remote sensing

MDPI

Article

Attention-Enhanced One-Shot Attack against Single Object Tracking for Unmanned Aerial Vehicle Remote Sensing Images

Yan Jiang and Guisheng Yin *

College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China; y.jiang@hrbeu.edu.cn

* Correspondence: yinguisheng@hrbeu.edu.cn

Abstract: Recent studies have shown that deep-learning-based models for processing Unmanned Aerial Vehicle (UAV) remote sensing images are vulnerable to artificially designed adversarial examples, which can lead to incorrect predictions of deep models when facing adversarial examples. Previous adversarial attack methods have mainly focused on the classification and detection of UAV remote sensing images, and there is still a lack of research on adversarial attacks for object tracking in UAV video. To address this challenge, we propose an attention-enhanced one-shot adversarial attack method for UAV remote sensing object tracking, which perturbs only the template frame and generates adversarial samples offline. First, we employ an attention feature loss to make the original frame's features dissimilar to those of the adversarial frame, and an attention confidence loss to either suppress or enhance different confidence scores. Additionally, by forcing the tracker to concentrate on the background information near the target, a background distraction loss is used to mismatch templates with subsequent frames. Finally, we add total variation loss to generate adversarial examples that appear natural to humans. We validate the effectiveness of our method against popular trackers such as SiamRPN, DaSiamRPN, and SiamRPN++ on the UAV123 remote sensing dataset. Experimental results verify the superior attack performance of our proposed method.

Keywords: adversarial attack; one-shot attack; object tracking; UAV remote sensing

check for updates

Citation: Jiang, Y.; Yin, G. Attention-Enhanced One-Shot Attack against Single Object Tracking for Unmanned Aerial Vehicle Remote Sensing Images. *Remote Sens.* **2023**, *15*, 4514. https://doi.org/10.3390/rs15184514

Academic Editor: Riccardo Roncella

Received: 30 June 2023
Revised: 11 September 2023
Accepted: 11 September 2023
Published: 14 September 2023

1. Introduction

Unmanned Aerial Vehicle (UAV) remote sensing [1] is widely employed in various fields, such as marine environmental monitoring, land use survey, and water resource development, due to its real-time image transmission, high resolution, and flexible maneuvering. Deep learning has played a pivotal role in enhancing the performance of UAV remote sensing image and video applications [2], encompassing tasks such as image classification [3], object detection [4], semantic segmentation [5], and single object tracking (SOT) [6]. SOT is an important research direction in the field of UAV remote sensing [7], and earlier approaches predominantly relied on traditional machine learning methods, such as correlation filters, to achieve accurate and stable results. A significant advancement in SOT was the introduction of the fully convolutional tracking model based on the Siamese Network, proposed by Bertinetto et al. [8]. This structural modification yielded substantial improvements in tracking performance. Subsequently, several Siamese-network-based trackers, including SiamRPN [9], DaSiamRPN [10], and SiamRPN++ [11], have emerged and have been extensively adopted for UAV remote sensing video object tracking, demonstrating exceptional tracking efficacy and efficiency. Moreover, recent advancements have witnessed the adoption of transformer-based architectures [12] as the backbone for object trackers [13], resulting in impressive performance gains in this domain.

However, the improved performance of UAV remote sensing image processing also entails security risks following the revelation by Goodfellow et al. [14] of the susceptibility of neural networks to adversarial samples. Since then, researchers have discovered that convolutional neural networks are highly vulnerable to adversarial samples across various tasks, leading to potential security threats to deep models. Adversarial examples are created by introducing imperceptible perturbations to the original data. These perturbed examples closely resemble the original data, yet they can cause unexpected and erroneous outcomes. The existence of adversarial samples underscores the need for robustness and security measures in UAV remote sensing image processing systems.

Since then, an increasing number of researchers have focused their attention on deep learning security and have proposed adversarial attack methods for various tasks. The most renowned among these are the fast gradient sign method (FGSM) [14], the project gradient descent (PGD) [15], and the Carlini and Wagner (C&W) attack method [16]. Building upon these classic approaches, subsequent researchers have introduced targeted adversarial attack methods in diverse domains, including detection models [17–19] and segmentation models [20–22]. More recently, there have been proposed adversarial attack methods specifically designed for SOT tasks [23–27]. While there are some adversarial attack methods for object tracking, the research on adversarial attacks in the context of UAV remote sensing [28] has primarily focused on classification and object detection tasks, with limited exploration of UAV remote sensing SOT. Attacking UAV remote sensing SOT is a challenging task due to several factors. Firstly, in continuous video, unlike classification and detection, the target is tracked by calculating the cross-correlation of features, meaning that no pre-labeled video tracking is available. Given that the video consists of a continuous sequence of image frames, the process of generating perturbations on a frame-by-frame basis necessitates extensive computational resources. Consequently, the attack method's effectiveness is limited as it cannot execute real-time attacks due to the substantial computational burden involved. Furthermore, the different heights of UAV remote sensing video data acquisition lead to different sizes of targets in the image. Smaller targets also present considerable challenges to attacking UAV remote sensing video.

To address the challenge of the adversarial attack in remote sensing object tracking we mentioned above, we propose a novel one-shot adversarial attack method for remote sensing object tracking, called attention-enhanced one-shot attack. Our proposed attack method is optimized to generate a unique perturbation that can successfully attack the tracker by only attacking the tracker's template frame. The generated perturbation is unique for each object tracking video. Only attacking the template frame can improve the computational efficiency compared to generating the adversarial samples frame by frame. Therefore, our proposed method is more practical in real-time attacks. In summary, the main contributions of this paper are as follows:

- We propose a novel one-shot adversarial attack method to explore the robustness of remote sensing object tracking models. Our attention-enhanced one-shot attack only attacks the template frame of each video. It generates a unique perturbation for each video which saves a batch of time and is more practical in real-time attacks.
- The effectiveness of our proposed attack method is verified on UAV remote sensing video. Our method optimizes perturbation via attention feature loss to force the generated adversarial samples dissimilar to the raw image and attention confidence loss to suppress or stimulate the different confidence scores, using these two loss functions to optimize an imperceptible perturbation to fool the tracker into getting wrong results.
- In addition, considering the high-altitude shooting of UAV remote sensing video leads to the characteristics of different sizes of targets, we also propose a background interference loss, which forces the tracker to consider the background information around the target, resulting in the tracker being interfered by redundant background features and unable to track the correct target. We also use TV loss to make the generated adversarial image more natural for the human eye.

We conduct experiments on UAV remote sensing benchmark datasets, e.g., UAV123 [28] on popular trackers. We mainly target representative trackers based on Siamese networks, such as SiamRPN [9], DaSiamRPN [10] and SiamRPN++ [11]. The experimental results illustrate our proposed method has a superior attack ability against the popular trackers.

The rest of the paper is organized as follows: Section 2 introduces recent adversarial attack methods for SOT and adversarial attacks for UAV remote sensing images. Section 3 explains the motivation and describes the details of our proposed attention-enhanced one-shot attack. Section 4 provides and analyzes the experimental results of our proposed attention-enhanced one-shot attack against popular trackers. Section 5 summarizes the conclusions.

2. Related Work

This section provides an overview of the SOT tracker. Then, we outline classical adversarial attacks for various visual tasks. Finally, we briefly analyze the adversarial attack methods specific to UAV remote sensing images.

2.1. Visual Object Tracking

Since Bertinetto et al. [8] introduced the fully convolutional network into the object tracking task, the performance of deep-learning-based trackers has been greatly improved. The introduction of the fully convolutional network by Bertinetto et al. [8] has led to remarkable advancements in deep-learning-based object tracking. This development has inspired researchers to propose diverse trackers based on deep learning, with the Siamese-network-based trackers emerging as prominent representatives. Inspired by this, researchers have proposed a variety of trackers based on deep learning. Among them, the most representative one is the tracker based on the Siamese network. However, it has the shortcomings of insufficient utilization of background information and limited use of specific field information. To address this issue, Li et al. [9] proposed SiamRPN, which combines the RPN and Siamese network to improve the problem of insufficient background information utilization. Furthermore, DaSiamRPN [10] addresses the limitation of distractors in the online tracking process by learning distractor-aware features during the offline training phase and suppressing distractors during inference. This approach improves tracking performance and enables long-term tracking. Additionally, SiamRPN++ [11] introduces a deep association layer and replaces the backbone network with a more robust feature extractor, ResNet50 [29], which significantly improves tracking performance by breaking the deep space limitation of deep networks. Recently, there have been notable explorations in applying transformer-based architectures [13] to Visual Object Tracking (VOT) [12], which have demonstrated exceptional performance. Despite the significant advancements achieved through deep learning techniques, it is crucial to acknowledge the susceptibility of these methods to adversarial samples, which presents a substantial security risk.

