Using Frequency Attention to Make Adversarial Patch Powerful Against Person Detector

XIAOCHUN LEI¹,², XIANG CAI¹, CHANG LU², ZETAO JIANG¹,², ZHAOTING GONG¹, AND LINJUN LU¹

¹School of Computer Science and Information Security, Guilin University of Electronic Technology (GUET), Guilin 541004, China
²Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin 541004, China

Corresponding author: Zetao Jiang (zetaojiang@guet.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61876049 and Grant 62172118; in part by the Natural Science Key Foundation of Guangxi under Grant 2021GXNSFDA196002; in part by the Sichuan Regional Innovation Cooperation Project under Grant 2021YFQ0002; in part by the Guangxi Key Laboratory of Image and Graphic Intelligent Processing under Grant GHIP2004; and in part by the Student’s Platform for Innovation and Entrepreneurship Training Program under Grant 202110595168, Grant S202110595168, and Grant 202110595025.

ABSTRACT Deep neural networks (DNNs) are vulnerable to adversarial attacks. In particular, object detectors may be attacked by applying a particular adversarial patch to the image. However, because the patch will lose information during the shrinking process, most existing approaches that employ adversarial patches to attack object detectors would diminish the attack success rate on small and medium targets. This paper proposes a Frequency Module (FRAN), a frequency-domain attention module for guiding patch generation to solve this problem. We use FRAN to allow the patched attack to rely on the low-frequency signals which are lost less in the shrinking process. This is the first study to introduce frequency domain attention to optimize the attack capabilities of adversarial patches. Our method increases the attack success rates of small and medium targets by 4.18% and 3.89%, respectively, over the state-of-the-art attack method for fooling the human detector while assaulting YOLOv3 without reducing the attack success rate of big targets.

INDEX TERMS Adversarial attack, frequency attention, object detection, deep learning.

I. INTRODUCTION

Deep neural networks (DNNs) have garnered much success in computer vision. DNNs have proven to be quite effective in data-driven computer vision tasks and multiple areas such as autopilot, aviation, sign gesture recognition [1], [2], and others [3], [4]. However, DNNs have shown flaws susceptible to perturbation by adversarial samples in image classification, object tracking, and object detection. Previous researches [5], [6], [7] explored attack classification networks. In the last years, several studies [8], [9], [10], [11] have recently demonstrated that object detection networks are similarly susceptible to perturbation by adversarial samples. Thys et al. [8] create a printed adversarial patch that successfully attacked the object detector YOLOv2 [12]. Liu et al. [9] successfully render the object detection network incapable of recognizing objects and generating incorrect prediction categories.

Wu et al. [10] formed wearable clothes for eluding the object detector. They use one adversarial patch to attack one-stage [13] and two-stage [14] object detectors. Wu et al. [10] create wearable clothes for evading the object detector. Xu et al. [11] use a thin-plate spline (TPS) base transformer to build a more robust patch that can withstand non-rigid deformation caused by a moving person’s posture changes. Huang et al. [15] proposed a universal physical camouflage attack method. All these methods have a characteristic that the success rate of attacks on small objects is lower than that of attacks on large objects.

Furthermore, network input size, JPEG compression [5], and network processing can all have impacted the efficacy of adversarial patches. The main reason for the above problem is the unbalance between decreasing the high-frequency and low-frequency signals when down sampling images to a specific size. When the picture is scaled, the loss of high-frequency signals significantly decreases faster than the loss of low-frequency signals. The reduction in picture size
resulted in an overall loss of high-frequency signals, leading to imbalanced representations. High-frequency information is also discarded by JPEG compression. As described, using patches to attack the object detector has an inherent difficulty: while patching, we frequently must shrink the patches and then paste them on the objects in the image. The patch’s attack ability is reduced due to the reduction in pixels caused by sampling.

