



Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World

Jia Tan*

Xiangtan University
Qian Xuesen Laboratory of Space Technology, China
Academy of Space Technology
201821511197@smail.xtu.edu.cn

Nan Ji*

Qian Xuesen Laboratory of Space Technology, China
Academy of Space Technology
jinan@qxslab.cn

Haidong Xie

Qian Xuesen Laboratory of Space Technology, China
Academy of Space Technology
xiehaidong@qxslab.cn

Xueshuang Xiang[†]

Qian Xuesen Laboratory of Space Technology, China
Academy of Space Technology
xiangxueshuang@qxslab.cn

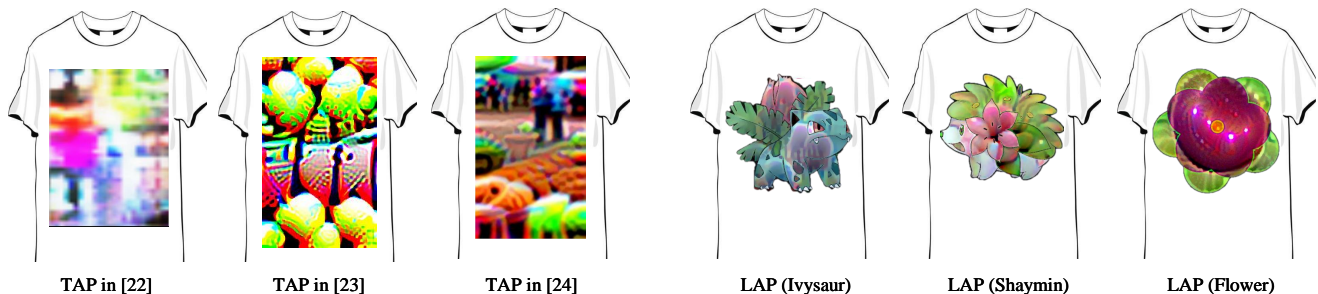


Figure 1: T-shirts with traditional adversarial patches (TAP) in [22–24] and our legitimate adversarial patches (LAP) generated from cartoon images Ivysaur, Shaymin and Flower. Adversarial patches are commonly used physical attacks to evade deep learning based object detection models. Obviously, the TAPs are unnatural and could not evade human eyes. Our LAPs look much natural and are hard to distinguish from general T-shirts.

ABSTRACT

It is known that deep neural models are vulnerable to adversarial attacks. Digital attacks can craft imperceptible perturbations but lack of the ability to apply in physical environment. To address this issue, efforts have been investigated to study physical patch attacks in the physical world, especially for object detection models. Previous works mostly focus on evading the detection model itself but ignore the impact of human observers. In this paper, we study legitimate adversarial attacks that evade both human eyes and detection models in the physical world. To this end, we delve into the issue of patch rationality, and propose some indicators for

evaluating the rationality of physical adversarial patches. Besides, we propose a novel framework with a two-stage training strategy to generate our legitimate adversarial patches (LAPs). Both in numerical simulations and physical experiments our LAPs have significant attack effects and visual rationality.

CCS CONCEPTS

• **Security and privacy** → *Social aspects of security and privacy.*

KEYWORDS

legitimate adversarial patches, human detection system, physical world

ACM Reference Format:

Jia Tan, Nan Ji, Haidong Xie, and Xueshuang Xiang. 2021. Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3475653>

1 INTRODUCTION

Despite their remarkable performance on many challenging tasks, deep learning models have shown to be vulnerable to adversarial

*Both authors contributed equally to this research.

[†] Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3475653>

attacks[21]. The So-called adversarial attacks are some artificial data that adding crafted tiny perturbations to images that are imperceptible to human eyes, but causing the deep learning model to output a wrong result. Current works generally focus on adversarial attacks in digital domain, like L_2 or L_∞ attacks that craft full-image perturbations injected into the input images[2, 7, 14], and L_0 attacks that modify some pixels after searching the full image extent[15]. While such attacks can confuse human eyes, they can be hardly transferred to physical world. That is due to the difficulty of making changes to entire pictures captured by the attacked target model and the lack of simulating physical environment (e.g., light, angle and noise).

To address the above issue, efforts have been investigated to study how to construct physical attacks in the real world, and patch attacks are commonly used means among them. Typically, this type of attacks work by altering the visible characteristics of an object, such as eyeglasses[18], mask[9], hat[10] and T-shirt [20, 23]. Though prior works have revealed the vulnerability of deep learning models to adversarial perturbations in the real world, but most of them do not adequately account for the semantics and the generated patches are unnatural and spotted easily to human observers (see Figure 1). **This discovery inspires us to re-examine the qualities that a qualified physical adversarial patch should have, i.e., evading both detection models and human eyes.** Similar to the attacks in the digital domain, the so-called evading human eyes is actually a rationality constraint, i.e., the physical adversarial patches should be closer to the patterns existing in the real world.

In this paper, we study the rationality of adversarial patches in the object detection model, specifically human detection systems. To this end, we first propose some rational indicators to analyze the rationality from three aspects of color features, edge features, and texture features. Then, we construct a novel framework to generate legitimate adversarial patches (LAPs). Unlike former works without semantics, considering that the patch needs to be attached to the clothes, we encourage the visual resemblance between generated patterns with people's familiar clothing decorations. In this paper, we choose cartoon pictures with natural perceptual properties and leverage a projection function as an additional optimization. Here, "projection" refers to keeping the edge features of cartoon pictures while approximating their color features. To further balance the attack effect and rationality of patches, we propose some training strategies for selection of initial inputs and paste methods. As shown in Figure 1, the LAPs are visually similar to cartoon images of Ivysaur, Shaymin and Flower, which look natural for human observers.