2.2. Adversarial Attacks

Despite the outstanding performance achieved by deep-learning¬-based trackers, they still show vulnerability to adversarial attacks. Early adversarial attack methods can be classified into gradient-based and optimization-based. Classic examples of gradient-based methods include Fast Gradient Sign Method (FGSM) [14] and Project Gradient Descent (PGD) [15]. These methods calculate the model's derivative with respect to the input, use the sign function to obtain the gradient direction, and apply single-step or multi-step update perturbation. Although gradient-based methods are fast and efficient, they may struggle when the update step is complex, such as when using momentum. The Carlini and Wagner (C&W) attack [16] is an optimization-based approach that uses clipped gradient descent to smooth the optimization process, accelerate convergence, and avoid getting stuck in extreme regions. Researchers have extended adversarial attacks to various fields. Xie et al. [21] proposed the DAG attack, which can fool both object detection and segmentation tasks. Li et al. [30] proposed a method for generating universal perturbations and analyzing their effects on different object detection models. Wang et al. [31] extended adversarial attacks to the physical world by solving an optimization problem that minimizes the target object's probability in the output of the object detection model, thereby hiding the target object. Recent research has extended adversarial attacks to object-tracking tasks. Cooling–shrinking attack, proposed by Yan et al. [32], interferes with the heatmaps by cooling down the hot regions where the target may exist, leading to the tracker losing the target. Guo et al. [33] used tracking video characteristics to propose a blurred attack. Ding et al. [34] proposed the first physical attack using the maximum textural discrepancy loss function to misguide visual trackers by de-matching the template and search frames at hierarchical feature scales. These examples demonstrate that general vision deep learning models are highly vulnerable to adversarial attacks, and UAV remote sensing image processing tasks are also threatened by adversarial samples.

2.3. Adversarial Attacks in Remote Sensing

Since the seminal discovery of adversarial examples by Goodfellow et al. [14], the research on adversarial attacks in remote sensing image processing has witnessed significant growth. These studies have contributed to the advancement of knowledge in the field. In the domain of remote sensing, adversarial attacks have been explored extensively. Ref. [35] introduced the artificial intelligence challenges and prospects in the field of remote sensing.

Firstly, several studies have focused on adversarial attacks for classification tasks in remote sensing. Czaja et al. [36] initially proposed targeted adversarial attacks on RS data. Chen et al. [37] evaluated the vulnerability of models with different structures trained on the same remote sensing dataset and found that the number of features affects model vulnerability. Bai et al. [38] first proposed a target universal adversarial sample generate method focused on white-box setting. Subsequently, Xu et al. [39] introduced the first black-box universal adversarial example generation method for classification in the remote sensing field. Additionally, defense methods against adversarial attacks in remote sensing classification have been investigated, e.g., Xu et al. [40] first explored targeted and untargeted attacks for remote sensing scene classification and investigated adversarial training to improve the robustness of the classifier. Ref. [41] proposed a defense method by comparing the confidence of the classifier output with the soft threshold of the category to distinguish whether it is an adversarial sample and to train the classifier to improve its robustness. Regarding object detection in remote sensing, Sun et al. [42] first revealed the threats of patch attacks on object detection in remote sensing. The Patch–Noobj framework [43] can interfere with remote sensing image target features, which can mislead the results of detectors and reduce the confidence of the predicted bounding box. Researchers also extended adversarial attacks to the physical world. Ref. [44] evaluated the attack ability on different scales objects, they also perform physical adversarial attacks on multi-scale objects. Rust-Style Patch [45] works on improving the natural and robust adversarial patches by utilizing style transfer on remote sensing, and the authors conducted experiments in both the digital and physical domains. Wei et al. [46] proposed a new way to utilize pan-sharpened images to attack object detectors. While there have been studies on adversarial attacks in the context of classification and detection tasks for remote sensing images, there has been little research on adversarial attacks for UAV remote sensing video for object tracking. Fu et al. [47] proposed an efficient attack approach that consists of downsampling the original frame directly and upsampling using a super-resolution to generate the adversarial frame. Therefore, To explore the robustness of object tracking models for UAV remote sensing videos in an adversarial environment, we propose a novel attention-enhanced one-shot adversarial attack method. Our proposed method generates adversarial examples that can successfully fool the tracker while only attacking template frames for each video. Our method is very efficient and can save a batch of time which is more practical in real-time applications. The results indicate the security of object tracking for UAV remote sensing videos should be the attention of researchers.

3. Attention-Enhanced One-Shot Attack

In this section, we will introduce the working mechanism of our proposed attention-enhanced one-shot attack method. Figure 1 shows the overview of our adversarial perturbation training process. The generated perturbations are updated by optimizing attention feature loss, attention confidence loss, TV loss, and background disturbance loss through backpropagation. Additionally, we describe the attention mechanisms that we utilize to generate the perturbation to enhance the attack power and loss functions.

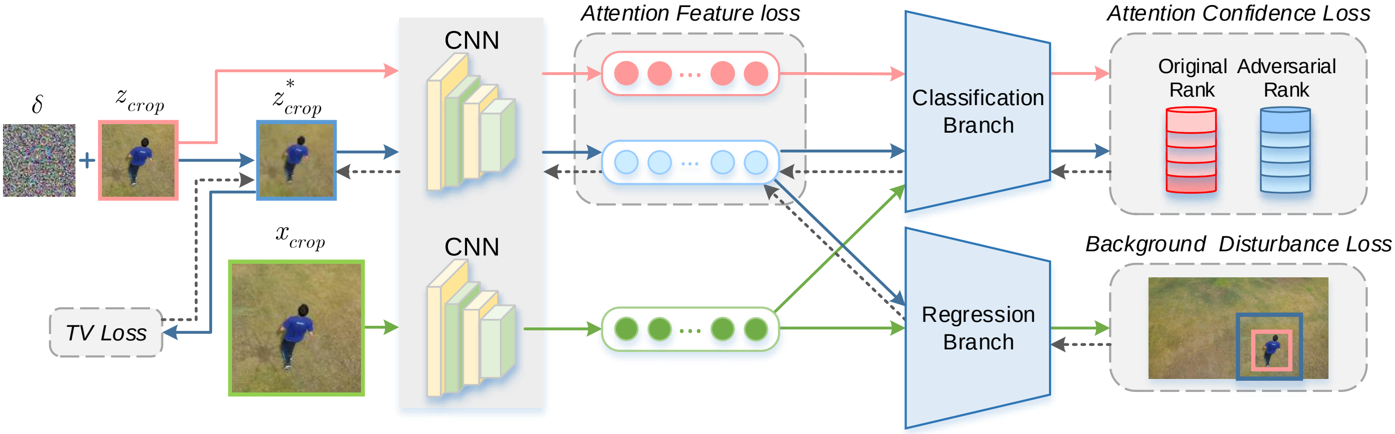


Figure 1. An illustration of the adversarial perturbation training process in our proposed attention-enhanced one-shot attack method. δ is perturbation. x_{crop} and z_{crop} denote the search patch and template patch, respectively. z^*_{crop} is adversarial template patch.

3.1. Motivation

In the tracking process, the tracker always tracks the target similarly to the template patch in the subsequent frame by comparing the cross-correlation information between these two frames. Given a remote sensing video from UAV dataset as $\mathcal{X}$, which can be expressed as $\mathcal{X} = \{x_i, i = 0, 1, 2, \cdots, I\}$, we input the template frame $z = x_0$ and ground truth B^i_{gt} to initialize the tracker first. Then, we feed the search frame $\{x_i, i = 1, 2, 3, \cdots, I\}$ to the tracker to get the search patch $\bar{x}$.

In this paper, we mainly focus on popular trackers based on Siamese networks. These trackers track the target by computing similarity maps between the template patch z and the subsequent search region $\bar{x}$ based on their respective feature representations. Accordingly, we aim to attack the matching process between the template and search features. Specifically, we simulate the tracking process exclusively within the template frame and optimize the perturbations to effectively interfere with the feature-matching mechanism of the tracker.

The tracker always uses Gaussian windows to suppress large displacements between consecutive frames. Hence, only attacking the output features of the tracker may not guarantee the desired attack effectiveness. To achieve a comprehensive attack, we consider additional strategies to optimize the perturbations. The first strategy involves incorporating confidence loss and feature loss, while the second strategy involves introducing a background disturbance loss. Since our attack method is designed based on remote sensing object tracking, we find that remote sensing videos' characteristics (e.g., presence of small targets) may affect the attack ability of adversarial samples. To address this problem, we design the background disturbance loss to consider the background information surrounding the target during perturbation generation. Consequently, during template initialization, the interference caused by redundant background features hinders the accurate matching of subsequent search frames' features. To further enhance the attack strength, we introduce an attention mechanism to both the feature loss and confidence loss. By combining all these factors, we propose an attention-enhanced one-shot attack method, which will be elaborated upon in detail in the subsequent section.

