One-shot Adversarial Attacks on Visual Tracking with Dual Attention

Xuesong Chen* [1] Xiyu Yan* [2] Feng Zheng† [3] Yong Jiang[2, 4]
Shu-Tao Xia[2, 4] Yong Zhao[1] Rongrong Ji [4, 5]
[1]Peking University, School of ECE [2]Tsinghua University
[3]Southern University of Science and Technology
[4]Peng Cheng Laboratory [5]Xiamen University
cedarchen@pku.edu.cn, yanqy17@mails.tsinghua.edu.cn, zhengf@sustech.edu.cn

Abstract

Almost all adversarial attacks in computer vision are aimed at pre-known object categories, which could be off-line trained for generating perturbations. But as for visual object tracking, the tracked target categories are normally unknown in advance. However, the tracking algorithms also have potential risks of being attacked, which could be maliciously used to fool the surveillance systems. Meanwhile, it is still a challenging task that adversarial attacks on tracking since it has the free-model tracked target. Therefore, to help draw more attention to the potential risks, we study adversarial attacks on tracking algorithms. In this paper, we propose a novel one-shot adversarial attack method to generate adversarial examples for free-model single object tracking, where merely adding slight perturbations on the target patch in the initial frame causes state-of-the-art trackers to lose the target in subsequent frames. Specifically, the optimization objective of the proposed attack consists of two components and leverages the dual attention mechanisms. The first component adopts a targeted attack strategy by optimizing the batch confidence loss with confidence attention while the second one applies a general perturbation strategy by optimizing the feature loss with channel attention. Experimental results show that our approach can significantly lower the accuracy of the most advanced Siamese network-based trackers on three benchmarks.

1. Introduction

Visual Object Tracking (VOT) plays a significant role in practical security applications such as intelligent surveillance systems. Recent years have witnessed many breakthroughs in visual object tracking algorithms [2, 25, 5, 17, 28, 16] brought by the progress of deep learning. For example, the SiamRPN++ tracker [16] based on Siamese network has reached 91% precision on the OTB100 benchmark [30]. However, whether the deep learning-based object tracking algorithms are as powerful as they seem is a question to worth pondering.

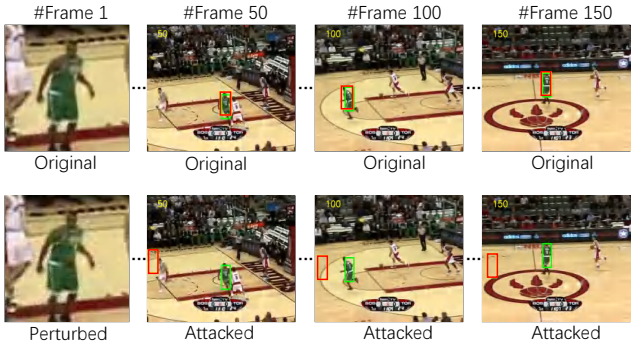

Figure 1. We only perturb slightly the target patch in the initial frame of a video, resulting in tracking failure in subsequent frames. First line: the original video frames are successfully tracked. Second line: attacking the target only in the initial frame could paralyze the tracker. The green boxes represent the ground truth, and the red boxes represent the tracking result of the tracker.

Adversarial attacks on deep learning models in computer vision have attracted increasing interest in the past years [1]. There are many adversarial attacks against deep networks successfully fooling image classifiers and object detectors. For example, Szegedy *et al*. demonstrated that putting small perturbations in images that remain (almost) imperceptible to the human visual system could fool deep learning models into misclassification [24]. Most recently, [26] created a small adversarial patch that is used as a cloaking device to hide persons from a person detector. Commonly, almost all these attacks are not aimed at free-models (*i.e*. arbitrary target) but the pre-known categories. Actually, adding adversarial perturbations on the free-model target patch in a cer-

*Equal contributions. This work is done when Xuesong Chen and Xiyu Yan visited to Feng Zheng Lab in SUSTech.
†Corresponding author.

tain frame may cause state-of-the-art trackers to lose the target in subsequent frames, which could be maliciously used to fool surveillance systems. Thus, it is necessary to study adversarial attacks on visual object tracking algorithms to help improve their prevention against these potential risks.