Because the frequency attention module can determine the relative importance of various locations within the patch frequency domain, this article built the Frequency Module (FRAN), a frequency attention module that guides the patch generation process. The FRAN allows the patched attack to rely on low-frequency signals and enhances the attack ability against small objects. The attack success rates of medium and small objects in our method are 3.89% and 4.18%, respectively, higher than the current best attack algorithm [8], which uses the patch to fool YOLOv3 while detecting people without affecting the attack success rate of large objects.

Our contributions are summarized as follows:
1. We used a frequency attention module to guide the patch generation process in the frequency domain, increasing the adversarial patch’s attack success rate against small and medium objects.
2. Our method raises the patch’s robustness when defending against adversarial attacks using JPEG compression technique.
3. We proposed three new indicators, ASRm, ASRI, and ASRs, to measure the patch’s attack effect on objects of various sizes.

The rest of this paper is structured as follows: Section II goes over the related work on adversarial attacks. Section III discusses how we generate patches and how we use the frequency attention module to guide the process of patch generation. In Section IV, we test the efficacy of the patch. In Section V, we reach a conclusion and look at the future directions of our research.

II. RELATED WORK
Attacks against object detectors have grown exceedingly prevalent due to the widespread use of object detection networks in various areas. In this section, we first go over the related works of the object detection network and the attacks against the object detection network, and then we go over the present challenges.

A. OBJECT DETECTION
The popular one-stage object detector and two-stage object detector are the most common types of modern detectors. The two-stage detector splits the detection task into two parts. In the first stage, the objects in the picture are located. And then, the object detector outputs the categories of the located objects acquired in the first stage. Faster RCNN [14] is a common two-stage detection network. It extracts the region of interest (ROI) using RPN, then produces and classifies detection boxes based on the ROI. Although two-stage object detectors have the advantage of accuracy, the detection speed of two-stage object detectors is often slower than that of one-stage object detectors since the prediction process is divided into two stages.

The one-stage detector uses the entire picture as input and outputs the position and category of the detection boxes directly in the output layer. Yolo-Series algorithms [12], [13], [16] are exceptionally representative one-stage object detectors. To acquire detection boxes, they do not need to extract ROI. Yolo-Series methods have been the most often used object detectors due to their quick detection speed and good detection effect. In this paper, we target YOLOv3 [13] as the main victim detector.

B. ADVERSARIAL ATTACKS ON OBJECT DETECTOR
According to the methods of adversarial attack, the job of attacking the object detector can be separated into two groups. One group is that the object detector can be attacked by rendering an adversarial patch over all of the objects in the picture. The other is to attack the detectors by changing the pixel value of the input image directly.

Thys et al. [8] first proposed using the adversarial patch to attack the object detection network in 2019. They use the loss function to guide the patch generation process, gradient descent to update the patch, and eventually, disable YOLOv2 [12] from detecting patched objects. Moreover, Thys et al. print out the patches to fool the object detector in the physical world by holding a small cardboard plate in front of their body. Liu et al. [9] interfere with the detection results of the entire picture by affixing patches to the image’s edge angle and implemented the object-vanishing and object-mislabeling attacks by designing different loss functions in the same year. Wu et al. [10] use one adversarial patch to attack one-stage [12], [13] and two-stage [14] object detectors. Fan et al. [17] proposed a new network to generate physical adversarial examples by simulating real-world transformations. Huang et al. [15] proposed a universal physical camouflage attack (UPC) which is effective for attacking non-rigid or non-planar objects in the physical world. They design wearable clothes that could attack the object detector in the physical world by printing the patch on clothes. Furthermore, they explore how patches’ attack ability can be applied among different detectors. According to their research, the patch’s attack has strong transferability among networks with the same backbone but is poor between one-stage and two-stage object detection networks. Nevertheless, there are several issues when using adversarial patches to attack surveillance systems in the physical world. One issue is that the object detector may observe the adversarial patches from different viewpoints in the physical world. There is much work [11], [18], [19] to be done in order for the patch to be sufficiently robust from several viewpoints. Xu et al. [11] utilize a TPS base transformer to model the non-rigid deformation of clothing induced by a moving person’s posture changes to create a more robust
adversarial T-shirt. Lennon et al. [19] apply an affine transformation to generate distinct viewpoints of the same patch, fuse the results of computation generated by different viewpoints patches, and then update the patch using gradient descent. As a result, the created patch is more robust from various viewpoints.

