. On physical adversarial patches for object detection. *CoRR*, abs/1906.11897, 2019. 8

[18] B. Li and Y. Vorobeychik. Scalable optimization of randomized operational decisions in adversarial classification settings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*. JMLR.org, 2015. 1

[19] Jianyi Liu, Yu Tian, Ru Zhang, Youqiang Sun, and Chan Wang. A two-stage generative adversarial networks with semantic content constraints for adversarial example generation. *IEEE Access*, 8:205766–205777, 2020. 2

[20] Jiajun Lu, Hussein Sibai, Evan Fabry, and David A. Forsyth. NO need to worry about adversarial examples in object detection in autonomous vehicles. *CoRR*, abs/1707.03501, 2017. 4

[21] Yan Luo, Xavier Boix, Gemma Roig, Tomaso A. Poggio, and Qi Zhao. Foveation-based mechanisms alleviate adversarial examples. *ArXiv*, abs/1511.06292, 2015. 4

[22] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 5

[23] Akshay Mangawati, Mohana, Mohammed Leesan, and H. V. Ravish Aradhya. Object tracking algorithms for video surveillance applications. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 0667–0671, 2018. 1

[24] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018. 1

[25] A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. 1

[26] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. *CoRR*, abs/1511.07528, 2015. 1

[27] Svetlana Pavlitskaya, Bianca-Marina Codău, and J. Marius Zöllner. Feasibility of inconspicuous gan-generated adversarial patches against object detection, 2022. 2, 8

[28] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. 5

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018. 5

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 5

[31] Alberto Sabater, Luis Montesano, and Ana C Murillo. Robust and efficient post-processing for video object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10536–10542. IEEE, 2020. 1

[32] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540. ACM, 2016. 5

[33] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. *CoRR*, abs/1904.08653, 2019. 2, 5, 6, 8

[34] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021. 3

[35] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022. 5

[36] Zuxuan Wu, Ser-Nam Lim, Larry S. Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *ECCV*, 2020. 8

[37] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, 2019. 2, 6, 8

[38] 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. page 1989–2004, New York, NY, USA, 2019. Association for Computing Machinery. 2