2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Model Agnostic Defense against Adversarial Patch Attacks on Object Detection in Unmanned Aerial Vehicles

Saurabh Pathak[1], Samridha Shrestha[1] and Abdelrahman AlMahmoud[1]

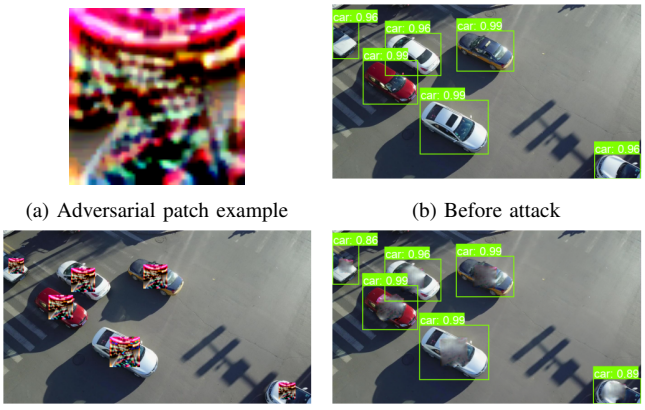
***Abstract*— Object detection forms a key component in Unmanned Aerial Vehicles (UAVs) for completing high-level tasks that depend on the awareness of objects on the ground from an aerial perspective. In that scenario, adversarial patch attacks on an onboard object detector can severely impair the performance of upstream tasks. This paper proposes a novel model-agnostic defense mechanism against the threat of adversarial patch attacks in the context of UAV-based object detection. We formulate adversarial patch defense as an occlusion removal task. The proposed defense method can neutralize adversarial patches located on objects of interest, without exposure to adversarial patches during training. Our lightweight single-stage defense approach allows us to maintain a model-agnostic nature, that once deployed does not require to be updated in response to changes in the object detection pipeline. The evaluations in digital and physical domains show the feasibility of our method for deployment in UAV object detection pipelines, by significantly decreasing the Attack Success Ratio without incurring significant processing costs. As a result, the proposed defense solution can improve the reliability of object detection for UAVs.**

I. Introduction

The utilization of UAVs has seen a substantial surge in recent times. A report suggests that the market for these vehicles is projected to quadruple by the year 2030 [1]. UAVs find applications in a variety of areas, including cargo transportation, remote sensing, and surveillance operations. Object detection is a crucial component in the automation of these vehicles [2]. The object detection module, typically powered by Deep Neural Networks (DNNs), constantly analyzes the camera feed from the UAV. The automation systems of the UAVs heavily depend on the accuracy and reliability of the DNN object detector for high-level tasks such as tracking a detected object and effectively communicating it to an operator [3]. Consequently, it is imperative to ensure that the object detectors in UAV deployments are reliable and robust against potential threats.

In recent years, it has been shown that DNNs are vulnerable to evasion attacks, which are triggered by the addition of specifically crafted input perturbations [4]. In the context of object detection, a common method of attack focuses on adding adversarial patches directly to the object of interest. An advantage of this type of attack is that an adversary with no access to the camera feed can focus on specific types of objects in an image and transfer the patches to the physical domain by printing the patches and attaching them to actual objects that are then imaged by the camera [5].

[1]The authors are with Secure Systems Research Center (SSRC) at Technology Innovation Institute (TII), Abu Dhabi {saurabh, samridha, abdelrahman}@ssrc.tii.ae



(a) Adversarial patch example (b) Before attack

(c) After attack (d) After defense

Fig. 1: Example of an adversarial patch being used to conceal vehicles from an EfficientDet-lite4 model on *VisDrone* dataset.

In the physical domain, adversarial patches serve as a form of "camouflage" (Fig. 1c) that can disrupt the functionality of DNNs-based object detectors, impairing their capacity to accurately identify objects. Given their aerial perspective, UAVs often need to survey extensive physical areas within a single image frame. This scenario exacerbates the issue, as the objects under consideration by a UAV appear smaller relative to the image size due to the altitude and viewpoint. Consequently, a patch-based attack on object detection can have devastating impacts on a higher-level task such as surveillance, tracking, or delivery. Despite this, a significant number of existing object detectors developed for UAV applications do not take into account the potential presence of adversaries, operating under the assumption that their systems will not be subjected to this threat. [6].

Adversarial patches have demonstrated their effectiveness in misleading object detectors for tasks such as causing misclassification of traffic signs [7], making people invisible to detection systems [5], or even obscuring cars [8]. However, adversarial patch generation against the UAV-based object detectors is still mostly overlooked in the literature [9]. Specifically, a patch generation process for UAV-based aerial imagery must take into account changes in the camera viewing angle and perspective and must be effective at a wide range of distances, significantly affecting the generation challenge [10]. Despite reports demonstrating the feasibility of adversarial patches against satellite imagery [8], there has been minimal focus in existing literature examining the degree to which adversarial patches can undermine the reliability of object detectors for UAVs.