However, attacking a tracker to lose the object in continuous video frames is a challenging task. First, online visual tracking is unable to pre-know the category of tracked and to learn beforehand because of the free-model tacked target and the continuous video frames. Secondly, it is difficult to set an optimization objective to generate adversarial examples since a successful attack on the tracking task is significantly different from that on the multi-classification task which could merely maximize the probability of the class with the second-highest confidence. Specifically, the tracking task in each frame is the same as that of classifying all candidate boxes into one positive sample and the others into negative samples. Such a special binary classification problem makes it difficult to perform a successful attack if only one candidate box is selected to increase its confidence.

To address these challenges, in this paper, we study the adversarial attacks against visual object tracking. Our attack target is a series of excellent trackers based on Siamese networks, in which the tracking accuracy and efficiency are well-balanced due to the unique advantages of off-line learning and the abandonment of similarity updating. For these trackers, we propose a one-shot attack framework—only slightly perturbing the pixel values of the object patch in the initial frame of a video, which achieves the goal of attacking the tracker in subsequent frames, *i.e*. failure to the tracking of SiamRPN (see Fig. 1).

Specifically, a novel attack method with dual losses and dual attention mechanisms is explored to generate adversarial perturbation on the target patch at the initial frame. Our optimization objective of the proposed attack method consists of two components, and each loss is combined with its corresponding well-designed attention weight to further improve attack abilities. On the one hand, we formulate such Siamese network-based tracking problems as a specific classification task—the candidates of tracking are treated as the labels of classification, for a successful matching of the target template and the candidate box with the maximum confidence. Thus, we can pertinently perturb with a tracker to make it match "the best box". Here, we optimize the batch confidence loss by suppressing the confidence of excellent candidates and stimulate that of moderate candidates. To further distinguish the high-quality candidate boxes, the distance-oriented attention mechanism is adopted to widen the distance between excellent candidates. On the other hand, we apply a general perturbation strategy by optimizing the feature loss that maximizes the distance between the clean image and its adversarial example in the feature space for a powerful attack. To further ensure the generalization ability of the one-shot attack, the feature channel-wise activation-oriented attention of feature maps is taken into account under limited perturbation conditions.

Eventually, we evaluate our attacks on three tracking benchmarks, including OTB100 [30], LaSOT [4], and GOT10K [11]. The experimental results show that our approaches can significantly lower the accuracy of the most advanced Siamese network-based trackers.

In summary, the key contributions of this paper are as follows.

- To the best of our knowledge, we are the first to study one-shot adversarial attacks against VOT. The proposed one-shot attack method against the trackers based on Siamese networks can make them fail to track in a video by only disturbing the initial frame.
- We present a new optimization objective function with dual attention mechanisms to generate adversarial perturbations for ensuring the efficiency of the one-shot attack.
- Experimental results on three popular benchmarks show that our method is able to significantly lower the accuracy of the state-of-the-art Siamese network-based trackers.

2. Background and Related Work

In this section, we first briefly describe the background of adversarial attack problems. Next, the development of adversarial attack methods in computer vision (CV) tasks is reviewed. Lastly, we discuss the trackers based on Siamese networks that are adopted as our attack targets in this work.

2.1. Background of Adversarial Attacks

It is necessary to introduce some common technical terms related to the adversarial attacks on deep learning models in computer vision and the remaining paper also follows the same definitions of the terms.

***Adversarial example*.** It is a concept related to a natural clean example and is obtained by a specific algorithm processing to make the incorrect decision of models. It can be generated by global pixel perturbations of clean samples, or by adding adversarial patches to clean samples. The global pixel perturbation is applied to our work.

***Adversarial attacks*.** According to the degree of the attacker's understanding of the model, it can be classified into *white-box attacks* and *black-box attacks*. Also, through the target attacked by the attacker, it can be divided into *targeted attacks* and *non-targeted attacks*.