3.2. Attention-Enhanced One-Shot Attack

The main purpose of our attention-enhanced one-shot attack method is to generate a one-shot imperceptible perturbation that can deceive the tracker away from the correct trajectory. Our method is efficient by only attacking the template frame of each video can successfully attack the tracker; therefore, we train our adversarial perturbations in the tracking process of the template frame. We define the adversarial example of attacking the tracker as follows:

$$\begin{aligned} &\arg\ \min\ \mathcal{L}(z, z+\delta) \\ &s.t.\ \|\delta\|_\infty \leq \epsilon \end{aligned} \tag{1}$$

where $\|\|_\infty$ denotes ℓ_∞ norm, and we limit the maximum perturbation range of the per-pixel value into the range $[-\epsilon, \epsilon]$, ensuring the perturbation is imperceptible to human eyes. We set ϵ to 16 in our experiments. δ represents the perturbation that we optimize to attack the tracker. $\mathcal{L}$ means the $\mathcal{L}_{total}$ which is detailed in Equation (10).

3.2.1. Attention Feature Loss

The primary objective of remote sensing object tracking is to track a target by comparing the similarity of paired images, specifically the features extracted from the template frame and the search frame. In our proposed method, we aim to attack all feature candidates extracted from CNNs to interfere with matching the template features.

Let $\varphi(\cdot)$ represent the extracted feature map from CNNs. We maximize the Euclidean distance between the clean template patch $\bar{z}$ and the adversarial template patch to ensure that the generated adversarial feature is dissimilar to the clean template patch feature. Thus, the feature loss function is defined as follows:

$$L_f = -\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\varphi(z^*)_{m,n} - \varphi(\bar{z})_{m,n}\right)^2 \tag{2}$$

where M and N denote the dimensions of the confidence candidates.

To further enhance the attack effectiveness of our proposed method, we incorporate attention mechanisms into the feature loss function. Specifically, we employ channel-wise activation-guided attention to penalize the less crucial channels within the feature maps. This operation not only aids in distinguishing the important channels but also strengthens the attack capability. Moreover, Equation (2) can be reformulated as follows:

$$L_f^A = -\frac{1}{MN}\sum_{j=1}^{C}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(w_j\left(\varphi_j(z^*)_{m,n} - \varphi_j(\bar{z})_{m,n}\right)\right)^2 \tag{3}$$

The feature attention weight w_i defined as:

$$w_j = \frac{1}{a + b \cdot \tanh\left(c \cdot \left(\varphi_j(z)_{mean} - \varphi_j(z)_{min-mean}\right)\right)} \tag{4}$$

where a, b and c are controlling hyper-parameters. $(\cdot)_{mean}$ and $(\cdot)_{min-mean}$ are each channel's mean and minimum mean values, respectively.

3.2.2. Attention Confidence Loss

Only attacking the remote sensing object tracking by feature loss is not enough to apply a successful attack. Therefore, we consider also attacking the confidence of the classification branch outputs at the same time based on the attack features. Assume the tracking target is present in the search patch $\bar{z}$. After extracting the feature map of the template $\varphi(z^*)$ and the search patch $\varphi(\bar{z})$, respectively, from CNNs, the tracker utilizes these features to calculate the cross-correlation between the two images through the RPN. The RPN consists of classification and regression branches. Specifically, the classification branch outputs the opposite and negative r confidence candidates. The opposite confidence of the r candidates can be ranked to $\mathcal{R}_{1:r} = \{\mathcal{R}_1, \cdots, \mathcal{R}_r\}$. To achieve the goal of activating the low-ranking confidence candidates and suppressing high-confidence candidates, the confidence loss function can be expressed as:

$$\mathcal{L}_c = \sum_{\mathcal{R}_{1:r1}} \{\mathcal{F}(\varphi(z^*), \varphi(\bar{z}))\} - \sum_{\mathcal{R}_{r2:r3}} \{\mathcal{F}(\varphi(z^*), \varphi(\bar{z}))\} \tag{5}$$

where $\mathcal{R}_{1:r1}$ denotes the ranking from 1st to $r1$, $\mathcal{R}_{r2:r3}$ denotes the ranking from $r2$ to $r3$ in the confidence candidates.

We also add attention mechanisms to confidence loss functions. The attention weight w_i can be used to distinguish candidates with different degrees of confidence, suppressing high confidence to excite low confidence. Thus, Equation (5) also can be rewritten as:

$$\mathcal{L}_c^A = \sum_{\mathcal{R}_{1:r1}} \{w_r \cdot \mathcal{F}(\varphi(z^*), \varphi(\bar{z}))\} - \sum_{\mathcal{R}_{r2:r3}} \{\mathcal{F}(\varphi(z^*), \varphi(\bar{z}))\} \tag{6}$$

We define attention confidence weight w_r as:

$$w_r = \frac{1}{a' + b' \cdot \tanh(c' \cdot d(\mathcal{R}_1, \mathcal{R}_r))} \tag{7}$$

where a', b', c' are controlling hyper-parameters. Specifically, a' and b' are use to constrain the attention weight w_r in the range $\left(\frac{1}{a'+b'}, \frac{1}{a'}\right)$, c' is shrinkage rate. $d(\mathcal{R}_1, \mathcal{R}_r)$ indicates the distance between the r-th confidence candidate $\mathcal{R}_r$'s corresponding predicted bounding box and the first confidence score's corresponding predicted bounding box in the confidence ranking.

3.2.3. Background Disturbance Loss

Considering that our attack method is specifically designed for remote sensing object tracking, we have observed that the characteristics of remote sensing videos, such as the presence of small targets, can impact the effectiveness of adversarial samples. To tackle this issue, we have introduced background disturbance loss, which takes into account the background information surrounding the target during the generation of perturbations. This allows us to mitigate the influence of the background and enhance the attack capability of the adversarial samples. We found that the regression branch of the tracker output feature maps $\mathcal{G}(\varphi(z^*), \varphi(\bar{z}))$ that measure the distance between pre-generated anchors and the corresponding ground truth. First, we select a set of K penalized scores of the predicted bounding box P_{TopK}. Then, we use the cosine window and $\mathcal{P}_{TopK}$ to re-rank the proposals' score to get the best one. In background disturbance loss, we selected bounding boxes set from outputs of the regression branch $\{(H_1, W_1), \cdots (H_K, W_K)\}$. We define the background interference loss as:

$$\mathcal{L}_b = \frac{1}{K} \sum_{k=1}^{k} (H_k + W_k) \tag{8}$$

where H_k and W_k are outputs of the regression branch. They refer to the height and width of the output candidate box, respectively.

3.2.4. Total Variation Loss

To optimize the generated perturbation to be imperceptive, we also introduce a total variation (TV) loss [48] to constrain each pixel to be smooth. TV Loss is often used as a regular term in the overall loss functions to constrain model learning, which can effectively promote the spatial smoothness of model output results. The TV loss can be represented as:

$$\mathcal{L}_{TV} = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left\{\left|{z^*}_{m+1,n} - {z^*}_{m,n}\right|^2 + \left|{z^*}_{m,n+1} - {z^*}_{m,n}\right|^2\right\}^{1/2} \tag{9}$$

where M and N denote the height and weight of the adversarial sample z^*, respectively. The above formula is only for a single image, and ${z^*}_{m,n}$ represents a pixel in the input image. We calculate each pixel ${z^*}_{m,n}$ of the horizontal direction (M and N) and vertical direction (image height N). Then, we calculate the square of the difference between each pixel and the next adjacent pixel ${z^*}_{m+1,n}$, ${z^*}_{m,n}$, as well as computing the square root and sum of all pixels.

Through the combined use of the above loss functions, our proposed attention-enhanced one-shot attack method can achieve the goal of quickly and efficiently generating one-shot perturbations that are imperceptive to the human eye. The generated perturbation is unique for each video. The total loss functions are as follows:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_f^A + \beta\mathcal{L}_c^A + \gamma\mathcal{L}_b + \lambda\mathcal{L}_{TV} \tag{10}$$

To date, the attention-enhanced one-shot attack proposed by us has been introduced. As shown in Algorithm 1, by using our method, after iteratively optimizing $\mathcal{L}_{total}$ and backpropagating the updated perturbation, we can obtain the final perturbation.