In addition to the threat of adversarial patches to UAVs, current defense strategies in the literature often involve the use of adversarial training methods, which may include the incorporation of adversarial patches during the training phase. While adversarial training can significantly mitigate the effects of adversarial attacks, it often comes at the expense of model accuracy. Furthermore, the increasing reliance on third-party solutions has led to the common use of open-sourced or pretrained models from external sources. These models, often frozen and quantized for deployment on resource-limited platforms, may not have been trained with a specific focus on adversarial threats. Consequently, additional defense mechanisms are needed to protect these models after deployment. This is often achieved by adding a preprocessing stage to counteract the effects of adversarial patches [11], [12], [13]. However, this solution necessitates the use of adversarial patches that are specific to the downstream model, while also requiring updates to the defense solution on the UAV to account for new patch generation techniques [14]. Additionally, if the model is replaced, the solution must also be retrained.

Contributions. This paper introduces a novel model-agnostic defense mechanism against adversarial patches for object detection in UAVs. This mechanism is implemented as a preprocessing stage, where the defense against adversarial patches is formulated as an occlusion removal task executed through a thin convolutional autoencoder. Consequently, our method does not rely on any model-specific assumptions about the generation of the adversarial patch, making it robust to changes in the patch generation process. Decoupling from the patch generation process allows our method the flexibility to work with various downstream object detection models without the need for modification or retraining. Experimental results demonstrate that our approach can effectively neutralize adversarial patches prior to the object detection task independently of the downstream model.

In summary, our paper's main contributions are:

- An assessment of the impact of adversarial patches on common UAV object detectors. Our experiments reveal that adversarial patches can significantly affect object detection performance in UAVs, achieving up to 84% attack success rate.
- A novel model-agnostic defense mechanism to facilitate reliable object detection in the presence of adversarial patches. Our defense approach can reduce the Attack Success Ratio (ASR) of adversarial patch attacks by $\approx 30\%$ on average, and can also reduce the impact of non-adversarial patch occlusions while registering an average per-image additional processing cost of only $\approx 4\%$ on the detection pipeline.

The remainder of this paper is organized as follows. Section II further describes the object detection reliability on UAVs and presents related works. Section III evaluates how adversarial patches affect UAVs object detectors. Section IV describes the proposed model-agnostic defense model. Section V evaluates the proposed scheme and Section VI concludes our work.

II. Background

A. UAV Object Detection

Object detection in UAVs typically follows a four-phase process [6]. Initially, the UAV camera feed is continuously collected through a *Data Acquisition* module. Subsequently, the ingested camera images undergo preprocessing before being utilized for object detection. This may involve encoding, resizing, and normalizing the ingested image appropriately. The processed image is then sent to an *Object Detection* module, which employs a Deep Neural Network (DNN) to identify predetermined object classes. These detected objects can be used for autonomous decision-making on the UAV, such as reporting to the operator or tracking the object with the UAV. Given the resource-limited nature of onboard electronics, it is generally preferable to maintain a lightweight object detection model and pipeline. This approach helps ensure low latency and power consumption, which are critical factors for the efficient operation of UAVs.

B. Adversarial Patches in UAV Applications

An adversarial patch is designed to alter the pixels of an input image before it is processed by the object detector of an UAV [8]. To accomplish this, an adversary undertakes an optimization process to create a patch that, when superimposed on a specific object, causes the target object detector to misidentify it. More precisely, given an object detection function $f(x) : x \in X \rightarrow y \in Y$ that outputs an object label y based on the input image x, the goal of the adversarial attack is to find a patch x^* that, when placed over the object x, biases the target detector toward an incorrect prediction.

An adversarial patch begins as a fixed-size arrangement of random noise. This is scaled to a specific proportion of the target object's bounding box size and positioned on each target object. The modified image is then inputted into the target object detection model. The backpropagated gradients from the object detector are utilized to update the pixel values of the patch. The optimization process is typically designed to minimize the classification confidence score of the target object detector with respect to the target objects [5]. Fig. 1 shows an example of an adversarial patch that can impact vehicle identification.

C. Related Works

Over the last years, several works have shown that adversarial patches can significantly affect the accuracy of state-of-the-art DNN-based object detectors [6]. In general, proposed techniques are evaluated in the digital domain, where the generated patch is digitally overlaid on the targeted object. T. B. Brown *et al.* [15] proposed one of the first approaches to generate adversarial patches. Their scheme generates patches to be digitally placed on a given image to bias the classifier. Their work significantly decreases the classifier accuracy, however, it does not address the printability aspects of the generated patch. K. Eykholt *et al.* [16] proposed a patch-based attack that generates stickers to be printed on traffic

signs. They showed that it significantly affects the reliability of object detection schemes in the physical domain. Since then, a plethora of works have shown the efficacy of adversarial patches against a wide range of applications, from traffic sign detection [7], obscuring people [5], or even cars [8]. Yet their threat to UAV-related applications is still in its beginnings. Andrew Du *et al.* [8] aimed the adversarial patch generation against cars on aerial imagery. The authors were able to affect the detection accuracy in the physical domain significantly. Unfortunately, they used a satellite image dataset that does not account for the challenges related to the UAV domain, in particular, the high variance in camera viewing angles, distances, and perspectives. Similarly, J. Lian *et al.* [17] showed the effectiveness of adversarial patches on satellite imagery to conceal airplanes. However, adversarial patch impact on UAV object detection remains largely unexplored, with some recent exceptions [18], [19] studying the YOLO variants in the context.