***White-box attacks*.** It means that when the attackers know all the knowledge of the model, including the structure, the parameters and the values of the trainable weights of the

neural network model, they can generate adversarial examples to mislead the model.

Black-box attacks. It means that, when the attackers only have limited or no information about the model, they construct adversarial examples that can fool most machine learning models.

Targeted attacks. It is usually used to attack classifiers. In this case, the attacker wants to change the prediction result to some specified target category.

Non-targeted attacks. On the contrary, in this case, the goal of attackers is simply to make the classifier give false predictions, regardless of which category the error classification becomes. Our attack is in the middle of these two cases.

Our work focuses on *white-box*, *test-time attacks* on visual object tracking algorithms, and other families of attacks not directly relevant to our setting are not discussed here.

2.2. Adversarial Attacks in CV Tasks

Szegedy et al. [24] first propose to generate adversarial examples for classification models that successfully misleading the classifier. Following that, Goodfellow et al. [7] extend this line and create a Fast Gradient Sign Method (FGSM) to generate adversarial attacks on images. Besides, attack methods based on the gradient include BIM [15], JSMA [22], DFool [20], Carlini and Wagner Attacks (C&W) [3], etc. Most of these attacks are directed at image classification that is the most basic visual task.

Recently, there are several explorations of the adversarial attacks on some high-level tasks, such as semantic segmentation and object detection. For example, [31] first transforms an attack task into a generation task and proposes a Dense Adversary Generation (DAG) method to optimize the loss function for the generation of adversarial examples, and then uses the generated adversarial example to attack the segmentation and detection models based on the deep network. This transformation makes the attacks no longer limited to the traditional gradient-based algorithms but also introduces more generation models, such as GAN. Then, [26] presents an approach to generate adversarial patches to targets with lots of intra-class variety and successfully hide a person from a person detector.

Most recently, PAT [29] and SPARK [9] generate adversarial samples against VOT through iterative optimization on video frames. However, the attack strategy with online iterative restricts their application scenarios. First, to generate adversarial sequences, they always need to access to the weights of the models during the attack. Second, the forward-backward propagation iteration is difficult to meet the real-time requirements of the tracking task.

2.3. Siamese Network-based Tracking

Visual Object Tracking (VOT) aims to predict the position and size of an object in a video sequence after the target has been specified in the first frame [18]. Recently, the Siamese network-based trackers [25, 2, 8, 32, 27, 10] have drawn significant attention due to their simplicity and effectiveness. Bertinetto et al. [2] first proposed a network structure based on Siamese fully convolutional networks for object tracking (SiamFC). Since then, many state-of-the-art algorithms of tracking have been proposed by researchers [32, 10, 17, 16, 28]. For example, the representative tracker—SiamRPN [17] introduces a regional recommendation network after the Siamese network and combines classification and regression for tracking.

These Siamese trackers formulate the VOT problem as a cross-correlation problem and learn a tracking similarity map from deep models with a Siamese network structure, one branch for learning the feature presentation of the target, and the other one for the search area. To ensure tracking efficiency, the offline learned siamese similarity function is often fixed during the running time. Meanwhile, the target template is acquired in the initial frame and remains unchanged in the subsequent video frames.

In the tracking phase of each frame, the target template and the search region including several candidate boxes are fed into the Siamese network to generate a confidence map that represents the confidences of the candidate boxes. It is worth noting that the Gaussian windows are widely applied to refine the confidences of the candidate boxes on the inference phase in tracking tasks. Different from the Non-Maximum Suppression (NMS) algorithm [21] used in detection tasks [23, 6] for suppressing the candidates with low confidences, the role of Gaussian windows in tracking is to weaken the confidence of the candidate boxes far from the center location of the predicted target in the last frame. The explanation for which the Gaussian window can be effectively used is based on the prior knowledge of the continuity of video frames in tracking tasks, that is, the target could not move too far in the two adjacent frames.

3. Methodology

In this section, we first introduce the problem definition of the proposed adversarial attack method for tracking algorithms. Then a one-shot attack framework setting against Siamese network-based trackers is detailed. Lastly, we elaborate on the proposed dual attention attack method.