Algorithm 1 Attention-Enhanced One-Shot Attack Method

```
Input: Original template target patch $z$ and template search patch $\bar{z}$ of the UAV video,
    Victim tracker (Siamese Network $\varphi(\cdot)$, classification branch $\mathcal{F}(\cdot,\cdot)$ and regression
    branch $\mathcal{G}(\cdot,\cdot)$ of RPN), Iteration
Output: Adversarial template target patch $z^*$;
Initialize Adversarial template target patch $z^* = z$, Iter = 0;
Feed $z$ and $\bar{z}$ to $\varphi(\cdot)$ to get the corresponding feature maps $\varphi(z)$ and $\varphi(\bar{z})$;
Feed $\varphi(z)$ and $\varphi(\bar{z})$ to $\mathcal{F}(\cdot,\cdot)$ to get the original score rank $\mathcal{R}_{1:r}^{Ori}$;
while Iter < Iteration do
    Feed $z^*$ into $\varphi(\cdot)$ to get the corresponding feature maps $\varphi(z^*)$;
    Feed $\varphi(z^*)$ and $\varphi(\bar{z})$ to $\mathcal{F}(\cdot,\cdot)$ to get the adversarial score rank $\mathcal{R}_{1:r}$;
    if $\mathcal{R}_{1:r}[1] < \mathcal{R}_{1:r}^{Ori}[45]$ then
        Break;
    else
        Calculate attention feature loss $\mathcal{L}_f^A$ by Equation (3) using $\varphi(z^*)$ and $\varphi(z)$;
        Calculate attention confidence loss $\mathcal{L}_c^A$ by Equation (6) using $\mathcal{F}(\varphi(z^*), \varphi(\bar{z}))$;
        Calculate background disturbance loss $\mathcal{L}_b$ by Equation (8) using $\mathcal{G}(\varphi(z^*), \varphi(\bar{z}))$;
        Calculate TV loss $\mathcal{L}_{TV}$ by Equation (9) using $z^*$;
        Calculate total loss $\mathcal{L}_{total}$ by Equation (10) and backward it to update $z^*$;
    end if
    Iter = Iter + 1;
end while
```

Since our attack only occurs on the template frame. By simulating the tracking process in template frames, a unique perturbation for a certain video can be quickly generated with no more than 100 iterations. Moreover, our proposed method can attack trackers offline, which is more suitable for applications in real scenarios than online attacks. After the perturbation is generated, we feed the perturbation back into the tracker. When the tracker initializes the template frame, it inputs the perturbation of the corresponding video, which interferes with the matching of the template frame and the subsequent search frame. Eventually, it leads to the failure of tracking the target. The test of our offline perturbation process is shown in Figure 2.

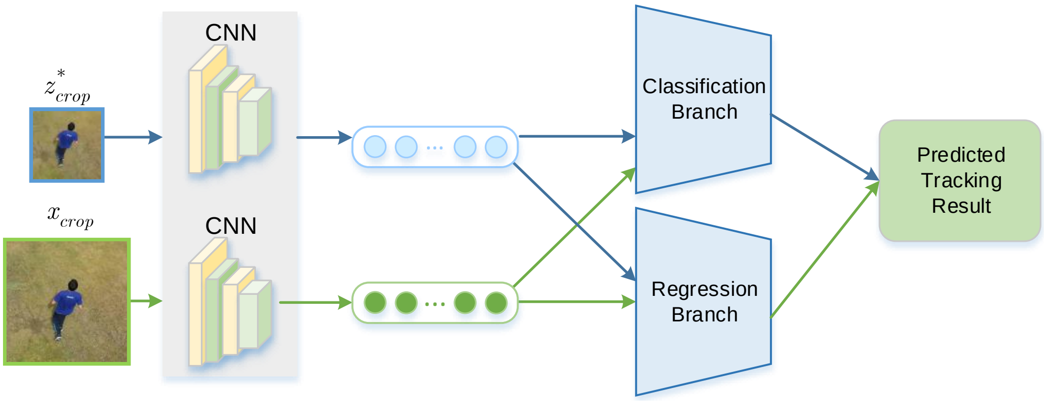


Figure 2. An illustration of the one-shot attack process on victim trackers using well-trained adversarial perturbations obtained by our proposed method.

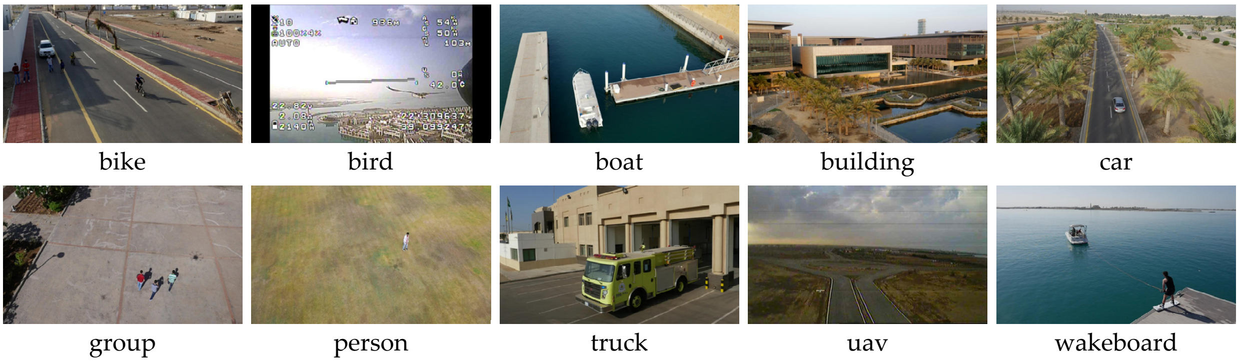
4. Experiments

This section presents our experimental setup and the results of our attention-enhanced one-shot attack on the UAV123 dataset. Next, we conduct ablation experiments to examine the impact of different loss functions on our attack method. Finally, we compare our method against classical attack methods.

4.1. Experimental Setup

4.1.1. Dataset

The UAV123 dataset [28] comprises 123 aerial UAV remote sensing video sequences captured by UAV, containing over 110,000 labeled image frames with target attributes and bounding boxes in each video sequence frame. The UAV123 dataset presents twelve types of challenge points for target tracking tasks due to variations in target scale, ambient illumination, and camera viewpoint across different video sequences, including target aspect ratio change, background interference, sudden camera movement, rapid target movement, complete target occlusion, illumination change, low resolution, target out of camera field of view, partial target occlusion, similar targets, target scale change, and camera viewpoint change. Furthermore, the dataset comprises various tracking targets in different scenes (e.g., city, road, and beach), including cars, persons, boats, and more, as illustrated in Figure 3.


bike bird boat building car

group person truck uav wakeboard

Figure 3. Sample images from different categories of videos in the UAV123 dataset.

4.1.2. Evaluation Metrics

Our proposed attention-enhanced one-shot attack method is evaluated based on two widely used metrics, namely success rate and precision. The success rate is calculated by measuring the intersection over union (IOU) between the predicted target bounding box and the ground truth bounding box and is deemed successful when the IOU value exceeds a pre-defined overlap threshold (default value of 0.5). Precision is measured by calculating the Euclidean distance between the center point coordinates of the predicted target frame and those of the ground truth bounding box and is considered successful if the distance is less than the set location error threshold (default is 20 pixels). The proposed method is deemed more effective for the attack of the tracker if the success rate and precision decrease more after the attack. Additionally, we utilize the one-pass evaluation (OPE) method provided by the PySoT toolkit, which can be accessed at https://github.com/StrangerZhang/pysot-toolkit (accessed on 20 July 2019). Specifically, we compare the tracking success rate under different overlap thresholds and tracking precision under different location error thresholds before and after the attack. Moreover, the PySOT toolkit enables separate evaluation of success rate and precision for various challenge points, such as illumination changes and low resolution, in the UAV123 dataset.

4.1.3. Implementation Details

The Adam [49] algorithm is used as a gradient descent optimizer to update the adversarial samples of the template frame images in each UAV video. The hyper-parameters a, b, and c used to calculate the feature attention weights (i.e., Equation (4)) are assigned values of 2, -1, and 20, respectively. The hyper-parameters a', b', and c' used to calculate the confidence attention weights (i.e., Equation (7)) are assigned values of 0.5, 1.5, and 0.2, respectively. In Equation (6), r_1, r_2, and r_3 are set to 45, 90, and 145, respectively. The maximum number of iterations is set to 2000 and the learning rate is set to 0.01 for the adversarial example generation for each video. To maintain training stability, α and β in Equation (10) are set to 0.01 and 10, respectively, for the first 200 iterations of each adversarial sample optimization training. For the remaining iterations, α and β are set to 0.5 and 0.1, respectively. γ and λ in Equation (10) are fixed to 0.1 and 0.01, respectively. We implement our proposed method using the Pytorch [50] framework with an NVIDIA GeForce GTX 3090 GPU.

4.2. Quantitative Attack Results

4.2.1. Overall Comparison

The attention-enhanced one-shot attack method is evaluated on widely used trackers, namely SiamRPN, DaSiamRPN, and SiamRPN++, employing diverse backbone networks, including AlexNet [51], ResNet-50 [29], and MobileNet-V2 [52]. The performance of these trackers is quantitatively assessed in terms of success rate and precision, as depicted in Table 1. Notably, the attack significantly diminishes both the success rate and precision of all the targeted trackers. Particularly, SiamRPN and DaSiamRPN are found to be more vulnerable to adversarial samples, as their success rates plummet from 59.5% and 57.8% to 17.6% and 26.7%, respectively. Additionally, their precision values also experience a substantial decrease, dropping from 79.3% and 78.7% to 25.6% and 39.2%, respectively. While SiamRPN++ exhibits a relatively lower susceptibility to the adversarial sample attack due to its intricate network architecture compared to SiamRPN and DaSiamRPN, our proposed method still exerts a significant influence on its tracking performance. Notably, even for SiamRPN++ (ResNet50), which boasts the most complex structure among the evaluated trackers, the success rate achieved after the attack remains below 40.0%.

Table 1. Overall comparison of the success rate and precision achieved by different trackers before and after being attacked by our proposed method.