As a result of the adversarial patch threat to DNN-based object detection, defense techniques are also the subject of several works in the literature [11]. In general, proposed schemes are implemented by adding a preprocessing stage [11], [12], [13], modifying the DNN architecture [14], or retraining the model with the adversarial patches included [20]. To this extent, current solutions usually do not consider their solution's processing costs. In practice, UAVs are resource-constrained devices that should execute their tasks with minimal processing footprint while maintaining their reliability. Consequently, as UAV applications are usually not considered in the adversarial patch literature, proposed solutions cannot be easily used for their defense. In such a case, deployed DNN-based object detectors are not easily updated and must execute their tasks with low processing needs while being resilient to adversaries.

III. Problem Statement

Adversarial patches can compromise the reliability of state-of-the-art object detection. However, their impact on UAV-related applications has been largely overlooked in the existing literature. In this section, we dive into how adversarial patch attacks can affect the reliability of the object detection task within UAV use-case. Specifically, we first outline the specifics of our object detection task, followed by the threat model, and then assess the accuracy degradation when an adversary employs adversarial patches.

A. Task

We consider a UAV-based vehicle detection task, often a crucial part of aerial surveillance and tracking. We use several one-stage multilevel object detectors for this task, namely, SSD [21], RetinaNet [22], EfficientDet [23], and YOLOv5 [24] in various configurations (see Tab. I). In consideration of the limited resources onboard and the fact that most of the aerial objects tend to be small relative to the image area, we discard the top-level feature map from the feature extraction backbones for all the object detectors and include an additional high-resolution feature map instead. Doing so reduces the size of the model and allows the model to focus on objects with a small camera footprint appropriately. In this manner, all the models that we evaluate have $\approx$ 7 to 21 million parameters, making them good candidates for deployment in UAVs. The object detectors are trained to detect vehicles belonging to four classes, namely *Car*, *Van*, *Bus*, and *Truck* on the VisDrone [25] dataset, which provides images in various resolutions, lighting conditions, heights, and camera angles relative to the ground objects acquired using UAV platforms. We use an input image size of 640×640 for all our experiments.

B. Threat Model

To attack the object detectors, we assume the following threat model:

Attacker's goal. The attacker's objective is to generate an adversarial patch that can be used to conceal vehicles from being identified by the object detector. The constructed patches can be attached either digitally or physically on vehicles that the attacker wishes to conceal.

Attacker's capabilities. The attacker operates in a white-box setting, where they have full access to a copy of the deployed model as well as the training dataset. Realistic scenarios where these assumptions apply include a UAV utilizing a publicly available object detection model that has been pretrained on a benchmark dataset, such as EfficientDet [26] trained on the VisDrone dataset [25].

C. The Adversarial Patch Threat

Our objective is to understand the impact of adversarial patch attacks on object detectors designed to work on UAVs used for vehicle surveillance. We use a 64×64 patch to learn adversarial information. Following an approach similar to Thys *et al.* [5], we consider three losses during the patch optimization, the Non-printability Score (NPS), Total Variation (TV), and the classification score. During training, the patch is dynamically scaled for each object in a uniform range of 15% to 35% of the target bounding box area. For evaluation, we use a fixed patch area of 20% relative to the object bounding box.

To ensure the robustness of the attack, the following transformations are applied to the input patch during training and evaluation:

- *Random flip*. Horizontal and vertical
- *Hue rotation*. Uniform range ± 0.08
- *Contrast multiplier*. Uniform range $[0.5, 1.5]$
- *Saturation multiplier*. Uniform range $[0.5, 1.5]$
- *Brightness adjustment*. Uniform range ± 0.3
- *Per-pixel additive noise*. Uniform range ± 0.1
- *Patch rotation*. Uniform range ± 20 degrees on camera axis

To ensure the availability of a reasonable patching area on all the objects, we preprocess the data before training by removing objects that occupy less than 0.1% of the original image area (i.e., area before resizing to a fixed size). After this prepossessing, around 90% of the images from the dataset are retained.

TABLE I: Impact of patch attacks on UAV-based vehicle detection task evaluated on the VisDrone test-set. Mean of 5 runs are reported for each attack method to account for stochasticity in patch transformations.