In previous research, people directly modified the picture pixels to attack the object detection network. This attack method originates from the study [5] that concerned attack classification networks, and its primary purpose is to create a disturbance that is difficult to notice by humans and causes the object detector to be disturbed at the same time. DAG [20] is proposed in Xie et al.’s work to attack the object detection network. It focuses on attacking Faster RCNN [14]. They first give each object an adversarial label selected randomly from other incorrect classes. In the next step, the image will be updated based on the detector’s output. There was also some work [21] against the transferability of this attack method across networks. According to the latest research, [22] can attack one-stage target detection networks and two-stages as well by changing the picture’s pixels. Their work describes three Targeted Adversarial Objectness Gradient (TOG) attack methods that can induce object-vanishing, object-fabrication, and object-mislabeling attacks in state-of-the-art object detectors.

There is an inherent difficulty with the strategy of using patches to attack the object detector: while patching, we frequently need to shrink the patches and then paste them on the objects in the image. The reduction in pixels caused by sampling will diminish the attack ability of the patches. As a result, the effect of current attack approaches using the adversarial patch for small and medium target objects needs to be improved. We guide the frequency domain information of the patch through the frequency attention module, such that the patch’s attack ability depends more on the low-frequency signal. Because the low-frequency signal of the picture will lose less than the high-frequency signal during image shrinking, the patch’s aggressiveness will be preserved even more.

### III. GENERATE MORE ROBUST ADVERSARIAL PATCH

The current section describes in detail our method for creating adversarial patches. Based on the problems outlined in Section II, we present a solution that employs a frequency attention module to guide the patch generating process. Our method directs the learning of the patch’s frequency information by FRAN. Figure 1 illustrates the primary process of generating the patch $\mathcal{P} \in \mathbb{R}^{W \times H \times 3}$. In Figure 1, Detector is the victim detector [10]. Frequency Module (FRAN) is the frequency domain attention module and Data is the training data set. We first add the patch and FRAN parameters to the optimizer and freeze the detector network parameters before training. We solely use gradient descent to update the patch’s pixel value and FRAN’s parameters during the training procedure. In the training procedure, the patches begin by walking through FRAN with the given data set. The transformation module will render the rescaled patch over every person from the ground truth in the input image. The image is then fed into the victim detector, and the loss function value is calculated based on the detection results. Finally, we update the patch’s pixel value and the FRAN’s parameters based on the calculation result of the loss function. Our ultimate goal is to create patches that minimize the detection success rate of pedestrians of various sizes throughout a dataset.

#### A. FREQUENCY MODULE

FRAN is a frequency domain attention module that assigns different weights in the frequency domain. Figure 2 depicts the construction of FRAN. The FRAN’s general procedure entails converting the patch to the frequency domain using Fast Fourier Transform (FFT) [23] and multiplying it by the attention mask. The output of FRAN is obtained by applying an Inverse Fast Fourier Transform (IFFT) to the results obtained in the previous step.
More formally, we take $\mathcal{F}_\theta$ as an attention function in the frequency domain. The attention function takes the patch $\mathcal{P} \in \mathbb{R}^{w \times h \times 3}$ as input and $\theta \in \mathbb{R}^{w \times h \times 3}$ as the parameter of the attention module. $\text{FFT}(x)$ is the fast fourier transform while $\text{FFT}^{-1}(x)$ represents the inverse fast fourier transform. $\mathcal{Y}(x)$ transforms the patch from RGB to YCbCr color format. $\mathcal{Y}^{-1}(x)$ is the inverse process of $\mathcal{Y}(x)$ FRAN’s output $\mathcal{F}_\theta(\mathcal{P})$ can be stated below:

$$
\mathcal{F}_\theta(\mathcal{P}) = \mathcal{Y}^{-1}(\text{FFT}^{-1}(\text{FFT}((\mathcal{P} \odot \theta))))
$$

Because the image’s high-frequency signal generally losses more than the low-frequency signal during scaling. Consequently, the attack success rate for small and medium target objects can be enhanced if the patch’s attack ability depends more on the low-frequency signal.