Tracker	Precision (%)		Success Rate (%)	
	Original	Attack	Original	Attack
SiamRPN	79.3	25.6	59.5	17.6
DaSiamRPN	78.7	39.2	57.8	26.7
SiamRPN++(A)	77.4	42.6	58.0	29.0
SiamRPN++(R)	79.8	54.7	60.5	39.0
SiamRPN++(M)	79.3	57.4	59.8	37.6

Figure 4 depicts the variation in success rate for each tracker, both before and after applying the attack, across different overlap thresholds. Additionally, it showcases the precision variation at different location error thresholds. The results unequivocally demonstrate a significant reduction in both the success rate and tracking precision of each tracker when attacked by our proposed method. Notably, among the evaluated trackers, SiamRPN exhibits the most substantial degradation in terms of tracking success rate and precision, while DaSiamRPN ranks as the second most affected. For instance, at an overlap threshold of 0.6, the attacked SiamRPN experiences a drastic decrease in tracking success rate, declining from approximately 70% to around 10%. Similarly, the attacked DaSiamRPN witnessed a significant decline in tracking precision, dropping from roughly 73% to about 17% when the location error threshold is set at 10 pixels.

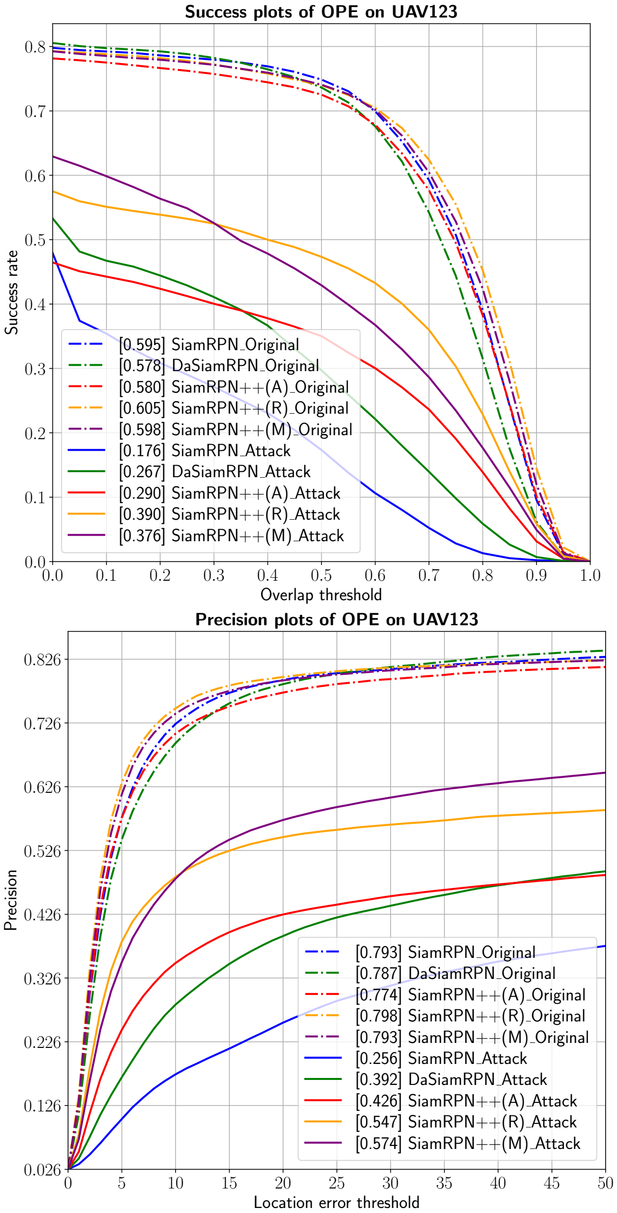


Figure 4. Comparison of the OPEs of various victim trackers before and after being attacked by our proposed method using the PySoT toolkit. (**Top**) Comparison of success rate with different overlap thresholds. (**Bottom**) Comparison of precision with different location error thresholds.

4.2.2. Comparison in Terms of Per-Class Success Rate and Precision

Figure 5 illustrates a detailed quantitative analysis of SiamRPN++(A)'s tracking success rate and precision for each video in the UAV123 dataset before and after the attack. The blue and orange lines represent the tracking results of the original and attacked SiamRPN++(A), respectively. The results indicate that the attack performance of SiamRPN++(A) is significantly impacted across most video categories, with the attack's effect being more noticeable on videos such as UAVs, wakeboards, trucks, and groups, and to a lesser extent on videos such as boats. The blue line, occupying a more significant proportion of the plot, indicates that our attack method is generally effective in tracking various types of videos.

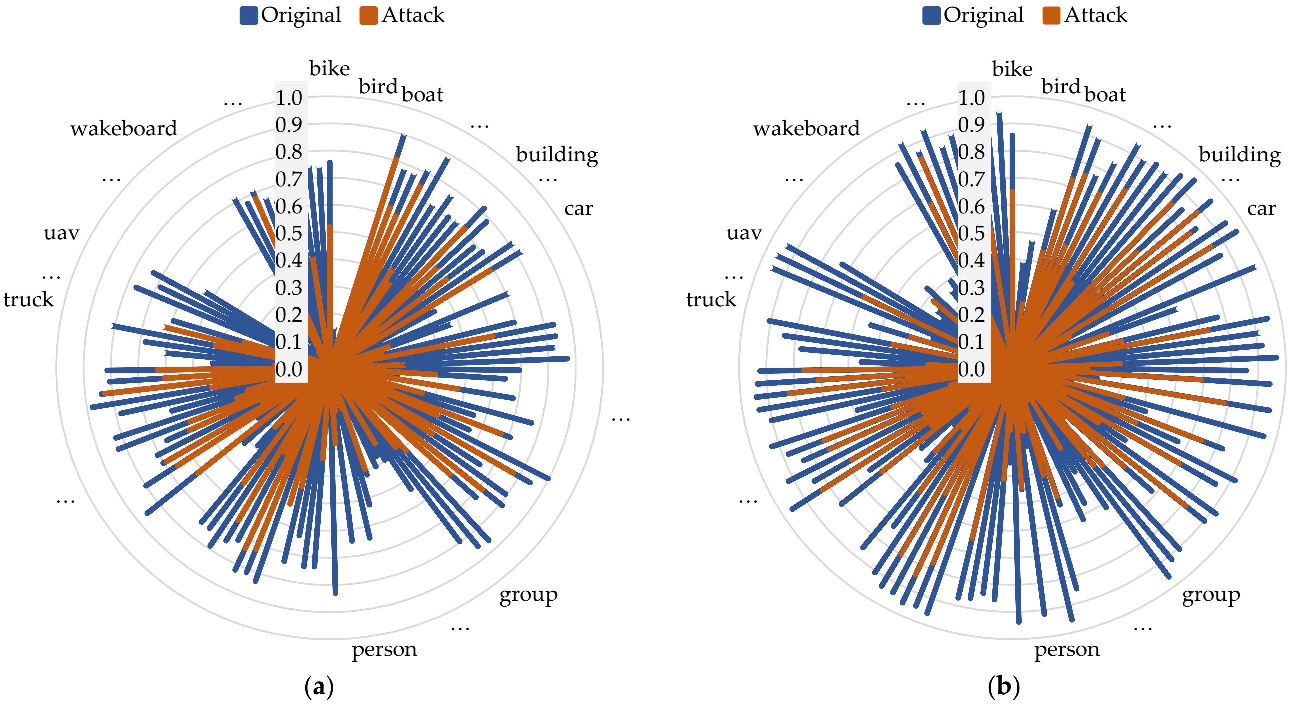


Figure 5. Per-video comparisons of the tracking (**a**) success rates and (**b**) precisions obtained by SiamRPN++ (A) before and after being attacked.

4.2.3. Comparison in Terms of Different Challenge Points

The UAV123 dataset is constructed by identifying multiple challenge points that encompass different aspects of the UAV remote video tracking task, as described in Section 4.1.1. Here, we present the impact of the proposed attack method on the tracker for each of these challenge points, as shown in Table 2. The results reveal that the proposed approach significantly decreases the success rate and precision of the tracker for all 12 challenge points. Notably, the attack the on camera motion has the most significant impact, reducing the tracking success rate from 58.8% to 25.9% and precision from 78.2% to 37.0%. Similarly, the effect of the attack on illumination variation, low resolution, viewpoint change, and other factors is also evident, with the tracking success rate decreasing from the original 53.0% to 22.8%, and the precision decreasing from 72.3% to 33.4%. While the impact of the attack on full occlusion is relatively smaller, the tracking success rate still drops to nearly half of the original. In summary, the proposed attack method is effective across the board and has a significant impact on different challenge points of the UAV remote sensing video tracking task.

Table 2. Comparison of the tracking success rate and accuracy achieved by SiamRPN++(A) for various challenge points of UAV remote sensing target tracking in the UAV123 dataset, before and after being attacked by our proposed method.