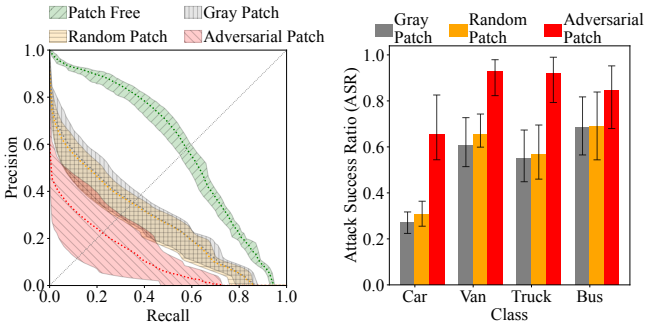
Target	Params(M)	Patch Free		Gray Patch			Random Patch			Adversarial Patch		
		AP	AR	AP	AR	ASR	AP	AR	ASR	AP	AR	ASR
Resnet50v2-SSD	13.4	0.57	0.82	0.20	0.57	0.63	0.19	0.56	0.66	0.04	0.26	**0.92**
Resnet50v2-RetinaNet	17.5	0.54	0.80	0.24	0.60	0.54	0.23	0.62	0.56	0.10	0.37	**0.84**
DenseNet121-RetinaNet	14.4	0.59	0.82	0.26	0.64	0.54	0.26	0.64	0.58	0.07	0.32	**0.90**
EfficientDet-D3	12	0.60	0.82	0.33	0.69	0.46	0.31	0.68	0.52	0.16	0.47	**0.76**
EfficientDet-Lite4	15.1	0.61	0.83	0.32	0.69	0.48	0.31	0.68	0.50	0.20	0.54	**0.72**
YOLOv5-Small	7	0.58	0.83	0.29	0.67	0.51	0.30	0.68	0.51	0.20	0.44	**0.85**
YOLOv5-Medium	20.9	0.61	0.83	0.30	0.66	0.53	0.29	0.67	0.55	0.11	0.37	**0.88**

We evaluate the model Average Precision (AP) and Average Recall (AR) at Intersection Over Union (IOU) threshold of ≥ 0.5 on the test dataset. Fig. 1 shows one example of an obtained adversarial patch overlayed on the target objects. We also evaluate the impact of adversarial patches compared to a randomly initialized patch and a gray patch as baselines, both of which employ the same size, scaling, and rotation approach as their adversarial counterparts. The objective of this comparison is to determine the extent to which the degradation in object detection accuracy is attributable to the features of the adversarial patch as compared to the occlusion caused by the addition of the patch.

We also investigate the Attack Success Ratio (ASR) of the generated patches. The ASR measures the ratio of objects successfully hidden from the object detector after adding the adversarial patch. More specifically, it measures the ratio of correctly detected objects before and after applying the adversarial patch. Objects that are not correctly identified due to the added patch cause a decrease in this ratio and therefore increase ASR. Similarly to the AP metric, we take the average of ASR at all recall thresholds for each class. We then report the ASR as the mean of average ASR across all classes.

Tab. I shows the object detection performance in the presence of patches. We note that the *Gray* and *Random* patch baselines perform similarly, registering on average, 53% and 55% ASR respectively, even with a patchable area of only 20% per object. This points to the fragility of UAV-based object detection and shows that ensuring robustness is in fact a challenging task in this scenario. It is possible to note that the adversarial patch severely impacts the accuracy of all object detection models considered in our work, achieving on average 84% ASR for all models considered in this paper, an increase of over $1.5\times$ the baseline results. In fact, as Tab. I shows, the ASR approaches 90% for some models.

Fig. 2a shows the consolidated impact of patch attacks on the accuracy of all object detectors considered in our evaluation. Notably, adversarial patches affect both *Precision* and *Recall* metrics catastrophically in our use case. In Fig. 2b, it can be observed that the *Car* class is least impacted by the baseline occlusion patches and the contrast between the effect of adversarial and non-adversarial patches is the strongest there. We argue that this is because the VisDrone dataset contains significantly more instances of the *Car* class, compared to others. This biases the detector to identify



(a) Averaged Precision *vs.* Recall (b) Mean ASR by Class

Fig. 2: Aggregated results across all models showing the impact of patch attacks on vehicle detection performance on the VisDrone test set. Each attack method was evaluated 5 times per model. Shaded regions and the error bars denote minimum and maximum values across all models for the respective attack method.

the majority class more effectively, reducing the efficacy of occlusion attacks. Similarly, the adversarial patch is biased to be more effective in concealing the instances of the *Car* class during training, since the patch is applied to them more frequently during training.