We demonstrate the performance of LAPs to evade both human eyes and detection models with a series of experiments. Our results on YOLOv2 show that our generated patches achieve considerably obvious attack effect, while being more legitimate compared to related work in terms of the rational indicators.

We summarize our contributions as follows:

- To the best of our knowledge, we first re-examine the rationality of patch attacks, i.e., LAPs should evade both human eyes and detection models.
- We construct a general optimization framework for generating LAPs and propose some rationality indicators for evaluating the rationality of physical adversarial patches.
- We conduct experiments in both digital and physical worlds. We find that with little decrease of aggression, our LAPs can greatly improve the rationality of the patterns, which shows great potential to avoid human eye surveillance. In addition, we verify the feasibility of LAPs in the physical world.

2 RELATED WORK

Object detection models Deep learning for object detection is a popular research problem because of its broad applications. Modern object detectors based on deep learning methods can be classified into two categories: two-stage strategy detectors such as R-CNN[17], and one-stage strategy detectors such as YOLOv2[16]. Likewise, the security of object detection models in practical applications attracts a great deal of attention. The adversary is required to attack object detectors by either hiding the object from being detected, or fooling the detector to output the target label. A well-known success of such attacks in the physical world is the generation of adversarial stop sign[3, 12, 19].

Patches for human detection systems We focus more on human detection systems in this paper. Recently, Thys et al.[20] demonstrate that a person can evade a detector by holding a card-board with an adversarial patch. On the basis of this work, Wu et al.[22] create wearable adversarial clothing, and consider training patches that fool an ensemble of detectors. In the same year, Xu et al.[23] develop a TPS-based transformer to model the temporal deformation of an adversarial T-shirt caused by pose changes of a moving person. However, these works mentioned above cannot be effectively applied to a supervised setting since they only focus on simulating external environment, e.g. lighting or viewpoint, but do not consider the impact of human visual senses.

Rationality of patches There are actually some studies on the rationality of adversarial patches. Liu et al.[11] use an attention model and GAN to generate visually natural patches with strong attacking ability. And Duan et al.[5] use neural style transfer for similar purposes. However, they both generate specific adversarial patches for each image classification task, which does not correspond to realistic applications. Different from the above work, our method aims to implement the rationality of adversarial patches and constructs a universal legitimate pattern that can be used in all kinds of human detection pictures.

3 RATIONALITY INDICATORS

The rationality of patch images, i.e., what kind of patches appear semantic and look natural to human observers, is actually a very subjective concept. To better evaluate a legitimate patch, we try to analyze its rationality quantitatively. Here we attempt to carry out the analysis of patch rationality from three aspects: color features, edge features, and texture features.

Color features The structure of RGB space does not conform to human subjective perception for colors and HSL space have better visual consistency, which are more suitable for human judgment to the naked eyes[13]. We discuss the color features of patches in HSL space, whose three components represent hue, saturation, and

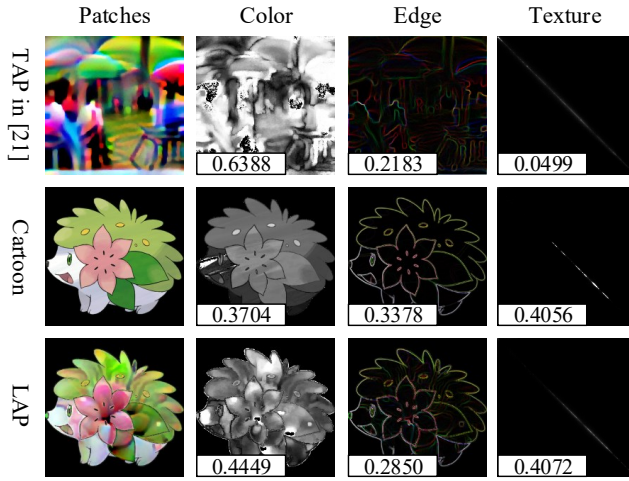


Figure 2: Comparison of the rationality for different patches which is justified by a combination of color features, edge features, and texture features. The first column contains the TAP in [20], the cartoon image Shaymin, and our LAP generated from Shaymin. The next three columns correspond to saturation maps, gradient maps and gray-level co-occurrence matrix maps whose indicators named in turn as ASI, AGI, ADE are noted in the lower left of each map.

lightness, respectively. Saturation refers to the purity of color, the higher the saturation, the purer the color, and the lower the closer to gray.

We propose a color saturation-based indicator, Average Saturation Intensity (ASI), to indicate the color rationality of different patches. Let the color of RGB be (L_R, L_G, L_B) , whose value ranges in $[0, 1]$. L_{max} and L_{min} are the maximum and minimum values of L_R, L_G, L_B , respectively. then ASI is calculated as:

$$ASI = \frac{\sum_{i,j} S_{i,j}}{\sum_{i,j} H(L_{max})} \quad (1)$$

Where S is the saturation value and $H()$ is piecewise function that defined as $H(t) = \{0, \text{if } t \leq 0; 1, \text{if } t > 0\}$.