### B. TV LOSS

TV loss [8] represents the total variation in the image, resulting in a patch with smooth color transitions and minimal noise. The following formula shows how TV loss is calculated in the previous works [8], [10], [11] of attacking object detectors with adversarial patches.

$$
L_{tv} = \sum_{i,j} \sqrt{\left((p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2\right)}
$$

The $p_{i,j}$ represents the pixel in patch $\mathcal{P}$. This approach only calculates the distance between each pixel and the pixel to its right and the pixel below it. To better smooth the resulting patch, we compute the distance between each pixel and the eight pixels. Thus, we calculate TV loss TV$_r$ as following:

$$
TV_r = \sum_{i,j} \sqrt{\sum_{k \in [-1,0,1]} \sum_{m \in [-1,0,1]} (p_{i,j} - p_{i+k,j+m})^2}
$$

TV$_r$ can better measure the smoothness of a picture since it measures the distance of surrounding pixels among each pixel.

### C. LOSS FUNCTION

Our optimization goal is divided into two parts:

- $TV_r$: The image’s total variance; this loss guarantees that the image is smooth and does not cause excessive noise. For further information, see Section III-B.
- $L_{\text{obj}}$: The highest objectness score from the prediction. As previously mentioned [8], [10], this loss can lower the detector output’s objectness score while detecting the person. Set $x$ to be the input image and $\mathcal{P}$ to be the patch. The transformation function $T(x, \mathcal{P})$ accepts photos and patches as input and scales and affixes the patches to the person in the image. The frequency attention function $\mathcal{F}_\theta$ was discussed in Section III-A. A detector takes a patched image $T(x, \mathcal{F}_\theta(\mathcal{P}))$ as its input and outputs the result with objectness scores $D_i(T(x, \mathcal{F}_\theta(\mathcal{P})))$ is defined as follows:

$$
L_{\text{obj}} = \sum_{i} D_i(T(x, \mathcal{F}_\theta(\mathcal{P})))
$$

Our total loss function is made up of these two parts:

$$
L = L_{\text{obj}} + \alpha TV_r
$$

We combine the two losses, scaled by empirically established factors $\alpha$ and use Adam [24] to optimize the loss. The $L_{\text{obj}}$ is the sum of the objectness scores of all the people in the image. If the detector outputs boxes with a high objectness score, the $L_{\text{obj}}$ will be high. The optimizer will try to minimize the $L_{\text{obj}}$ and make detectors output boxes with a low objectness score. All of the calculated operations are recorded on the PyTorch computation graph. Therefore the optimizer can...
backpropagate the gradient to the adversarial patch’s values and FRAN’s parameters.

IV. EXPERIMENTS AND RESULTS

We demonstrate the efficiency of our strategy in this section by comparing our attack effect against several object detectors. While there are many researches that have optimized the aggressiveness of the patch, their research has focused on optimizing the aggressiveness of the patch in the physical world [10], [11], [19] or generating new attack effects [9], [22]. The core idea to generate the adversarial patch of those researches is still based on the baseline [8]. We choose three methods: Baseline [8], Adv-Tshirt [11] and UPC [15] for comparison.

The rest of this section is structured as following. In Section IV-A, we introduce the basic settings of the experiment, including the data set, the attacked object detectors and the evaluation index. Section IV-B introduces our method for attacking objects of various sizes and different detectors then evaluates its robustness against JPEG compression resistance. Finally, we use ablation experiments to illustrate the efficiency of our strategy. Figure 3 depicts several examples of attacking the object detector. We compared four attack methods in the figure. The first three methods employ the FRAN to guide the patch generation process, and their impacts on small and medium objects are more significant than the Baseline. In the situation of object overlap in a multi-person scene, our method can also have a high attack success rate. The last two lines of the image show the results of the object detector when detecting the original image and the image with a random patch. People in those images can be successfully detected.