D. Discussion

In this section, we have evaluated the impact of patch attacks on commonly used object detectors in the context of a UAV-based vehicle detection task. Our evaluation has yielded two key findings that underscore the need for a defense mechanism to counteract the threat posed by patches. First, our experiments have shown that adversarial patches can significantly impair the performance of UAV object detection, posing a considerable threat to upstream tasks. Second, we also observe a notably sharp decline in performance even in the presence of random or gray patch occlusions, although not as severe as with adversarial patches. These findings suggest that an ex-situ defense mechanism may be necessary to ensure reliability in UAV object detection pipelines when in-situ protection, such as an adversarially trained model, is unavailable. Object detection plays a critical role in enabling automation on UAV applications. Therefore, the provision of a reliable and resilient object detection procedure that is feasible to implement on UAVs is a must to ensure their reliability.

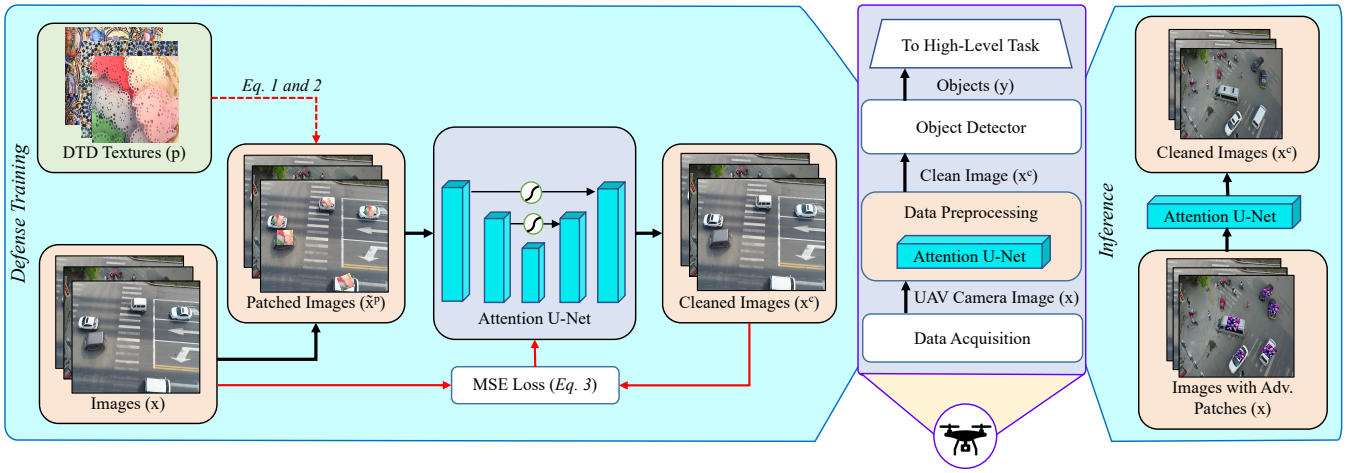


Fig. 3: Proposed model-agnostic mechanism for adversarial patch defense on UAV object detectors. The Attention U-Net model is trained to reconstruct regions masked by DTD patches in the input image. The inference phase uses the built model to preprocess the input image and restore the object regions altered by an adversary to the estimated object pixels without the patch.

IV. A Defense Mechanism Against Adversarial Patch Attacks

Current adversarial patch defense approaches in the literature generally focus on defending against a specific patch generation approach, lacking robustness to be applied in real-world conditions. In such a case, the attacker can generate new adversarial patches as needed to evade detection. Consequently, these solutions must be updated as new patch generation approaches arise or when the downstream model is changed.

Our proposal aims to defend UAV object detection against adversarial patch attacks that aim to conceal an object from the detector. We know that a significant portion of the adversarial patch threat on object detection performance comes from the presence of adversarial patterns on the patch. However, as we have shown in Sec. III, even non-adversarial patches can potentially impact the performance of an object detector due to their occlusion effect alone. The impact due to occlusion is expected to be severe for UAV based object detection since the profile of objects visible to the UAV camera tends to be very small during missions. It motivates us towards a defense approach that is capable of mitigating both these aspects.

We rethink adversarial patch defense from the viewpoint of image restoration. We aim to achieve this in a model-agnostic manner. To that effect, our approach does not require prior access to the details of the adversarial attack, such as adversarial patterns that might be specific to an object detector. Our only assumption is that the attacker uses patches that are positioned on the objects of interest, occluding them partially. Our objective is to recover the object pixels occluded by the patch and in doing so, mitigate the effect of the patch on the object detector. In this manner, we aim to abstract away the details on the patch itself, modeling our defense approach as an occluded object reconstruction problem. Doing so helps us in two ways. First, our defense mechanism is able to defend against adversarial and non-adversarial patch occlusions on objects of interest. Second, our approach allows us to maintain a model-agnostic nature, and once deployed does not require to be updated in response to changes in the object detection pipeline. Additionally, our solution is simple and lightweight in order to form the preprocessing stage in the real-time object detection pipeline onboard UAVs.