3.1. Problem Definition

Our attack targets the most popular VOT pipeline—Siamese network-based trackers described above, which formulate the VOT as learning a general similarity map by cross-correlation between the feature representations learned for the target template and the search region (see Fig. 2). In these trackers, the offline learned siamese similarity function and the target template given in the first frame are fixed during the running time. Such a tracking

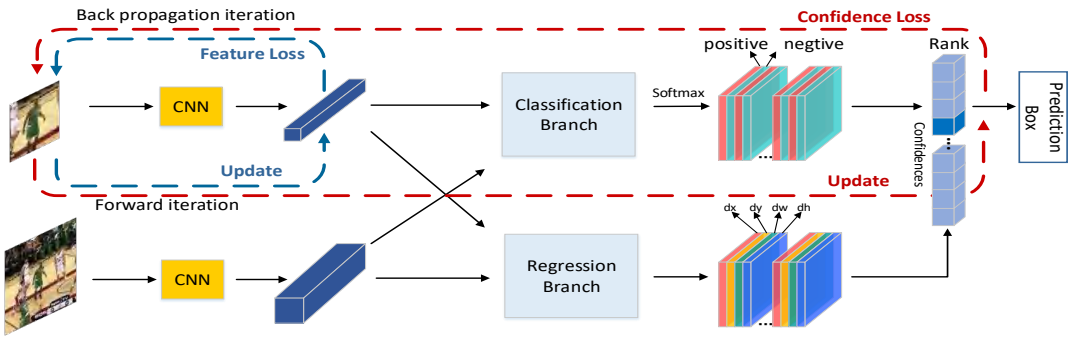

Figure 2. The framework of the one-shot attack against Siamese trackers by dual attention mechanisms.

process without model updates and template updates makes it possible to encounter attacks. Note that other trackers with updating, such as CREST, MDNet and ATOM, are more easily to be attacked because the adversarial information will lead the feature of the model to draft to the wrong space. Then they almost cannot work after attacks due to the wrong update. They are thus not discussed in this paper.

Although there are many existing attack methods for other high-level CV tasks such as detection and classification, it is quite a challenge to attack tracking tasks because tracking tasks are very different from these tasks. Specifically, we analyze the characteristics of Siamese trackers compared with detection and classification.

First, online visual tracking is unable to pre-know the category of the tracked objects because the target position is only given in the first frame of a video for the model training. Therefore it cannot off-line learn a mechanism to perturb the pixel values in advance while it is impossible to generate general class-level adversarial perturbations, which are just commonly used in attack algorithms against classification and detection.

Second, the concept of failure tracking, which is different from misclassification, that maximizing the probability of the category with the second-highest confidence to exceed the probability of the correct category for the targeted attack. As was explained above, the Siamese trackers output confidence maps that metric the similarity of the target and the candidates in the search region. The candidate with the highest confidence in the rankings is selected for the prediction of the object. Only simple maximizing the box with the second-highest confidence does not lead to failed tracking. For example, all anchors (candidate boxes) in SiamRPN are employed for regression to the location of the target which enables a considerable number of anchors to accurately return to the target location.

Last but not least, different from the NMS algorithm used in the detection, the Gaussian windows are widely applied to refine the box confidences in the tracking task, which induces difficulties to balance the strength and the success of the attack. For example, when only considering the power of the attack, the box farthest from the object is the best perturbation target. However, the confidences of distant boxes are more suppressed by Gaussian windows and selecting these boxes as the target may result in a failed attack.

In response to these challenges, we propose several criteria to generate the adversarial perturbations.

Firstly, it is necessary to generate matching adversarial perturbations for arbitrary targets of tracking because of the unknown category. Therefore, we propose to only add an adversarial perturbation in the initial frame of each video, namely one-shot attack.