A. EXPERIMENTAL SETUP

1) EVALUATION METRICS

This section will introduce the evaluation metrics ASR [25], ASRs, ASRm, and ASRI to analyze adversarial patch attacks.

The attack success rate (ASR) [25] is a metric to better evaluate the attack object detector’s performance. [25] proposes three different types of ASRs. The ASR is used in this work to counter the object-vanishing attack, which is defined as following:

$$\text{ASR} = \frac{\sum_{x \in D} \sum_{\hat{o} \in \hat{O}(x)} [\exists \hat{o}' \in \hat{O}(x') | \text{IOU} (\hat{o}, \hat{o}') \leq \text{tIOU}]}{\sum_{x \in D} ||\hat{O}(x)||}$$

(6)

If $[condition]$ is 1 or 0, the condition is satisfied. We consider $x$ as the input image, $D$ as the set of all $x$. $x'$ represents the image after adversarial disturbance. We use $\hat{o}$ to denote a ground truth box and $\hat{o}'$ to denote the predicted box. Let $\hat{O}(x)$ be the set of ground-truth objects of $\hat{o}$. Similarly, let $\hat{O}(x')$ be the set of predicted objects of $\hat{o}'$. $\text{IOU}(\hat{o}, \hat{o}')$ represents the intersection over union of $\hat{o}$ and $\hat{o}'$. $\text{tIOU}$, a specified super parameter, represents the IOU threshold. For the calculations in the experiment, we choose 0.5 [25].

However, calculating ASR alone does not account for the attack effect of adversarial patches on objects of numerous sizes; consequently, we proposed ASRs, ASRm, and ASRI based on the ideas of APs, APm, and API [26]. ASRs, ASRm, and ASRI represent the attack success rates of attacks on small, medium, and large objects, respectively.

We utilized height to differentiate between three distinct item sizes. Let $\hat{H}$ stand for a height function, $\hat{h}(\hat{o})$ for the detection box’s height, and $\hat{h}(x)$ for a picture’s height. Next, the height ratio $e$ can be defined as follows:

$$e = \frac{\hat{h}(\hat{o})}{\hat{h}(x)}$$

(7)

We made two boundaries $e_0$ and $e_1$ to separate detection boxes of different sizes. Small objects have an $e$ value less than $e_0$, while medium objects have an $e$ value between $e_0$ and $e_1$. A big object is one with an $e$ value larger than $e_1$. In our research, $e_0$ is 0.3, and $e_1$ is 0.6. While dividing the data set, this two values can balance the number of three sizes of objects.

2) DATA COLLECTION

We used two datasets when training the object detector: COCO [26] and VOC [27]. Similarly, we chose 16000 images from COCO dataset to generate the adversarial patch. After randomly shuffling the dataset, we used 12000 images as the training dataset and 4000 images as the test dataset. There is at least one person in each of the images. The number of the small, medium and big size of human boxes is nearly the same in the training and test datasets. We utilized ASR, ASRs, ASRM and ASRI described in Section IV-A1 as the evaluation metrics of the attack effectiveness.

3) VICTIM OBJECT DETECTORS

Our work selected YOLOv3 [13], YOLOv4 [28] and RetinaNet [29] as the victim detectors, and the input image size is $3 \times 612 \times 612$. We first trained pre-training models on the COCO dataset and then trained the final models using the VOC dataset.

4) IMPLEMENTATION DETAILS

We randomly initialized a $3 \times 950 \times 950$ patch in the patch generation process. We used FRAN to guide patch production in the first seven training epochs. Moreover, Adam [24] was used in our experiments. For training, we used a single Tesla A100. Our batch size in the training process was four, and the learning rate was set at 0.005. Each training stopped after about 45 hours when ASR rarely rises. Ours, Baseline [8], Adv T-shirt [11] and UPC [15] take roughly the same time to generate a patch. We used the patch with the highest ASR in the test dataset as the final training result.