The second column of Figure 2 shows the saturation maps of three patches (e.g. TAP in [20], cartoon image Shaymin and our generated LAP). We can see that the saturation maps of TAP has a lot of white parts, which means it is highly saturated in color, and harsh to human eyes. a legitimate patch tends to have a low ASI, indicating that its colors are not too vibrant, which is in line with our real-life visual perception tendencies.

Edge features The gradient map is a commonly used method to detect image edge[1] and patch images with significant semantic information often have distinct outlines.

Based on the above observations, we propose an indicator to portray the edge features in patch images, named as Average Gradient

Intensity (AGI). We can describe it as follows:

$$AGI = \frac{\sum_{i,j} g_{i,j} * H(g_{i,j} - \theta)}{\sum_{i,j} H(g_{i,j} - \theta)} \quad (2)$$

where g is the gradient map of the patch image, the threshold θ is used to filter out the blurred edge contours and $H()$ is a piecewise function for calculating the number of pixels beyond θ in value.

The third column of Figure 2 shows gradient maps of three images with $\theta = 0.1$. Compared to TAP in [20], the edge contours of cartoon Shaymin tend to be stronger, containing obvious visual perceptual information. The AGI value noted in the lower left corner of each patch image demonstrates that a legitimate patch tends to have a larger AGI value, which means it has a clean and clear edge profile.

Texture features Gray-level co-occurrence matrix[8], abbreviated as GLCM, is a common method for describing textures by studying the spatial correlation properties of grayscales. Since texture is formed by the recurrence of grayscale distribution in spatial locations, there will be a certain grayscale relationship between two pixels separated by a certain distance in the image space, i.e., the spatial correlation property of grayscale in the image.

GLCM energy[8] reflects texture coarseness and high energy occurs when the distribution of gray level changes is uniform and regular, which fits our need for a legitimate patch. We choose the statistics calculated based on GLCM energy as the indicator for texture features of patches, named Average Directional Energy (ADE):

$$ADE = \frac{1}{n} \sum_{k=1}^n \sqrt{\sum_{i,j} (G_{i,j}^k)^2} \quad (3)$$

where $\sqrt{\sum_{i,j} (G_{i,j}^k)^2}$ is GLCM energy with the direction k . ADE measures the multi-directional average stability of patch texture grayscale variation and reflects the multi-directional coarseness of texture.

The fourth column of Figure 2 shows the GLCM maps when we set $n = 4$ with angles $0, \pi/4, \pi/2$, and $3\pi/4$. The texture rule transformation of cartoon Shaymin and LAP tend to be more stable, so their ADE values tend to be larger as well.

4 METHODOLOGY

4.1 Framework

Our goal is generating LAPs that assist to hide a person from being detected by both object detectors and human observers in the physical world. In this paper, we describe the problem in general terms in the following way:

$$\min_p \mathbb{E}_{(x,y) \sim \mathcal{D}, t \sim T} L(A(x, M(p), t), y) \quad (4)$$

where \mathcal{D} is a distribution over samples, T is a distribution over patch transformations to simulate the various influences in the physical world, and A is a patch application function that transforms the masked patch $M(p)$ with t and applies the result to the image x . The definition of $M(p)$ is showed below:

$$M(p) = p - \text{Grad}(p, \theta) - Bg(p) \quad (5)$$

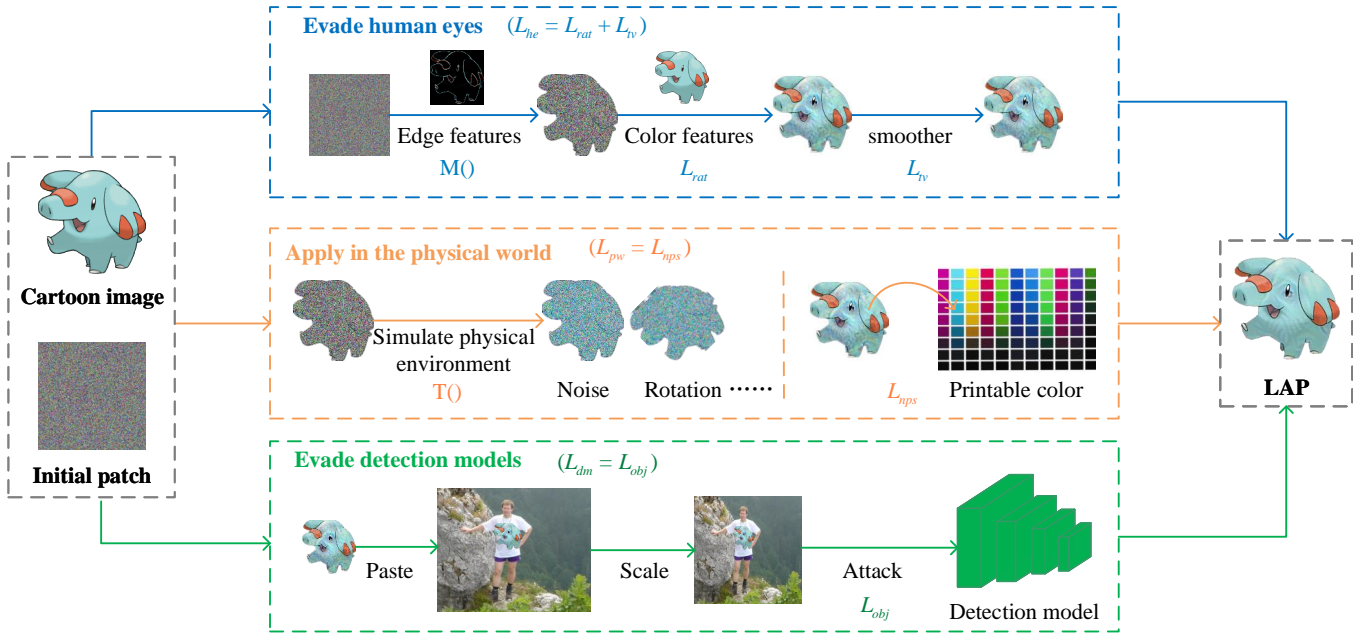


Figure 3: Overview of the framework to generate LAPs with the cartoon image and initial patch as inputs. Evading human eyes, applying in the physical world and evading detection models together enable the effectiveness of LAPs. The edge feature, color feature and smoother are considered in part A. Simulations for physical environment and printability are center roles in part B. In part C, we take detector attacks in consider.

where p is the original cartoon image, $Grad(p, \theta)$ is the gradient map of p with θ as the threshold, and $Bg(p)$ is the black background outside the outline of p . $M(p)$ is the first step in the training process to manipulate the inputs, so all patches in the following formulas are masked by $M()$ and we set $q = M(p)$.

Figure 3 shows the framework of our paper to generate LAPs with the cartoon image and initial patch as inputs. The function to be optimized in (4) can be described as follows:

$$L = \underbrace{\alpha L_{rat} + \beta L_{tv}}_{\text{human eyes}} + \underbrace{\gamma L_{nps}}_{\text{physical world}} + \underbrace{L_{obj}}_{\text{detection models}} \quad (6)$$

where α , β and γ are parameters that control the weight of different optimization items. The design of optimization function reflects our thinking on the problem of patch attacks from three aspects: evade human eyes ($L_{he} = L_{rat} + L_{tv}$), apply in the physical world ($L_{pw} = L_{nps}$) and evade detection models ($L_{dm} = L_{obj}$).

4.2 Optimization function

Evade human eyes Inspired by the imperceptible perturbation constrained in the digital attacks, we use L_{rat} and L_{tv} to encourage the generated patch look similar to the cartoon image. Specifically, the gradient map of the cartoon image is used to keep the edge features clearly. We further enhance the semantics by decreasing the color space distance between the patch and the cartoon image and highlighting the texture information. In summary, our objective

of evading human eyes can be expressed as

$$L_{rat} = \mathcal{L}(q, c) = \sqrt{\sum_{i,j} (q_{i,j} - c_{i,j})^2} \quad (7)$$

$$L_{tv} = \sum_{i,j} \sqrt{(q_{i,j} - q_{i+1,j})^2 + (q_{i,j} - q_{i,j+1})^2}. \quad (8)$$

where q is the masked patch, c is the cartoon image and \mathcal{L} is the Euclidean distance loss. L_{rat} constrains the color features similarity and L_{tv} stands for total variation loss and it is important for smoothing the patch and reducing the noise.

Apply in the physical world To keep the generated adversarial patterns more robust in the physical world, we use a series of transformations to simulate the external environment, such as light, noise and angle, as described in the center second box of Figure 3. To effectively handle the difference between digital images and printed objects, we also introduce additional measurement function as follows:

$$L_{nps} = \sum_{q_{pixel} \in q} \min_{c_{print} \in C} \|q_{pixel} - c_{print}\|_2, \quad (9)$$

where q_{pixel} is a pixel in the masked patch q , the c_{print} is a color in a set of printable colors C . L_{nps} is the non-printability loss that represents how well the colors in adversarial patches can be implemented by a common printer.

Evade detection models We attack YOLOv2 [16], a one-stage strategy detector, under white-box settings in this paper. The detector divides the input image into multiple grids and predicts bounding boxes, confidence score and class probabilities in a single

step. We propose to craft a universal pattern to fool YOLOv2 by lowering confidence score, i.e. reducing the probability that boxes contain persons.

YOLOv2 predicts N boxes for each input image. We denote the output boxes of each images x as $\mathcal{B} = \{b_i | b_i = (\vec{d}_i, s_i, \vec{c}_i), i = 1, 2, \dots, N\}$, where \vec{d}_i represents the coordinates of i -th bounding box, s_i is the confidence score of i -th bounding box and \vec{c}_i is the conditional class probability. We define the objective function to attack the detector as follows:

$$L_{obj} = \sum_{b_i \in \mathcal{B}} Top_k(s_i), \quad (10)$$

where $Top_k(s_i)$ presents top- k confidence coming from k boxes ordered by their confidence scores and k is the upper bound on the number of persons we can think of in a image. By minimizing L_{obj} , our goal is lowering the valid boxes that containing persons and assisting for hiding a person from the detector YOLOv2.

4.3 Training strategy

Initial patches Our physical adversarial attack against human observers and detection models is a difficult task because of the combination of attack effectiveness, environmental robustness, and visual rationality. We make some attempts on how to achieve a better attack under these conditions and find that the choice of initial patch has a great impact. Different initial patches tend to generate different looking adversarial patches with different attack effects. The specific comparison experiments are described in Section 5.2.