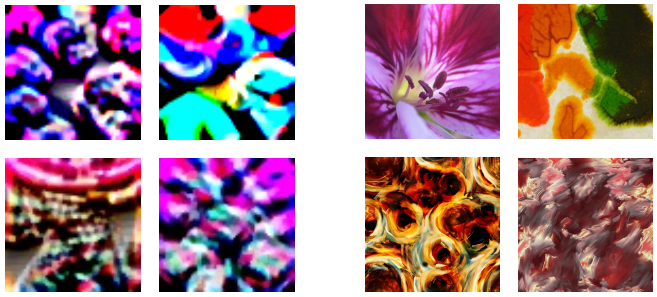
The implementation of our proposed scheme is shown in Fig. 3. We use the Attention-UNet [27] autoencoder architecture for reconstructing the input image in the preprocessing stage. During training, we encourage the model to identify and ignore the patch pixels and reconstruct the image with the object pixels instead. In the inference pipeline, our preprocessing stage recovers the object pixels occluded by patches in the input image before sending it to the downstream object detection model.

A. Training

We train our autoencoder on the VisDrone dataset for the vehicle restoration task. To occlude vehicles, we dynamically add patches to ground-truth bounding boxes following a similar approach and set of transformations described in Sec. III. For our task, we desire to be independent of the adversarial patterns on a patch yet be able to detect and remove them. To that end, we realize that adversarial patterns are typically comparable to textures. Therefore, we use the Describable Textures Dataset (DTD) as a source of textures that are applied to objects as patches during training of our defensive scheme. Fig. 4 shows an example of images from the DTD dataset bearing textural similarities to the adversarial patches learned in Sec. III.

At training time, let x be an image from a given training batch D such that $x \in [-1, 1]^{H \times W \times 3}$ where H and W denote the height and width of the image. We generate the autoencoder input (Fig. 3, *Patched Images* ($\tilde{x}^p$)) based on the following functions:

$$\tilde{p_y}, s_y, l_y = T(p, y) \tag{1}$$



(a) Adversarial patches (b) DTD textures

Fig. 4: Example images from the Describable Textures Dataset (DTD) (right) with textural similarities to adversarial patches (left).

$$\tilde{x}^p = A(\tilde{p_y}, s_y, l_y, x, y) \tag{2}$$

where $T(p, y)$ is a set of stochastic transformations that are applied on a texture patch p obtained from the DTD dataset for each object $y \in x$. Subsequently, a patch application function $A(\tilde{p_y}, s_y, l_y, x, y)$ applies the transformed patch $\tilde{p_y} \in \mathbb{R}^{s_y \times s_y \times 3}$ at location l_y for each object y, where s_y is the size. The location l_y is placed within the bounding box b_y of the object y, such that $b_y > s_y$.

The objective is to restore a given patched image $\tilde{x}^p$ to its unpatched counterpart x (see Fig. 3). We use the pixel-wise Mean Squared Error (MSE) loss

$$\mathcal{L}(x, x^c) \propto \sum_{i}^{H} \sum_{j}^{W} (x_{ij} - x_{ij}^c)^2 \tag{3}$$

where x^c is the reconstructed output of the autoencoder.

B. Defense Implementation for UAVs

It is important to keep the defensive model lightweight to reduce the computational cost incurred due to the addition of a preprocessing stage while maintaining the feasibility of the solution to be deployed in a real-time object detection pipeline onboard a UAV. To that extent, we modify the standard model architecture. We use an EfficientNet-B0 [28] backbone pre-trained on the ImageNet benchmark as the encoder for the Attention-UNet. Taking into account the small area footprint of ground objects relative to the UAV camera, we further discard the top-level feature map from the backbone and instead include a high-resolution map from the prior layers to enable the availability of small object details to the decoder. Doing so also reduces the number of parameters in the encoder layers. We use a slim decoder to further reduce the inference time, with only one decoder convolution per level. We use multiplicative attention in the attention module. We do not use any pooling layers and use the Hard-Swish [29] as an activation function in the decoder for faster computation. Our decoder has a 5-level configuration with 16, 32, 64, 128, and 256 filters, respectively. As a result, our model has only $\approx$ **1.2 million** parameters, making it suitable for deployment in UAVs.

We do not freeze the backbone during training; the complete architecture is trained on the VisDrone dataset. The model training was executed for 200 epochs with a batch size of 16 images. SGD with momentum was used for optimization, with a cosine annealed learning rate schedule.

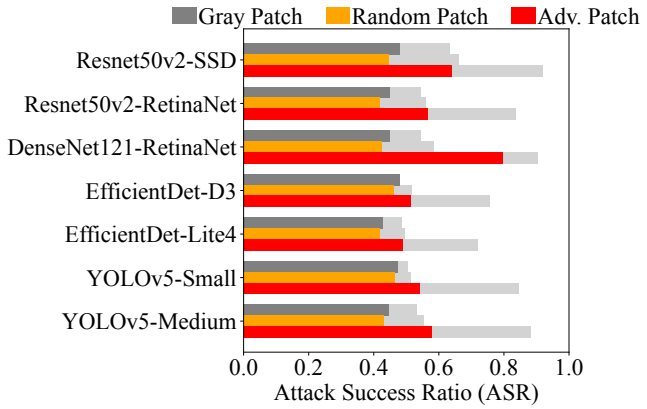


Fig. 5: Performance of the proposed scheme evaluated against individual object detectors, as compared to the performance without defense shown here as a light-gray shadow. Mean of 5 runs are reported for each attack method to account for stochasticity in patch transformations.