B. RESULTS

Our method achieved 93.68% ASR when attacking YOLOv3 [13]. Simultaneously, our method successfully raised ASRm and ASRs without affecting ASRI. Our ASR improved by
FIGURE 3. The effectiveness of an adversarial patch in evading YOLOv3 detection. We classify the images from the test set into four groups based on the size and number of people. Except for the last rows, each row employs a distinct attack method. First row: use FRAN to direct the generation of adversarial patches in the first seven epochs. Second row: use FRAN to direct the generation of adversarial patches in the first seven epochs but did not convert the patch to YCbCr color space. Third row: use FRAN throughout the generating process. Fourth row: use the baseline [8] method to generate the patch. Fifth row: use random generated patch. Sixth row: original image without patch.

2.56%, ASRm increased by 3.89%, and ASRs amplified by 4.18% compared with the Baseline [8]. Moreover, our method achieved a higher ASR on small and middle objects compared with the Adv T-shirt [11] and the UPC [15].

1) ATTACK ON DIFFERENT SIZE OBJECTS
In this paper, we hope that the patch can successfully attack objects of various sizes, which means that it must remain adversarial even after being shrunk in several times. The most commonly used image reduction algorithm is bilinear interpolation. Because this algorithm is a Low-pass filter, patch’s high-frequency information will be lost during the reduction process. As a result, patch attack capability is dwindling. We used FRAN to guide the patch generation process, and the experimental results demonstrate the workability of the method. ASRm and ASRs can achieve 96.41% and 87.21% for small objects. Compared to the Baseline [8], the ASRm and ASRs is only 92.52% and 83.03%. For small objects,
C. ABLATION STUDY

This section performed additional experiments to better understand our method’s robustness. In the first set of experiments (clean), the images were used without a patch. In the second set of experiments (random), the images were used with a random noise patch. The third set of experiments (baseline) employed the baseline [8] method. The fourth set of experiments (Ours(nonYCbCr)) utilizes FRAN without converting it to YCbCr color space. The fifth set of experiments (Ours(keep)) used FRAN throughout the generating process. The sixth set of experiments (Ours) only utilized FRAN in the first seven epochs. The results of ablation experiments were depicted in Table 3.

1) ROBUSTNESS AGAINST JPEG COMPRESSION

According to Ian et al., JPEG compression [5], which removes the image’s high-frequency signals, can significantly reduce the perturbability of attacks. The adversarial image’s effectiveness can be lowered by JPEG compression, which removes the image’s high-frequency signals.

To identify how JPEG compression affects adversarial patch attacks, we saved the patch in JPEG format and reloaded it to calculate their attack success rates on the test dataset. The results were displayed in Table 2.

The experiment results reveal that our method outperforms the baseline when attacking small and medium-sized objects. The attack success rate can be increased by directing patch creation with FRAN. However, the attack success rate decreases if FRAN is used throughout the training process. The reason is that FRAN suppressed frequency information excessively during the later training period, reducing patch aggression. So we chose to use FRAN during the first several epochs. When attacking different detectors we will choose different epoch to stop using FRAN. At the same time, converting the patch into YCbCr color space and extracting

TABLE 1. The ASR(%) against different object detectors. Attacking three object detectors separately and comparing the effect of different methods (Ours, Baseline [8], Adv T-shirt [11] and UPC [15]).

| Detector            | ASR(%) | ASRl(%) | ASRm(%) | ASRv(%) |
|---------------------|--------|---------|---------|---------|
| YOLOv3(Ours)        | 93.68  | 95.73   | 98.41   | 87.21   |
| YOLOv3(Baseline)    | 91.18  | 96.98   | 92.52   | 83.03   |
| YOLOv3(Adv T-shirt) | 91.74  | 97.84   | 92.24   | 84.13   |
| YOLOv3(UPC)         | 92.06  | 98.36   | 93.45   | 83.27   |
| RetinaNet(Ours)     | 84.31  | 90.37   | 85.94   | 75.55   |
| RetinaNet(Baseline) | 81.78  | 89.11   | 80.85   | 72.99   |
| RetinaNet(Adv T-shirt) | 81.34  | 80.04   | 81.25   | 71.34   |
| RetinaNet(UPC)      | 82.39  | 92.04   | 82.25   | 72.99   |
| YOLOv4(Ours)        | 63.57  | 70.43   | 62.15   | 57.12   |
| YOLOv4(Baseline)    | 61.44  | 70.53   | 56.14   | 53.34   |
| YOLOv4(Adv T-shirt) | 61.02  | 71.34   | 55.97   | 54.43   |
| YOLOv4(UPC)         | 62.66  | 73.34   | 57.99   | 55.23   |