Secondly, our adversarial attack must be able to perturb a certain number of boxes, which can increase the success rate of the attack. Specifically, adding the perturbations can reduce the confidence of several high-quality boxes and raise the confidence of many low-quality boxes, resulting in that tracker outputs wrong prediction boxes with large deviations. Thus, we propose to learn attack perturbations by optimizing a batch confidence loss. Besides, we need to consider general attack by designing a feature loss function to ensure the attack power. Therefore, two optimization strategies are introduced by us. One is the batch confidence loss while the other is attacking generally all candidates from the feature space of CNNs.

Lastly, to further improve the attack power, we add both attention mechanisms to these two loss functions. On the batch confidence loss, we distinctively suppress different candidates using confidence attention. On the feature loss, we add attention to the channels of the feature map to distinguish the importance of different channels by feature attention, which is inspired by [12].

Considering these criteria, we propose the one-shot attack based on the dual attention, which is detailed in the next two subsections.

3.2. One-shot Attack with Dual Loss

Given the initial frame and the ground-truth bounding box of the tracking target, we could obtain the target patch z. The goal of our one-shot attack is to generate an adver-

sarial target image $\boldsymbol{z}^*$ ($\boldsymbol{z}^* = \boldsymbol{z}+\Delta\boldsymbol{z}$) with slight pixel value perturbation $\Delta\boldsymbol{z}$ only in the initial frame, which could make the tracking results away from the ground-truth (*i.e.* failure tracking). We define the adversarial example attacking the tracker as follows:

$$\boldsymbol{z}^* = \underset{|z_k - z_k^*| \leq \varepsilon}{\arg\min} \mathcal{L}(\boldsymbol{z}, \boldsymbol{z}^*) \tag{1}$$

where z_k denotes the per pixel of clean image $\boldsymbol{z}$ while z_k^* refers to which of the adversary $\boldsymbol{z}^*$, and ε stands for the maximum perturbation range of the per-pixel value in the image. In our experiments, ε is set to 16, for which the global perturbations with such intensity are considered to be an imperceptible change in the human visual system. The batch confidence loss function $\mathcal{L}_1$ and feature loss function $\mathcal{L}_2$ are detailed below.

Batch Confidence Loss. Our one-shot attack only occurs in the initial frame of each video, so we simulate the tracking process in the initial frame (given the tracking template) to generate the adversarial example. Note that the test has not yet started in this phase for the general tracking task.

Follow the Siamese trackers, we assume that the search region $\boldsymbol{X}$ is around the target and twice the size of it, includes n candidates $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. Let $f(\boldsymbol{z}, \boldsymbol{x}_i)$ denotes the tracking model which takes $\boldsymbol{z} \in R^m$ and $\boldsymbol{x}_i \in R^m$ as inputs and the confidence of each candidate as the output. The output confidences $f(\boldsymbol{z}, \boldsymbol{x}_i)$ of the n candidates have a ranking $\mathcal{R}_{1:n}$. Thus the batch confidence loss function can be defined as follows:

$$\begin{aligned} \mathcal{L}_1 = & \sum_{\mathcal{R}_{1:p}} f(\boldsymbol{z}^*, \boldsymbol{x}_i) - \sum_{\mathcal{R}_{q:r}} f(\boldsymbol{z}^*, \boldsymbol{x}_i), \\ & s.t. \quad |z_k - z_k^*| \leq \varepsilon. \end{aligned} \tag{2}$$

where $\mathcal{R}_{1:p}$ denotes the sort in the first p, $\mathcal{R}_{q:r}$ denotes the sort from q to r in the confidence ranking. The purpose of this approach based on batch confidence is to suppress the candidates with high confidence and stimulate the candidates with moderate confidence.

Feature Loss. Considering the challenges come from the Gaussian window and to balance the strength and success of the attack power, we apply another strategy that is generally attacking all candidates from the feature space of CNNs.

Let $\phi(\cdot)$ represents the feature map of CNNs, then the Euclidean distance of the feature maps of $\boldsymbol{z}$ and $\boldsymbol{z}^*$ are maximized. Thus the loss function is defined as follows:

$$\begin{aligned} \mathcal{L}_2 = & -\sum_{j=1:C} ||\phi_j(\boldsymbol{z}^*) - \phi_j(\boldsymbol{z})||_2, \\ & s.t. \quad |z_k - z_k^*| \leq \varepsilon. \end{aligned} \tag{3}$$

where C denotes the channel of the feature maps.