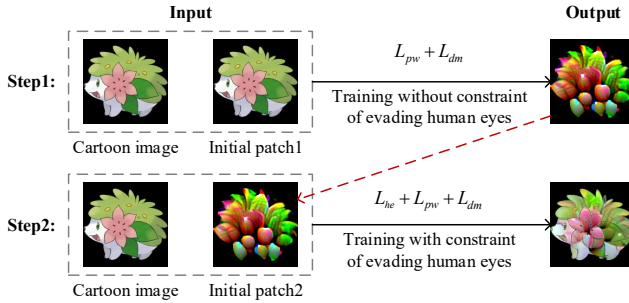


Figure 4: The two-stage training process we use to generate our LAPs. The first stage uses Shaymin as input cartoon image and initial patch, and the second stage corresponds to the input of Shaymin and transition patch, which output from the first stage, to finally output our LAP.

Due to the limitations of visual similarity, as described in (7), it was natural to choose the cartoon image itself as initial patch. Our training process divide into two steps in order to generate the more attack effective and visual rational patches. As shown in Figure 4 step1, we uses Shaymin as the input cartoon image and initial patch, and temporarily disregard the human eyes loss for enhancing the attack effects. In the step2, the transition patch output from the first step is used as the new initial patch, and finally output our LAP with constraint of evading human eyes.

We can see from the Figure 4 that the finally output LAP looks similar to the cartoon image and its attack effect is also very significant which is illustrated in Section 5.2.

Actually, gray and random images are commonly used initial patches and previous work have proved their effectiveness. When adding the consideration of rationality, adversarial patches generated with initial gray and random images do not attack well, while better attack effect is achieved with the initial cartoon images. Then we use the transition patch generated from the first step as a secondary initial, and the generated LAP is both adversarial and rational.

Paste method Another point of interest in our training strategy is the way the adversarial patches pasted on sample images. In fact, regarding how to paste the adversarial patch to human images during the training process, the operation in [20][22] is to pad the different image data into squares and paste the patch without scale (See the left image in Figure 5).

However, in real world applications the object detection model receives the adversarial sample and scales it directly to a square, which causes the patch to be scaled along with it. On the other hand, padding operation itself significantly reduce the accuracy of detection model, which makes the patch itself much less attack effective in physical applications.

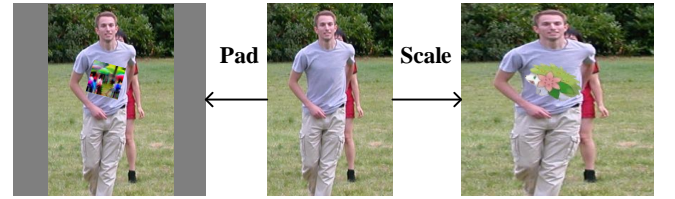


Figure 5: Comparison of pad paste method[20, 22] and scale paste method.

In our training framework the adversarial samples are directly scaled to squares required by the detection model and the patch is equally scaled as shown in the right image of Figure 5. This scale operation makes it more difficult to optimize the adversarial patch because of extra deformation, resulting in a less intense attack compared to pad method for numerical experiments, but a better physical application.

5 EXPERIMENTAL RESULTS

In this section, we empirically demonstrate the effectiveness of our proposed LAPs in both digital and physical worlds.

5.1 Setup

We use the images of the Inria dataset[4] for experiments. These images are targeted more towards full body pedestrians, which are better suited for our surveillance camera application. We evaluate the performance of LAPs to attack YOLOv2[16] trained on the Pascal VOC 2007+2012 [6] dataset. The detection minimum threshold is set as 0.7 for YOLOv2 by default.

The default optimizer is Adam, and the learning rate is initialized as 0.03 decayed by a factor of 0.1 every 50 epochs. We use the input cartoon image and initial patch of size $3 \times 300 \times 300$, and the patch is dynamically scaled by the transform function during the forward pass, where it is rotated up to 10 degrees each way.