ASR of the Baseline [8] and our method is 96.75% and 96.98%, respectively. We improved the attack success rate of small and medium objects while not affecting large objects’ attack success rate. Figure 4 depicts the spectrum attention map when attack YOLOv3. Because Cb and Cr channels in YCbCr color space represent color offset, we only show Y channel’s frequency attention mask.

TABLE 2. The impact of JPEG compression on attacks.

| Methods            | ASR(%) | ASRl(%) | ASRm(%) | ASRv(%) |
|--------------------|--------|---------|---------|---------|
| Ours(JPEG)         | (5.65)  | (2.45)  | (5.53)  | (2.91)  |
| Baseline [8]       | 91.12  | 96.98   | 92.52   | 83.03   |
| Baseline [8](JPEG) | (7.58)  | (4.03)  | (7.55)  | (4.18)  |
| Adv T-shirt [11]   | 91.74  | 97.84   | 92.24   | 84.13   |
| Adv T-shirt [11](JPEG) | (7.74)  | (4.32)  | (6.34)  | (4.98)  |
| UPC [15]           | 92.06  | 98.36   | 93.45   | 83.27   |
| UPC [15](JPEG)     | (8.52)  | (5.20)  | (6.13)  | (3.47)  |

The experiment results reveal that our method outperforms the baseline when attacking small and medium-sized objects. The attack success rate can be increased by directing patch creation with FRAN. However, the attack success rate decreases if FRAN is used throughout the training process. The reason is that FRAN suppressed frequency information excessively during the later training period, reducing patch aggression. So we chose to use FRAN during the first several epochs. When attacking different detectors we will choose different epoch to stop using FRAN. At the same time, converting the patch into YCbCr color space and extracting

TABLE 3. The ASR (%) of adversarial patches generated from Ours, Ours(non-YCbCr), Ours(keep), Baseline [8], Adv T-shirt [11] and UPC [15] in digital-world against YOLOv3. FRAN is the option of whether to use the frequency domain attention mask, YCbCr represents whether to convert the patch into YCbCr color space, Keep use FRAN indicates whether the frequency attention mask is used to guide the generation of patches along all the training process.

| method               | FRAN | YCbCr | Keep use FRAN | ASR(%) | ASRl(%) | ASRm(%) | ASRs(%) |
|----------------------|------|-------|---------------|--------|---------|---------|---------|
| Ours                 | ✓    | ✓     |               | 93.68  | 96.75   | 98.41   | 87.21   |
| Ours(non-YCbCr)     | ✓    |       |               | 85.33  | 89.27   | 92.07   | 78.93   |
| Ours(keep)           | ✓    | ✓     |               | 88.02  | 94.29   | 90.88   | 77.70   |
| Baseline             |      |       |               | 91.12  | 96.98   | 92.52   | 83.03   |
| Adv T-shirt          |      |       |               | 91.74  | 97.84   | 92.24   | 84.13   |
| UPC                  |      |       |               | 92.06  | 98.36   | 93.45   | 83.27   |
| Random               |      |       |               | 21.76  | 23.62   | 16.60   | 25.01   |
| Clean               |      |       |               | 17.83  | 19.26   | 13.61   | 20.69   |
frequency information using Fourier transform during the training phase result in a patch with a greater attack impact.