3.3. Dual Attention Attacks

Furthermore, we add attention mechanisms to both two loss functions to further improve the attack power.

Confidence Attention. By applying the confidence attention mechanism to the loss function, we can distinguish the degree of suppression and stimulation for the candidates with different confidences. The Eq. (2) is rewritten as

$$\begin{aligned} \mathcal{L}_1^* = & \sum_{\mathcal{R}_{1:p}} \{w_i \cdot f(\boldsymbol{z}^*, \boldsymbol{x}_i)\} - \sum_{\mathcal{R}_{q:r}} \{f(\boldsymbol{z}^*, \boldsymbol{x}_i)\} \\ & s.t. \quad |z_k - z_k^*| \leq \varepsilon. \end{aligned} \tag{4}$$

with w_i defined as

$$w_i = \frac{1}{a + b \cdot \tanh(c \cdot (d(\boldsymbol{x}_i) - d(\boldsymbol{x}_1)))}, \tag{5}$$

where $d(\boldsymbol{x}_i)$ denotes the coordinates distance between the any i-th candidate $\boldsymbol{x}_i$ and first $\boldsymbol{x}_1$ in the sorted confidence list. Eq. (5) is inspired by the Shrinkage loss [19], in which a, b, and c are controlling hyper-parameters. Specially, c stands for the shrinkage rate, and both a and b limit the the weight w_i to the range of $(\frac{1}{a+b}, \frac{1}{a})$.

Feature Attention. Because of the limited perturbation conditions, we further consider the channel-wise activation-guided attention of feature maps to distinguish the importance of different channels, which will guarantee the generalization ability of the one-shot attack. Similarly, the Eq. (3) is rewritten as:

$$\begin{aligned} \mathcal{L}_2^* = & -\sum_{j=1:C} ||w_j' \cdot \{(\phi_j(\boldsymbol{z}^*) - \phi_j(\boldsymbol{z})\}||_2, \\ & s.t. \quad |z_k - z_k^*| \leq \varepsilon. \end{aligned} \tag{6}$$

and w_j is defined as

$$w_j = \frac{1}{a' + b' \cdot \tanh(c' \cdot (m(\phi_j(\boldsymbol{z})) - m(\phi_j(\boldsymbol{z}))_{min}))}, \tag{7}$$

where $m(\cdot)$ and $m(\cdot)_{min}$ stand for the mean of each channel $\phi_j(\boldsymbol{z})$ and the minimum mean value, a', b' and c' are controlling hyper-parameters.

Dual Attention Loss. We combine the advantages of $\mathcal{L}_1^*$ with accurate attacks and $\mathcal{L}_2^*$ with general attacks, and eventually obtain the dual attention loss:

$$\mathcal{L} = \alpha\mathcal{L}_1^* + \beta\mathcal{L}_2^*, \tag{8}$$

where the factors α and β will be determined empirically.

The goal of our optimizer is to minimize the total loss $\mathcal{L}$. In the implementation, we use Adam optimizer [13] to minimize the loss by iteratively perturbing the pixels along the gradient directions within the patch area, and the generation process stops when the number of iterations reaches 100 or the first candidate of the ranking $\mathcal{R}_\tau[1]$ behinds p in the initial ranking $\mathcal{R}_0$. The whole attack process is presented in Algorithm 1.

Algorithm 1: One-shot White-box Attack for VOT

Input: The target crop $\boldsymbol{z}$ in the first frame image of a video; The tracker with Siamese network $f(\cdot,\cdot)$
Output: An adversarial example $\boldsymbol{z}^*$.