Challenge Point	Success Rate (%)		Precision (%)	
	Original	Attack	Original	Attack
Aspect Ratio Change	54.2	24.9	74.1	36.3
Background Clutters	41.9	16.7	61.5	30.0
Camera Motion	58.8	25.9	78.2	37.0
Fast Motion	49.0	20.9	67.7	30.9
Full Occlusion	32.9	18.3	54.9	37.8
Illumination Variation	53.0	22.8	72.3	33.4
Low Resolution	42.3	12.0	65.0	25.8
Out of View	51.3	27.8	69.4	41.7
Partial Occlusion	48.0	24.3	67.3	38.0
Scale Variation	55.5	27.5	74.6	40.3
Similar Object	50.8	27.5	71.1	43.7
Viewpoint Change	59.1	29.5	77.5	39.9

4.3. Visual Results

Figures 6 and 7 depict the visual tracking results of attacking SiamRPN++(A) using our attention-enhanced one-shot method on the car and wakeboard videos, respectively. The bounding boxes depicted in green represent the ground truth, while the blue bounding boxes correspond to the predicted target bounding box. In the car video, the bounding box predicted by the post-attack tracker deviates from the correct trajectory, sometimes incorrectly tracking the car as an unspecified tracking target, and at other times incorrectly predicting it as a non-target-related background area next to the highway. In addition, possibly influenced by our proposed background interference loss, the post-attack tracker mainly predicts the bounding box on the wave region around the labeled target in the wakeboard video. Overall, our proposed attack method can cause the tracker to underperform significantly and effectively on the UAV remote sensing video target.

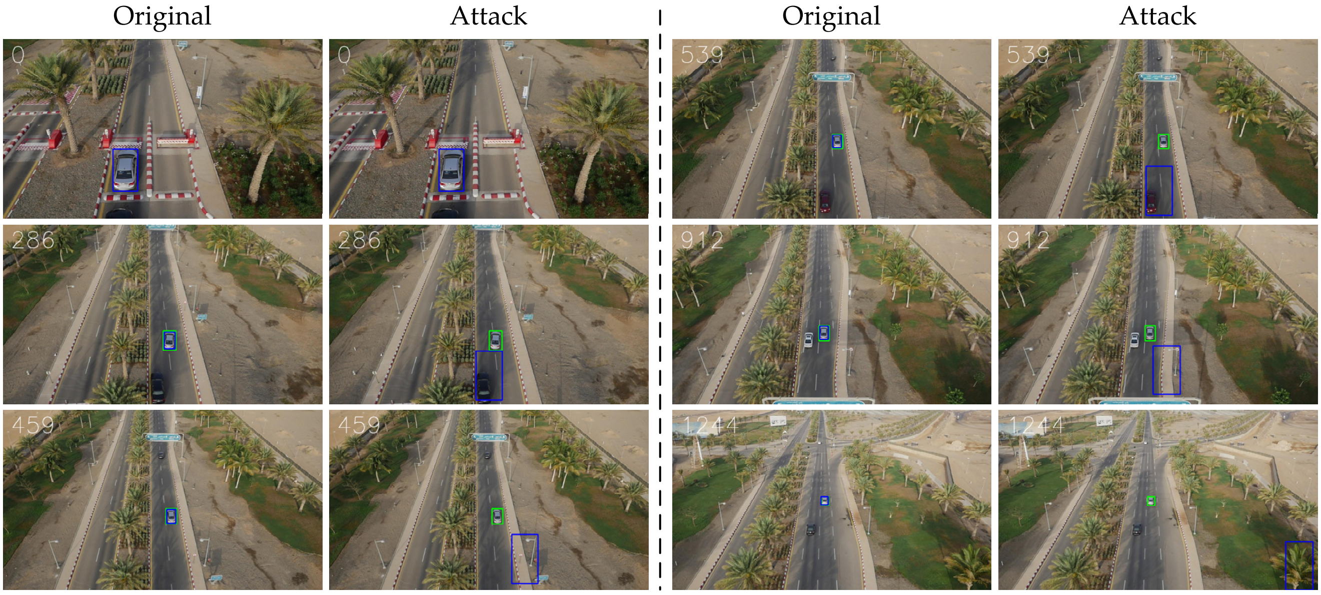


Figure 6. Visual comparison of the tracking results of the car video using SiamRPN++ (A) before and after being attacked by our proposed method. The green and blue boxes indicate the ground truth bounding box and the bounding box predicted by the tracker, respectively.

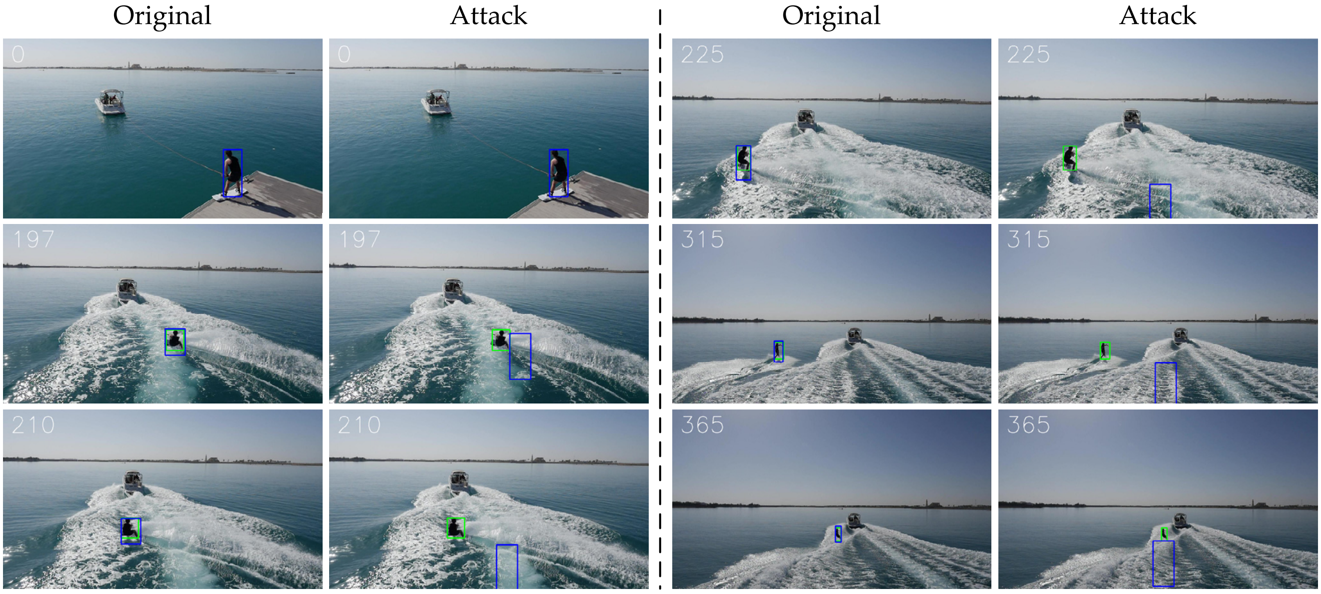


Figure 7. Visual comparison of the tracking results of the wakeboard video using SiamRPN++ (A) before and after being attacked by our proposed method. The green and blue boxes indicate the ground truth bounding box and the bounding box predicted by the tracker, respectively.

Figure 8 provides a comparison between the heatmaps generated from the original search frame images and the heatmaps obtained from the search frame images attacked by our proposed method. The top row displays the original search frame images, while the second row exhibits the corresponding heatmaps derived from these images. The bottom row shows the adversarial heatmaps generated from the attacked search frame images. It is evident that the resulting adversarial heatmaps exhibit minimal activity or relevance in the target region, leading to the subsequent failure of the tracker. This observation highlights the effectiveness of our attention-enhanced one-shot attack method in both perturbing the search region and deceiving the tracker. By significantly reducing the heatmap values within the target area, our approach successfully compromises the tracker's ability to accurately locate and track the intended target. This comparative analysis demonstrates the improved performance of our attention-enhanced one-shot attack method in terms of its impact on the search region and its ability to deceive the tracker, ultimately leading to diminished tracking accuracy.

4.4. Ablation Studies

We perform a series of ablation experiments to assess the impact of each loss function on our attack effectiveness, using SiamRPN++ (R) as the victim model. Table 3 presents the results, which indicate that using only $\mathcal{L}_f^A$ or $\mathcal{L}_b$ significantly reduces success rate and precision. The success rates are reduced by 15.1% and 18.9%, respectively, and the precisions are 62.2% and 57.4%, respectively. Moreover, $\mathcal{L}_b$ has shown better attack performance than $\mathcal{L}_f^A$, suggesting that $\mathcal{L}_b$ has a stronger attack effect. Combining $\mathcal{L}_c^A$ and $\mathcal{L}_f^A$ also contributed to a further reduction in precision and success rate. Our experiments reveal that using all three loss functions simultaneously yields the best performance, further demonstrating the effectiveness of our attention-enhanced one-shot attack. Additionally, we examine the visual effect of $\mathcal{L}_{TV}$. Figure 9 illustrates the visual difference between adversarial samples generated with and without $\mathcal{L}_{TV}$. In our preliminary experiments, it became evident that the utilization of the L norm led to perturbations that exhibited a high degree of pixelation. The presence of pixelation significantly impacted the natural appearance of the generated adversarial samples. To address this issue, we investigated the application of $\mathcal{L}_{TV}$, as it offered the potential for generating smoother perturbations. Figure 9 visually demonstrates the impact of employing $\mathcal{L}_{TV}$ in comparison to the adversarial sample without using $\mathcal{L}_{TV}$.