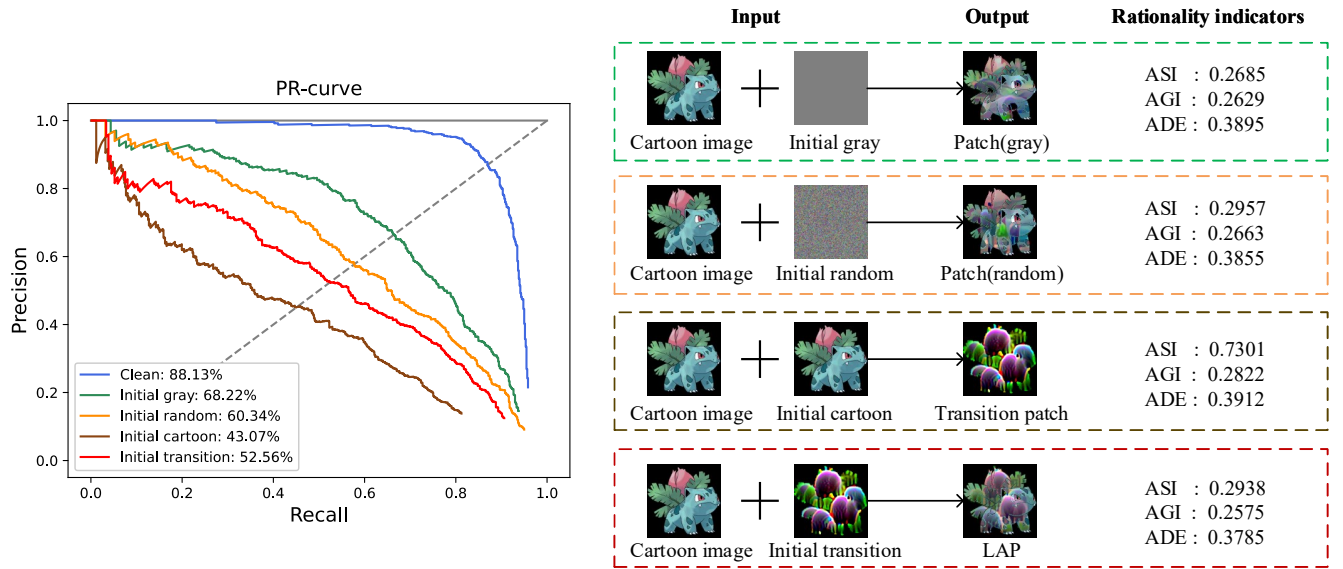


Figure 6: Comparison of attacking effects(left) and rationality(right) with different initial patches of gray, random, cartoon and transition. The lower the average precision (AP) in the PR-curve, the better the attack effect. The corresponding color boxes in the right image show the input-to-output process.

5.2 Attack in digital world

Comparison of initial images As we discussed in the training strategy part, different initial patches have a significant impact on both the rationality and attack effect of final generated adversarial patches. In order to verify this impacts, we keep the parameter α and β constant respectively as 0.0001 and 0.2, and choose gray, random, cartoon and transition patch as different initial inputs.

Figure 6 shows the generated adversarial patches and their attack effects with different initial inputs. As we can see in the left image, the adversarial patches generated initially for gray and random images demonstrate poor performance, which only make the average precision (AP) drop to 68.22% and 60.34% with both loss functions converged. Actually, it is a big difference from previous work, and we think it's may due to rationality constrains that lead to the lost of attack ability. on the contrary, the adversarial patches generated with the cartoon image and the transition patch as inputs demonstrate their good results, making the AP decrease to 43.07% and 52.56%. The better performance gives me a good choice of initial images for training

From the perspective of attacking ability, we can conclude that the cartoon image and the transition patch are better choices as initial inputs. Next, we make further analysis and comparison from the perspective of rationality.

Figure 6 shows the rationality indicators of adversarial patches generated with different initial inputs in the right part. We mostly focus on the results of the latter two patches. It can be seen that the performances of both patches on AGI and ADE are basically the same compared with the big gap in the ASI indexes. Obviously, the transition patch case outperforms the cartoon case by a large margin (0.2938 vs 0.7301). Actually, this observation can also be clearly seen in the transition patch and LAP image in Figure 6(the

middle part of the brown and red boxes). The color of transition patch is too saturated, and its similarity to the input cartoon image is not good enough.


To sum up, by comprehensively considering the influence of different initial patch on the attack ability and rationality of adversarial patches, in this paper, we use a two-stage training process to generate LAPs and this method is adopted in the following experiments.

Rationality vs attack effect To investigate the relationship between rationality and attack effectiveness for LAPs, we compare the performance with three different input cartoon images under the following setting (Table 1): (1) parameter β and γ are fixed in (6) at 0.2 and 0.01; (2) only the degree parameter of rationality is modeled, i.e., α is changed; (3) the selected three typical cartoon images sorted as Ivysaur, Shaymin and Flower, have gradually decreasing outline complexity.

The performance of 5-pattern scheme is recorded in Table 1, and the implications have the following aspects. First, we can see that aggression is at odds with rationality in general. From a visual perspective, as the degree of rationality α increases, the generated patches become more and more visually similar to the original cartoon images. Especially in the case of cartoon Shaymin that the picture with $\alpha = 0$, the transition patch generated in the first step, lacks visual perceptual rationality, while as α increases it is very close to a cartoon image.

From a data point of view, the corresponding aggression is generally declining. However, this is not the case for a simple pattern such as cartoon Flower, where the AP value decreases and then increases as α increases. This means that increasing a little L_{rat} constraint instead enhances the attack effect, while the attack effect gradually decreases when the constraint is too large. We think it

Table 1: Comparison of attack capability and rationality indicators under varying degrees of rationality weight α . Ivysaur, Shaymin and Flower are used as different input cartoon images, and the two-stage training strategy is adopted to generate LAPs for different α .

$\alpha(10^{-3})$	Cartoon Ivysaur					Cartoon Shaymin					Cartoon Flower				
	0	0.1	0.5	0.7	∞	0	0.1	0.5	0.7	∞	0	0.1	0.5	0.7	∞
AP(%)	43.07	52.56	64.24	77.62	83.77	49.05	50.59	56.50	58.64	72.87	47.14	37.34	36.15	41.10	82.32
ASI(10^{-2})	73.01	29.38	28.48	32.95	34.49	83.39	44.49	41.55	40.15	37.04	92.58	78.00	65.55	64.16	67.78
AGI(10^{-2})	28.22	25.75	26.61	28.16	29.98	22.90	28.50	30.66	32.10	33.78	23.85	26.10	22.22	21.91	39.44
ADE(10^{-2})	39.12	37.85	37.64	38.06	35.79	45.43	40.72	40.47	40.48	40.56	42.52	36.56	31.85	31.86	55.99
LAP															

is because the images with simple patterns are easy to generate the adversarial patches with voids in the first stage of optimization, and the attack effect is enhanced after adding a small amount of L_{rat} constraint instead to fill the voids.