Initialize the adversary $\boldsymbol{z}^* = \boldsymbol{z}$;
Initialize the iterative variable $\tau = 0$;
Feed clean $\boldsymbol{z}$ and search area $\boldsymbol{X}$ containing n candidates $\boldsymbol{x}_i$ into the network to get confidence map $f(\boldsymbol{z}, \boldsymbol{x}_i)$;
Sort $f(\boldsymbol{z}, \boldsymbol{x}_i)$ and obtain the initial Rank $\mathcal{R}_0[1:n]$;
Save the candidate indexes in original Rank $\mathcal{R}_0[1:n]$;
while *Number of iterations* τ ++ < 100 **do**
    Sort $f(\boldsymbol{z}^*, \boldsymbol{x}_i)$ and get the new Rank $\mathcal{R}_\tau[1:n]$;
    if *the sort of the candidate in* $\mathcal{R}_\tau[1] > p$ *in* $\mathcal{R}_0$ **then**
        break;
    else
        dual attention attack ;
        $\boldsymbol{z}^* := \boldsymbol{z}^*_\tau$;
    end
end

4. Attack Evaluation

In this section, we describe our experimental settings and analyze the attack results of the proposed dual attention attack algorithm against different trackers on 3 challenge tracking datasets, including OTB100 [30], LaSOT [4], and GOT10K [11]. Then we evaluate the effectiveness of the proposed method by ablation studies on various contrast experiments.

4.1. Experimental Setting

Attacked Targets. We show our adversarial attack results on four representative Siamese network based trackers, including SiamFC [2], SiamRPN [17], SiamRPN++ [16], and SiamMask [28]. Besides, our experiments employ SiamRPN++ with two different backbones, including ResNet-50 and MobileNet-v2, which are called SiamPRN++(R) and SiamPRN++(M) respectively below.

Evaluation Metrics. For fair evaluation, the standard evaluation methods are applied to measure our attack effect. In the OTB100 and LaSOT datasets, we applied the one-pass evaluation (OPE) with the precision plot and success plot metrics. The precision plot reflects the center location error between tracking results and ground-truth. The threshold distance is set to 20 pixels. Meanwhile, the success rate measures the overlap ratio between the detected box and ground-truth which could reflect the accuracy of tracking in scales. In the GOT10K dataset, we applied Average of Overlap rates (AO) between tracking results and ground-truths over all frames and Success Rate (SR) with a threshold of 0.50. We view successful attacks and failed trackings as consistent. Specifically, the lower the accuracy of the tracking is, the higher the success rate of the attack is.

Table 1. Comparison of results with original, Random noise, and our attack of different Siamese trackers on the OTB100 dataset in terms of precision and success rate.

Trackers	Precision (%)			Success Rate(%)		
	Org	Noise	Ours	Org	Noise	Ours
SiamFC	76.5	73.4	27.1	57.8	56.0	32.3
SiamRPN	87.6	83.1	27.8	66.6	63.3	20.4
SiamRPN++(R)	91.4	85.0	33.7	69.6	64.9	25.2
SiamRPN++(M)	86.4	80.7	35.3	65.8	58.0	26.1
SiamMask	83.7	83.6	65.0	64.6	62.6	48.1

Implementation Details. Our algorithm is implemented by Pytorch and runs on the NVIDIA Tesla V100 GPU. For each attacked video, we use Adam optimizer [13] to optimize the generated adversarial perturbation, with 100 iterations and a learning rate of 0.01. Based on the different purposes of the attention modules, we use different hyper-parameter settings. Specifically, for the confidence attention module, we set $a = 0.5$, $b = 1.5$ and $c = 0.2$. Meanwhile, for the feature attention module, we set $a' = 2$, $b' = -1$, and $c' = 20$, respectively. To balance the weight parameters α and β, we set $\beta = 1$ while α is a model-sensitive parameter in the range of 0.2 to 0.8 in our experiments. In Eq. (2), the hyper-parameters of p, q, r are set to 45 ($9 \cdot 5$ anchors), 90, and 135, respectively. In addition, all the results presented below are the averaged value of repeated five times experiments under these settings.