In the first row of the figure, the adversarial samples generated without $\mathcal{L}_{TV}$ are presented, highlighting the presence of noticeable ripples, particularly in the building and person images. Conversely, the second row of the figure showcases the same set of adversarial samples that are optimized using $\mathcal{L}_{TV}$, resulting in significantly improved naturalness and a reduction in the aforementioned ripples. These findings underscore the effectiveness of employing $\mathcal{L}_{TV}$ as a means to enhance the visual quality and authenticity of adversarial samples, mitigating the issues associated with pixelation that arise when relying solely on the L norm. Furthermore, the $\mathcal{L}_{TV}$ has notably enhanced the visual quality while generally preserving a slight reduction in attack effectiveness, as demonstrated in the final row of Table 3.

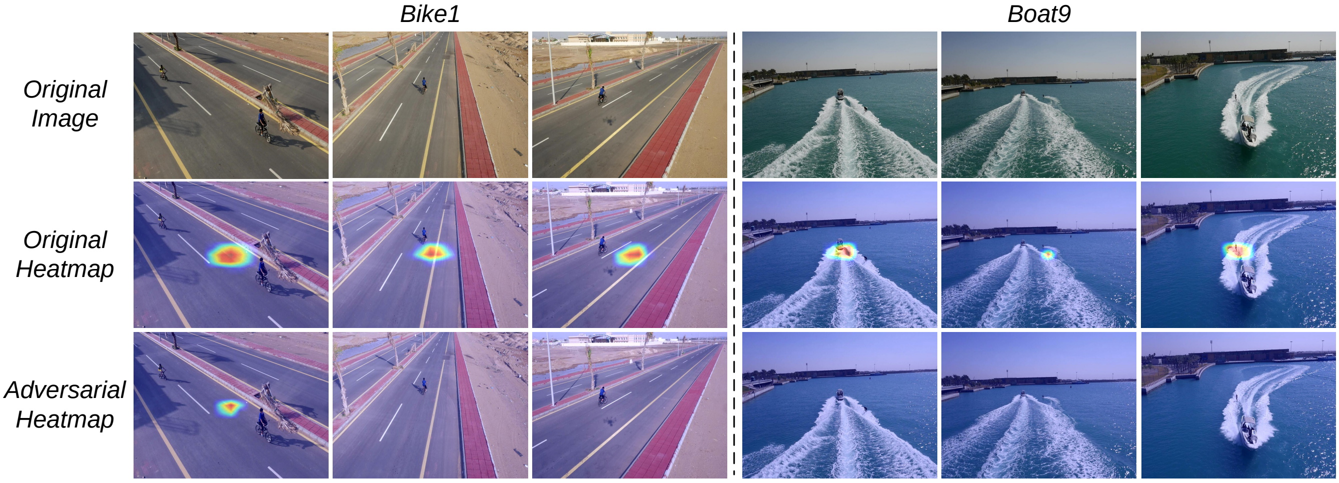


Figure 8. Heatmap comparison of SiamRPN++ (A) for tracking bike targets and boat targets, before and after being attacked by our proposed method.

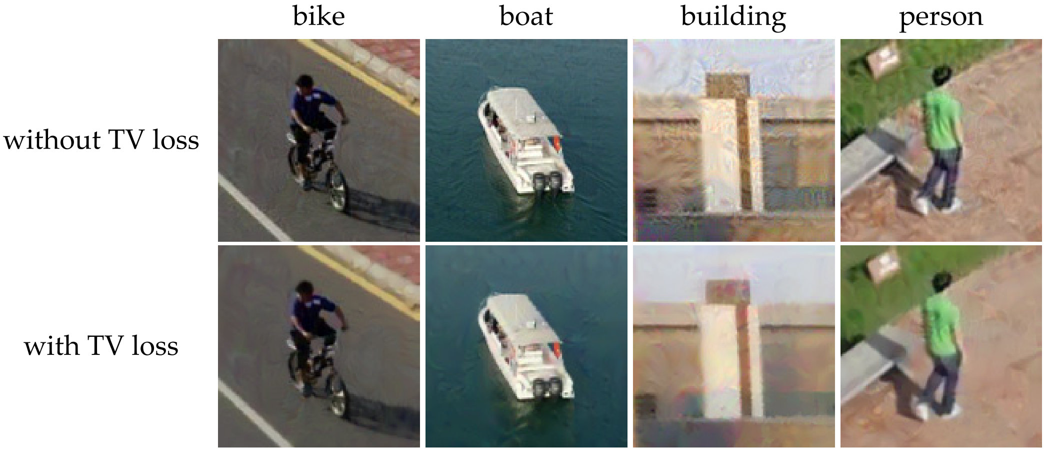


Figure 9. Effect of $\mathcal{L}_{TV}$ on the visual appearance of adversarial examples generated by our proposed method.

Table 3. Ablation comparison of $\mathcal{L}_f^A$, $\mathcal{L}_c^A$, $\mathcal{L}_b$ and $\mathcal{L}_{TV}$ in our proposed method, in terms of their impact on tracking success rate and precision.

SiamRPN++ (ResNet-50)	Precision (%)	Success Rate (%)
Original	79.8	60.5
Random Noise	77.1	58.3
Attack by $\mathcal{L}_f$	62.2	45.4
Attack by $\mathcal{L}_c$	74.6	54.7
Attack by $\mathcal{L}_b$	57.4	41.6
Attack by $\mathcal{L}_f$+$\mathcal{L}_c$	59.4	43.1

Table 3. *Cont.*

SiamRPN++ (ResNet-50)	Precision (%)	Success Rate (%)
Attack by $\mathcal{L}_f$+$\mathcal{L}_b$	55.7	39.8
Attack by $\mathcal{L}_c$+$\mathcal{L}_b$	55.6	39.7
Attack by $\mathcal{L}_f$+$\mathcal{L}_c$+$\mathcal{L}_b$	54.5	38.4
Attack by $\mathcal{L}_f$+$\mathcal{L}_c$+$\mathcal{L}_b$+$\mathcal{L}_{TV}$	54.7	39.0

4.5. Comparison with SOTA Methods

To assess the attack capability of our attention-enhanced one-shot attack, we conducted a comparative analysis with the classic algorithms FGSM and C&W. The evaluation was performed on the UAV123 dataset, employing SiamRPN and SiamRPN++(R) as the victim models. Table 4 presents the results of this comparative study, demonstrating the strong attack performance of our proposed method. The success rates achieved 17.6% for SiamRPN and 39.0% for SiamRPN++(R) by being attacked by our attention-enhanced one-shot method. Notably, these success rates are lower than those attacked by FGSM and C&W. Therefore, our method exhibits superior attack strength in comparison to the FGSM and C&W algorithms. These findings underscore the effectiveness of our proposed method in generating adversarial perturbations that can successfully evade the tracking capabilities of the targeted SiamRPN and SiamRPN++(R) models.

Table 4. Comparison of tracking success rate and precision obtained by SiamRPN and SiamRPN++(R) before and after being attacked by FGSM, C&W, and our proposed method.

Method	SiamRPN		SiamRPN++(R)	
	Precision (%)	Success Rate (%)	Precision (%)	Success Rate (%)
Original	79.3	59.5	79.8	60.5
FGSM	56.2	45.7	71.8	51.4
C&W	47.9	44.8	59.3	43.6
Proposed	25.6	17.6	54.7	39.0

5. Conclusions

In this paper, we proposed a novel attention-enhanced one-shot attack on UAV remote sensing images for generating template adversarial examples that can deceive trackers. The proposed method generates optimized unique perturbations for each video by perturbing the features near templates and suppressing the confidence scores that activate the tracker. The attention-enhanced one-shot attack mainly used the template frame to simulate the tracking process and generated adversarial samples by injecting perturbations into the template frame, causing the tracking of subsequent frames to fail. The experimental results demonstrated that our attention-enhanced one-shot attack is more effective than classical algorithms, can rapidly generate offline adversarial samples, and deceives the widely used trackers such as SiamRPN, DasiamRPN, and SiamRPN++. In our future work, we plan to study the adversarial attack on UAV remote sensing images of multiple tracking.