Second, compared with AGI or ADE, the ASI changes significantly with the change of α . Specifically, the value gap of ASI is the greatest with or without reasonableness ($\alpha = \infty$ or $\alpha = 0$), and the constrains for rationality ($0 < \alpha < \infty$) have significantly improved the indicators.

Besides, the AGI and ADE values of the TAP in Figure 2 are only 0.2183 and 0.0499, but that of patches in Table 1 all reach a higher level because of the edge constraint of $M()$. Obviously the strength of L_{rat} has a greater impact on ASI values, and $M()$ has a more significant effect on AGI and ADE values.

Third, for different initial input images, the appropriate parameters to weigh attack power and rationality are different. Ideally, we would certainly choose the one with the smallest AP and ASI, and the largest AGI and ADE. Due to the contradiction between attack and rationality and the less impact of α changes for AGI and ADE as analyzed above, we focus on the trade-off between AP value and ASI value. Specifically, for patches that weaken quickly but have little change in rationality, we give priority to the former. So in the case of cartoon Ivysaur and cartoon Shaymin, we choose $\alpha = 0.1 \times 10^{-3}$. Otherwise, we consider the opposite direction and choose $\alpha = 0.5 \times 10^{-3}$ for the case of cartoon Flower. The index values for the most legitimate adversarial patches are marked in bold in Table 1. We can also find by cross-sectional comparison that the simpler the pattern, the better the attack effect of LAP generated based on this, and here LAP(Flower) reduces the AP at most from a clean 88.13% to 36.15%.

Comparison with previous work In this part, we further compare our LAPs with the previous work [20] in paste methods and rationality constrains. The paste methods represent pad and scale method as we discussed in Figure 5 and rationality constrains mean the edge and color constrains demonstrated in Figure 3.

Figure 7 shows the results for comparing the TAP with pad method (pad TAP), the TAP with scale method (scale TAP) and our LAP with scale method (scale LAP). By comparing pad TAP and scale TAP, we can see that pad method could significantly weaken the capability of the detector, reducing the clean AP from 88.13%

to 55.03%. This means that the detection ability of the detector on padded images (images with two gray bars) is insufficient, which also results in a significant attack effect for the pad TAP, starting from a position closer to the end after all. The scale method has no effect on paste results and the scale TAP performs a little better than the scale LAP (34.33% vs 36.15%) in attack results. However, the scale LAP demonstrates much more natural as shown in the last column.

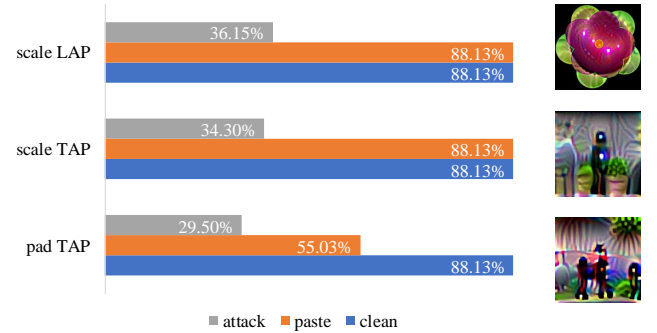


Figure 7: Comparison with previous work with different paste methods (pad or scale) and rationality constrains (with or without rationality for LAP and TAP). Clean represents the AP for Inria test dataset, paste represents the AP for Inria test dataset with pad or scale transformation method, and attack means the detection AP under corresponding AP (the last column) attacks.

5.3 Attack in the physical world

We next evaluate our proposed method in the physical world. We use Thermal Dye Sublimation to print three LAPs (left part in Figure 8) onto T-shirts of quick-drying material to attack object detector YOLOv2. It is worth noting that white T-shirts are used as the carrier of the patch because they are easy to wear and convenient for physical attacks. In fact, having the patch printed directly on the chest can also be used to attack.