V. Evaluation

In the context of the UAV vehicle detection task, our evaluation aims to answer the following Research Questions (RQs) in this section:

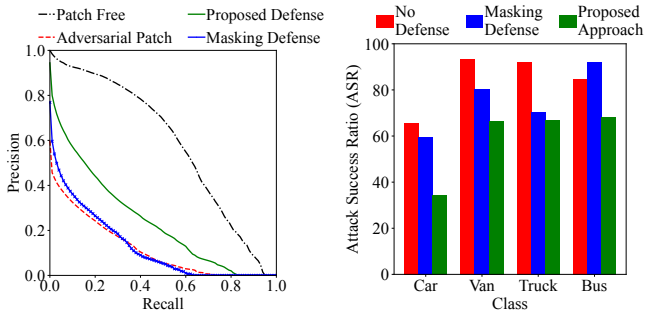
- (**RQ1**) *How well does the proposed defense scheme perform?*
- (**RQ2**) *What are the processing costs of our proposal?*
- (**RQ3**) *How well does our approach perform in the physical domain?*

A. Defending Against Adversarial Patches

To answer RQ1, we evaluate object detection performance when using the cleaned images generated by the autoencoder.

Fig. 5 shows that the proposed defense approach can, in practice, reduce adversarial and non-adversarial patch occlusions in the input image at the preprocessing stage, resulting in previously hidden objects being detected by the downstream object detector, reducing the ASR of an adversary. On average, our method reduces the ASR of adversarial patch attacks in all models from 84% to 59%, a relative reduction of $\approx$ 30%. As a result, our model significantly reduces the impact of adversarial patch attacks on the reliability of UAV object detection. Qualitatively, a translucent effect can be seen in place of the patch due to object pixels restored by our approach (see Fig. 1). For non-adversarial patches, our method registered a $\approx$ 13% and $\approx$ 21% relative improvement on ASR averaged across all models for *Gray* and *Random* patch attack baselines, respectively.

Comparision: Our solution is closely related to *Patch Masking* defense techniques similar to [12], [13], in which the identified patch is masked out from the input image rather than restoring it. For a fair comparison with a pixel-masking-based solution, we keep our approach model-agnostic by avoiding knowledge about the downstream object detector in contrast to [12], [13]. We construct a single-stage pixel



(a) Averaged Precision *vs.* Recall (b) Mean ASR by class

Fig. 6: The proposed method in comparison to a model-agnostic patch-masking approach based on segmentation.

(a) Without defense (b) With defense

Fig. 7: Example of a physically printed patch (left) affecting the accuracy of a YOLOv5s object detector. Our defense approach can identify and remove the physical patch (right) from the camera image, making the vehicle detectable.

masking approach, using Attention-UNet as before, but for a patch segmentation task and train it in an identical manner. The obtained segmentation mask is inverted and multiplied pixel-wise with the input image in the preprocessing stage. Fig. 6 shows the average Attack Success Ratio (ASR) impact on all object detectors in the presence of adversarial patches. It is evident that a standalone model-agnostic approach based on segmentation-based masking alone is insufficient for defense and requires additional measures or assumptions to be effective. On the other hand, the proposed defense solution can defend the UAV object detectors in a model-agnostic manner without having prior access to the adversarial patches or object detectors. Having a tiny single-stage defense mechanism also helps in deployment onboard UAVs.

To evaluate the processing costs of our adversarial patch defense solution (RQ2), we compare the average execution time of the object detection pipeline with and without defense. Without any quantization or optimization, our solution incurred an average additional processing time of only $\approx 4\%$ per image during inference on the VisDrone test set. Consequently, we are able to provide reliable object detection for UAVs without incurring high additional processing costs.

B. Physical Domain Experiments

For the physical domain scenario (RQ3), we constructed a controlled environment laboratory setup wherein we used toy model vehicles (1:50 downscaling ratio) and imaged them from various angles at heights ranging from 1 to 6 feet using a standard high-definition camera. In practice, our laboratory setup closely mimics the aerial views of vehicles taken from a UAV-based camera such that the images of the vehicles achieve the same camera footprint as observed in the VisDrone dataset. Our setup also allows us to quickly print patches on a standard high-definition printer and place them directly on the toy models, instead of using actual cars which requires a much larger and complex printing setup and overlaying process. A YOLOv5s model was trained on the collected toy dataset comprising around a hundred images collected for the task. Considering the limited number of images, a VisDrone checkpoint was used for finetuning, and data augmentations were heavily employed. Similarly, we trained adversarial patches to hide the cars. Fig. 7 shows the performance of our approach on an image from the hold-out validation set with a physically printed adversarial patch. As can be seen, the physical patch conceals the car from the object detector successfully. This scaled-down laboratory experiment serves as a proof of concept that the effectiveness of our approach extends to the physical domain in the case of UAV-based object detection.