Author Contributions: Conceptualization, Y.J.; methodology, Y.J.; software, Y.J.; validation, Y.J.; formal analysis, Y.J.; investigation, Y.J.; resources, Y.J.; writing—original draft preparation, Y.J.; writing—review and editing, Y.J.; visualization, Y.J.; supervision, G.Y.; project administration; G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The UAV123 dataset used in this study can be accessed on https://cemse.kaust.edu.sa/ivul/uav123 (accessed on 17 September 2016).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gaffey, C.; Bhardwaj, A. Applications of Unmanned Aerial Vehicles in Cryosphere: Latest Advances and Prospects. *Remote Sens.* **2020**, *12*, 948. [CrossRef]
2. Osco, L.P.; Junior, J.M.; Ramos, A.P.M.; de Castro Jorge, L.A.; Fatholahi, S.N.; de Andrade Silva, J.; Matsubara, E.T.; Pistori, H.; Gonçalves, W.N.; Li, J. A Review on Deep Learning in UAV Remote Sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102456. [CrossRef]
3. Cherif, E.; Hell, M.; Brandmeier, M. DeepForest: Novel Deep Learning Models for Land Use and Land Cover Classification Using Multi-Temporal and -Modal Sentinel Data of the Amazon Basin. *Remote Sens.* **2022**, *14*, 5000. [CrossRef]
4. Qian, X.; Zhang, N.; Wang, W. Smooth GIoU Loss for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1259. [CrossRef]
5. Wang, Z.; Wang, B.; Liu, Y.; Guo, J. Global Feature Attention Network: Addressing the Threat of Adversarial Attack for Aerial Image Semantic Segmentation. *Remote Sens.* **2023**, *15*, 1325. [CrossRef]
6. Wu, D.; Song, H.; Fan, C. Object Tracking in Satellite Videos Based on Improved Kernel Correlation Filter Assisted by Road Information. *Remote Sens.* **2022**, *14*, 4215. [CrossRef]
7. Wu, J.; Cao, C.; Zhou, Y.; Zeng, X.; Feng, Z.; Wu, Q.; Huang, Z. Multiple Ship Tracking in Remote Sensing Images Using Deep Learning. *Remote Sens.* **2021**, *13*, 3601. [CrossRef]
8. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully Convolutional Siamese Networks for Object Tracking. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Volume 9914, pp. 850–865. ._56. [CrossRef]
9. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking With Siamese Region Proposal Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8971–8980. [CrossRef]
10. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-Aware Siamese Networks for Visual Object Tracking. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11213, pp. 103–119. [CrossRef]
11. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4282–4291. [CrossRef]
12. Cui, Y.; Song, T.; Wu, G.; Wang, L. MixFormerV2: Efficient Fully Transformer Tracking. *arXiv* **2023**, arXiv:2305.15896. [CrossRef].
13. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020—Demos, Online, 16–20 November 2020; Liu, Q., Schlangen, D., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA , 2020; pp. 38–45. [CrossRef]
14. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
15. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
16. Carlini, N.; Wagner, D.A. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57. [CrossRef]
17. Thys, S.; Ranst, W.V.; Goedemé, T. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 49–55. [CrossRef]
18. Gao, R.; Guo, Q.; Juefei-Xu, F.; Yu, H.; Fu, H.; Feng, W.; Liu, Y.; Wang, S. Can You Spot the Chameleon? Adversarially Camouflaging Images from Co-Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 2140–2149. [CrossRef]
19. Liao, Q.; Wang, X.; Kong, B.; Lyu, S.; Yin, Y.; Song, Q.; Wu, X. Fast Local Attack: Generating Local Adversarial Examples for Object Detectors. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
20. Mei, H.; Ji, G.; Wei, Z.; Yang, X.; Wei, X.; Fan, D. Camouflaged Object Segmentation With Distraction Mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 19–25 June 2021; pp. 8772–8781. [CrossRef]
21. Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A.L. Adversarial Examples for Semantic Segmentation and Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1378–1387. [CrossRef]
22. Yang, J.; Xu, R.; Li, R.; Qi, X.; Shen, X.; Li, G.; Lin, L. An Adversarial Perturbation Oriented Domain Adaptation Approach for Semantic Segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New York, NY, USA, 7–12 February 2020; pp. 12613–12620.

23. Jia, S.; Ma, C.; Song, Y.; Yang, X. Robust Tracking Against Adversarial Attacks. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Volume 12364, pp. 69–84. [CrossRef]
24. Guo, Q.; Xie, X.; Juefei-Xu, F.; Ma, L.; Li, Z.; Xue, W.; Feng, W.; Liu, Y. SPARK: Spatial-Aware Online Incremental Attack Against Visual Tracking. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Volume 12370, pp. 202–219. [CrossRef]
25. Liang, S.; Wei, X.; Yao, S.; Cao, X. Efficient Adversarial Attacks for Visual Object Tracking. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Volume 12371, pp. 34–50. [CrossRef]
26. Yan, X.; Chen, X.; Jiang, Y.; Xia, S.; Zhao, Y.; Zheng, F. Hijacking Tracker: A Powerful Adversarial Attack on Visual Tracking. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2897–2901. [CrossRef]
27. Nakka, K.K.; Salzmann, M. Temporally Transferable Perturbations: Efficient, One-Shot Adversarial Attacks for Online Visual Object Trackers. *arXiv* **2020**, arXiv:2012.15183.
28. Mueller, M.; Smith, N.; Ghanem, B. A Benchmark and Simulator for UAV Tracking. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 445–461. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
30. Li, D.; Zhang, J.; Huang, K. Universal Adversarial Perturbations Against Object Detection. *Pattern Recognit.* **2021**, *110*, 107584. [CrossRef]
31. Wang, Y.; Lv, H.; Kuang, X.; Zhao, G.; Tan, Y.; Zhang, Q.; Hu, J. Towards a Physical-World Adversarial Patch for Blinding Object Detection Models. *Inf. Sci.* **2021**, *556*, 459–471. [CrossRef]
32. Yan, B.; Wang, D.; Lu, H.; Yang, X. Cooling-Shrinking Attack: Blinding the Tracker With Imperceptible Noises. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 987–996. [CrossRef]
33. Guo, Q.; Cheng, Z.; Juefei-Xu, F.; Ma, L.; Xie, X.; Liu, Y.; Zhao, J. Learning to Adversarially Blur Visual Object Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10819–10828. [CrossRef]
34. Ding, L.; Wang, Y.; Yuan, K.; Jiang, M.; Wang, P.; Huang, H.; Wang, Z.J. Towards Universal Physical Attacks on Single Object Tracking. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Online, 2–9 February 2021; pp. 1236–1245.
35. Xu, Y.; Bai, T.; Yu, W.; Chang, S.; Atkinson, P.M.; Ghamisi, P. AI Security for Geoscience and Remote Sensing: Challenges and Future Trends. *IEEE Geosci. Remote. Sens. Mag.* **2023**, *11*, 60–85. [CrossRef]
36. Czaja, W.; Fendley, N.; Pekala, M.J.; Ratto, C.; Wang, I. Adversarial Examples in Remote Sensing. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL), Seattle, WA, USA, 6–9 November 2018; pp. 408–411. [CrossRef]
37. Chen, L.; Zhu, G.; Li, Q.; Li, H. Adversarial Example in Remote Sensing Image Recognition. *arXiv* **2019**, arXiv:1910.13222.
38. Bai, T.; Wang, H.; Wen, B. Targeted Universal Adversarial Examples for Remote Sensing. *Remote Sens.* **2022**, *14*, 5833. [CrossRef]
39. Xu, Y.; Ghamisi, P. Universal Adversarial Examples in Remote Sensing: Methodology and Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
40. Xu, Y.; Du, B.; Zhang, L. Assessing the Threat of Adversarial Examples on Deep Neural Networks for Remote Sensing Scene Classification: Attacks and Defenses. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1604–1617. [CrossRef]
41. Chen, L.; Xiao, J.; Zou, P.; Li, H. Lie to Me: A Soft Threshold Defense Method for Adversarial Examples of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
42. Sun, X.; Cheng, G.; Pei, L.; Li, H.; Han, J. Threatening Patch Attacks on Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–10. [CrossRef]
43. Lu, M.; Li, Q.; Chen, L.; Li, H. Scale-Adaptive Adversarial Patch Attack for Remote Sensing Image Aircraft Detection. *Remote Sens.* **2021**, *13*, 4078. [CrossRef]
44. Zhang, Y.; Zhang, Y.; Qi, J.; Bin, K.; Wen, H.; Tong, X.; Zhong, P. Adversarial Patch Attack on Multi-Scale Object Detection for UAV Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5298. [CrossRef]
45. Deng, B.; Zhang, D.; Dong, F.; Zhang, J.; Shafiq, M.; Gu, Z. Rust-Style Patch: A Physical and Naturalistic Camouflage Attacks on Object Detector for Remote Sensing Images. *Remote Sens.* **2023**, *15*, 885. [CrossRef]
46. Wei, X.; Yuan, M. Adversarial pan-sharpening attacks for object detection in remote sensing. *Pattern Recognit.* **2023**, *139*, 109466. [CrossRef]
47. Fu, C.; Li, S.; Yuan, X.; Ye, J.; Cao, Z.; Ding, F. Ad2Attack: Adaptive Adversarial Attack on Real-Time UAV Tracking. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022.
48. Mahendran, A.; Vedaldi, A. Understanding Deep Image Representations by Inverting Them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5188–5196. [CrossRef]
49. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.

50. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Anitiga, L.; Desmaison, A.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
51. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
52. Sandler, M.; Howard, A.G.; Zhu, M.; Zhmoginov, A.; Chen, L. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.