V. CONCLUSION AND FUTURE WORK

A frequency attention module is proposed in this work to guide the patch generating process. When attacking the YOLOv3 object detector, our attack approach has a success rate of 93.68%. The attack success rate of medium and small targets has improved without compromising the attack success rate on big targets. Our findings show that frequency attention can control the frequency distribution when creating patches, resulting in more robust patches.

In the future, we would like to extend this work by making it more powerful. One way to do this is by optimizing the frequency attention map’s generating process. (i.e. use a general model to produce the frequency attention map) Another way is to optimize the training process of the frequency attention module. Our current method will suppress too much frequency information; using dynamic weight might improve upon this. In addition, we plan to extend our work to the multi-modal models to test the impact of attacking RGB modes or multi-mode attacks on multi-modal models.

REFERENCES

[1] R. Kaluri and C. H. Pradeep, “An enhanced framework for sign gesture recognition using hidden Markov model and adaptive histogram technique,” Int. J. Intel. Eng. Syst., vol. 10, pp. 11–19, Jun. 2017.

[2] R. Kaluri and P. R. Ch, “Optimized feature extraction for precise sign gesture recognition using self-improved genetic algorithm,” Int. J. Eng. Technol. Innov., vol. 8, no. 1, pp. 25–37, 2018.

[3] R. Zhang, L. Xu, Z. Yu, Y. Shi, C. Mu, and M. Xu, “Deep-IRTarget: An automatic target detector in infrared imagery using dual-domain feature extraction and allocation,” IEEE Trans. Multimedia, vol. 24, pp. 1735–1749, 2022.

[4] R. Zhang, S. Yang, Q. Zhang, L. Xu, Y. He, and F. Zhang, “Graph-based few-shot learning with transformed feature propagation and optimal class allocation,” Neurocomputing, vol. 470, pp. 247–256, Jan. 2022.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, arXiv:1412.6572.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” 2016, arXiv:1611.01236.

[7] B. Phan, F. Mannan, and F. Heide, “Adversarial imaging pipelines,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 16046–16056.

[8] S. Thys, W. V. Ranst, and T. Goedemé, “Fooling automated surveillance cameras: Adversarial patches to attack person detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 49–55.

[9] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, “DPatch: An adversarial patch attack on object detectors,” 2018, arXiv:1806.02299.

[10] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, “Making an invisibility cloak: Real world adversarial attacks on object detectors,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 1–17.

[11] K. Xu, X. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, “Adversarial t-shirt! Evading person detectors in a physical world,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 665–681.

[12] J. Redmon and A. Farhadi, “YOLO9000: Better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6517–6525.

[13] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.
CHANG LU is currently pursuing the bachelor's degree with the School of Computer and Information Security, Guilin University of Electronic Technology, where he is focusing on several computer vision tasks, including semantic segmentation and adversarial learning. He has published one IEEE conference paper.

ZETAO JIANG received the Ph.D. degree. He is currently a National Professor of China, a Doctoral Supervisor, and the Director of the Guangxi Key Laboratory of Image Graphics and Intelligent Processing. He is also a Ph.D. Supervisor of the Nanjing University of Aeronautics and Astronautics and the Guilin University of Electronic Technology, the Director of the Chinese Society of Stereology, and the Deputy Director of the Image Analysis Branch of the Chinese Society of Stereology. He has published more than 120 papers, including more than 70 papers collected by SCI and EI. He has presided over five general projects of the National Natural Science Foundation of China. His current research interests include computer vision technology, the application in deep learning, and the Internet of Things. He won the First Prize in Jiangxi Provincial Science and Technology Progress Achievement Award, and in 2010, he won the Second Prize in Jiangxi Provincial Science and Technology Progress Achievement Award.

ZHAOTING GONG is currently pursuing the bachelor’s degree with the School of Computer and Information Security, Guilin University of Electronic Technology, where he is focusing on applied deep learning techniques and several computer vision tasks, including semantic segmentation, and image analytic in the frequency domain.

LINJUN LU is currently pursuing the bachelor’s degree with the School of Computer and Information Security, Guilin University of Electronic Technology. His current research interests include computer vision and deep learning technology applications.