VI. Conclusion

Object detection is indispensable for UAV applications. Therefore, deployed solutions must be resilient and reliable against adversaries. In this paper, we have shown that adversarial patches significantly degrade the accuracy of UAV object detection, posing a significant threat to their complete automation. To address this challenge, we have proposed a defense model that formulates adversarial patches as an image restoration task.

The proposed scheme is implemented without using adversarial information from the object detectors during the training phase and is executed as a preprocessing task during the inference phase. This makes our method effective on multiple object detectors as shown in the paper. The evaluations carried out have shown the efficacy of the approach, significantly decreasing the success of attacks without incurring a significant impact on processing costs. Notably, the effectiveness of our method is determined by the richness and variety of textural patterns the defensive model is exposed to during training. Adversarial patterns that do not follow the textures understood by the model are likely to escape the defense mechanism.

In future work, we would investigate our approach in additional tasks. A key insight of our method comes from using a relatively simple but lightweight image restoration technique in the form of pixel-wise image reconstruction. It would be beneficial to employ more advanced techniques from the image inpainting literature to improve the effectiveness of this method. A single-stage approach that combines masking to detect and abstract away the adversarial patterns followed by inpainting is an interesting direction to explore in the future.

REFERENCES

[1] *Unmanned Aerial Vehicle (UAV) Drones Market*, 2023. [Online]. Available: https://www.precedenceresearch.com/unmanned-aerial-vehicle-drones-market

[2] D. Ebrahimi, S. Sharafeddine, P.-H. Ho, and C. Assi, "Autonomous uav trajectory for localizing ground objects: A reinforcement learning approach," *IEEE Transactions on Mobile Computing*, vol. 20, no. 4, pp. 1312–1324, 2021.

[3] M. Franke, C. Reddy, D. Ristić-Durrant, J. Jayawardana, K. Michels, M. Banić, and M. Simonović, "Towards holistic autonomous obstacle detection in railways by complementing of on-board vision with uav-based object localization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7012–7019.

[4] M. Yin, S. Li, C. Song, M. S. Asif, A. K. Roy-Chowdhury, and S. V. Krishnamurthy, "Adc: Adversarial attacks against object detection that evade context consistency checks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 3278–3287.

[5] S. Thys, W. Van Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[6] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude uav datasets: A survey," *Image and Vision Computing*, vol. 104, p. 104046, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885620301785

[7] X. Wei, Y. Guo, and J. Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2711–2725, 2023.

[8] A. Du, B. Chen, T.-J. Chin, Y. W. Law, M. Sasdelli, R. Rajasegaran, and D. Campbell, "Physical adversarial attacks on an aerial imagery object detector," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1796–1806.

[9] M. Klingner, V. R. Kumar, S. Yogamani, A. Bär, and T. Fingscheidt, "Detecting adversarial perturbations in multi-task perception," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 13 050–13 057.

[10] H. Chawla, A. Varma, E. Arani, and B. Zonooz, "Adversarial attacks on monocular pose estimation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 12 500–12 505.

[11] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, "Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking." in *USENIX Security Symposium*, 2021, pp. 2237–2254.

[12] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, "Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 14 973–14 982.

[13] P.-H. Chiang, C.-S. Chan, and S.-H. Wu, "Adversarial pixel masking: A defense against physical attacks for pre-trained object detectors," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1856–1865.

[14] A. Shafahi, M. Najibi, Z. Xu, J. Dickerson, L. S. Davis, and T. Goldstein, "Universal adversarial training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5636–5643.

[15] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[16] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[17] J. Lian, S. Mei, S. Zhang, and M. Ma, "Benchmarking adversarial patch against aerial detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[18] S. Shrestha, S. Pathak, and E. Kugler Viegas, "Towards a robust adversarial patch attack against unmanned aerial vehicles object detection," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10 2023.

[19] Y. Zhang, Y. Zhang, J. Qi, K. Bin, H. Wen, X. Tong, and P. Zhong, "Adversarial patch attack on multi-scale object detection for uav remote sensing images," *Remote Sensing*, vol. 14, no. 21, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/21/5298

[20] S. Rao, D. Stutz, and B. Schiele, "Adversarial training against location-optimized adversarial patches," in *Computer Vision – ECCV 2020 Workshops*, A. Bartoli and A. Fusiello, Eds. Cham: Springer International Publishing, 2020, pp. 429–448.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[23] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[24] Ultralytics, "YOLOv5: A state-of-the-art real-time object detection system," https://docs.ultralytics.com, 2021.

[25] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2022.

[26] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 10 778–10 787. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01079

[27] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[28] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[29] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.