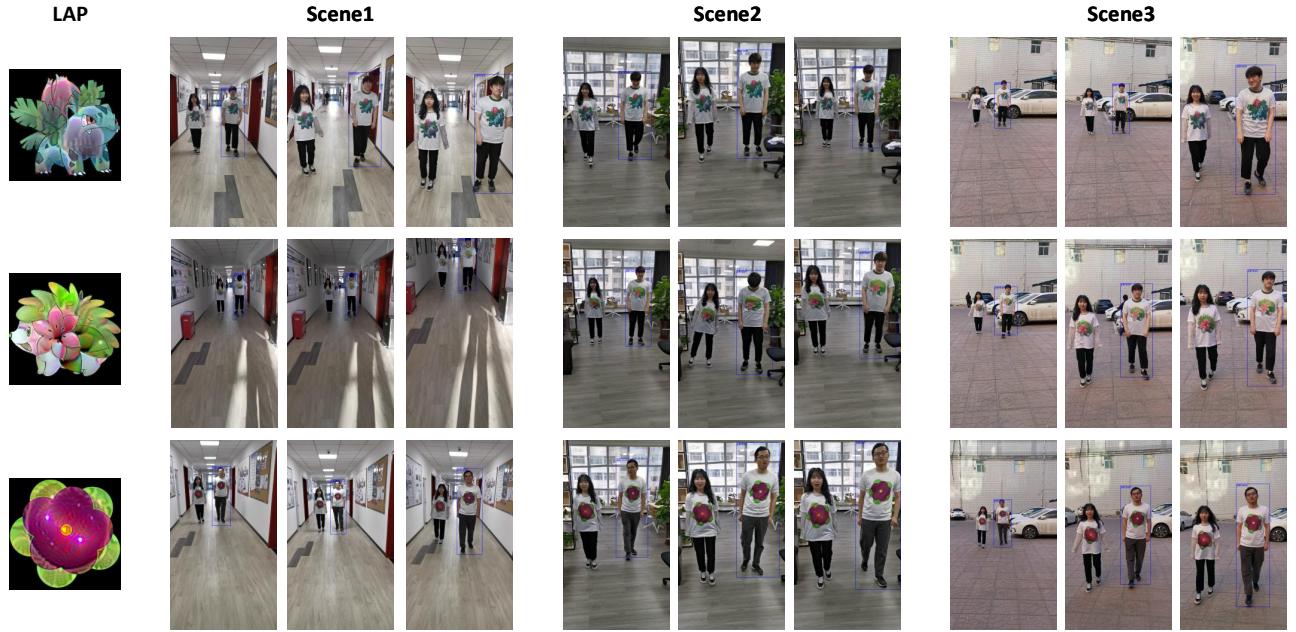


Figure 8: Display of physical-world attacks. We test the effectiveness of our LAPs (the first column) printed on white T-shirts in three scenes. Each row in any scene corresponds to a specific LAP and each picture shows an individual frame in a video. All frames are performed by two persons wearing T-shirts with the corresponding LAP and its cartoon input.

During concrete operation, we also print the corresponding original cartoon images for fair comparison. To collect material for testing, we use Huawei Mate20 to record videos for tracking two moving persons wearing T-shirts with LAPs(left) and cartoon images(right) in three different scenarios (corridor, indoor and outdoor).

In Figure 8, we demonstrate our physical-world attack results in nine groups of comparative experiments corresponding the three LAPs selected and three scenarios in our paper. We can see that in all frames with different lighting, distance and angles, the detector can recognize persons who are decorated with cartoon pictures, implying the strong detection ability of YOLOv2 and the ineffectiveness of cartoon pictures. On the contrary, our method outperforms cartoon images, assisting for hiding persons from the detector YOLOv2.

6 CONCLUSION

Current physical patch attacks do not adequately account for the semantics and the generated patches are unnatural and spotted easily to human observers. In this paper, we study legitimate adversarial attacks that evades both human eyes and detection models in the physical world. Specifically, we propose a novel framework to generate LAPs with visual similarity to cartoon pictures. To achieve the above purpose, we use a leverage a projection function to constraint the edge features and color features. In order to balance the effectiveness and rationality of patters, we introduce a two-stage training process to initialize input patches. We additionally impose some rationality indicators to quantify the tendency of conflicts

between attack capability and rationality. The experimental results show that with little decrease in attack ability, our proposed LAPs look much more natural compared with previous work. Further, we also carry out several groups of physical attacks to verify the effectiveness of our LAPs. The successful application of LAPs in the physical world exposes the potential security risks of deep learning models when applied in the real world. In the future, we call on scholars to research defenses against naturally realistic physical adversarial attacks like LAPs.

ACKNOWLEDGMENTS

This work was supported in part by Beijing Nova Program of Science and Technology under Grant Z191100001119129, and in part by National Natural Science Foundation of China under Grant No. 12004422.

REFERENCES

- [1] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.
- [2] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.
- [3] Shang Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. 2019. *ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector: Recognizing Outstanding Ph.D. Research*.
- [4] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.
- [5] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. 2020. Adversarial Camouflage: Hiding Physical-World Attacks with Natural

- Styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1000–1008.
- [6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.
 - [7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
 - [8] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. 1973. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics* 6 (1973), 610–621.
 - [9] Edgar Kaziakhmedov, Klim Kireev, Grigori Melnikov, Mikhail Pautov, and Aleksandr Petiushko. 2019. Real-world attack on MTCNN face detection system. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*. IEEE, 0422–0427.
 - [10] Stepan Komkov and Aleksandr Petiushko. 2019. Advhat: Real-world adversarial attack on arcface face id system. *arXiv preprint arXiv:1908.08705* (2019).
 - [11] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. 2019. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 1028–1035.
 - [12] Jiajun Lu, Hussein Sibai, and Evan Fabry. 2017. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494* (2017).
 - [13] Rong-Hui Miao, Jing-Lei Tang, and Xiao-Qian Chen. 2015. Classification of farmland images based on color features. *Journal of Visual Communication and Image Representation* 29 (2015), 138–146.
 - [14] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.
 - [15] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 372–387.
 - [16] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.
 - [17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence* 39, 6 (2016), 1137–1149.
 - [18] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. 2018. Adversarial Generative Nets: Neural Network Attacks on State-of-the-Art Face Recognition. *arXiv preprint arXiv:1801.00349* (2018).
 - [19] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. 2018. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*.
 - [20] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
 - [21] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker. 2019. Adversarial examples in modern machine learning: A review. *arXiv preprint arXiv:1911.05268* (2019).
 - [22] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. 2020. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*. Springer, 1–17.
 - [23] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*. Springer, 665–681.
 - [24] Darren Yu Yang, Jay Xiong, Xincheng Li, Xu Yan, and Zhenyu Zhong. 2018. Building Towards "Invisible Cloak": Robust Physical Adversarial Attack on YOLO Object Detector. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*.