4.2. Overall Attack Results

Results on OTB100. Table 1 compares the overall results of theses trackers in OTB100 dataset. We compare random noises with our adversarial examples on the target patch in the initial frame and observe that they impact very little on tracking results, but our adversarial attack can cause almost devastating results to the tracking methods. Specifically, the precision of adding random noises to SiamFC, SiamRPN, SiamRPN++(R), SiamRPN++(M) are reduced by 3.1%, 4.5%, 6.4%, and 5.7% respectively. While the precision of adding adversarial perturbations to corresponding trackers are greatly reduced by 49.4%, 59.8%, 57.7%, and 51.1%, respectively.

Fig. 3 shows the success and precision plots on OTB100 dataset with the comparison of the results by original trackers and the results after our corresponding attacks. We can see that the precision results and success rates of the five trackers are significantly reduced after being attacked. In precision plots, we observe that the proposed attack method has the best and second attack effects on SiamRPN and SiamRPN++(R), which reduces the precision by 59.8% and 57.7%, respectively. Similarly, our attack method reduces the success rate on SiamRPN and SiamRPN++(R) by 46.2% and 44.4% respectively.

Results on LaSOT. We compare our attack against these

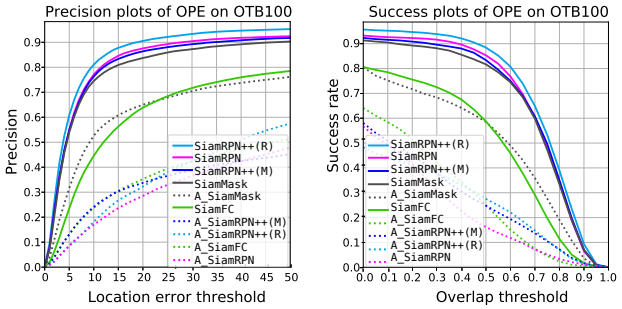

Figure 3. Evaluation results of trackers with or without adversarial attacks on OTB100 dataset.

Table 2. Comparison of results with original, Random noise, and our attack of different Siamese trackers on the LaSOT dataset in terms of precision and success rate.

Trackers	Precision (%)			Success Rate(%)		
	Org	Noise	Ours	Org	Noise	Ours
SiamFC	34.4	33.7	12.0	35.2	34.7	16.7
SiamRPN	42.4	42.2	10.8	43.3	43.1	14.7
SiamRPN++(R)	50.2	49.3	12.2	49.6	48.5	14.9
SiamRPN++(M)	45.5	45.5	11.4	45.2	44.9	14.7
SiamMask	46.3	46.0	34.3	46.5	46.3	37.1

Table 3. Comparison of results with original, Random noise, and our attack of different Siamese trackers on the GOT10k dataset in terms of AO and $SR_{0.50}$.

Trackers	AO (%)			$SR_{0.50}$ (%)		
	Org	Noise	Ours	Org	Noise	Ours
SiamFC	53.8	50.2	34.6	57.8	54.3	28.4
SiamRPN	60.8	56.1	31.2	71.4	65.2	26.5
SiamRPN++(R)	65.1	65.0	31.2	76.7	75.7	26.5
SiamRPN++(M)	64.1	61.0	39.4	75.0	70.2	34.7
SiamMask	64.4	64.1	55.6	76.5	75.9	64.1

trackers on the LaSOT dataset [4]. Table 2 shows the overall results of these trackers after attacks perform poorly. Through the results, we can see that precision of these five trackers has a significant decline, accounting for 34.9%, 25.5%, 24.3%, 25.1%, and 74.1% of the original results in SiamFC, SiamRPN, SiamRPN++(R), SiamRPN++(M), and SiamMask respectively.

Results on GOT10K. We also implement our attack against these five trackers on the large tracking dataset GOT10K [14]. Table 3 shows that there are significant declines in overall results on these trackers after attacks. Through the results, we can see that AO these five trackers decreased by 64.3%, 51.3%, 61.5%, 47.9%, and 86.3% respectively.

Analysis. From the attack results on these trackers in various datasets, we found an interesting phenomenon that the simplest SiamFC shows good robustness on both OTB100 and LaSOT, which we believe is due to the under-fitting of the algorithm. More specifically, to some extent, SiamFC can be considered as a SiamRPN with only one anchor.

Table 4. Ablation comparison studies of dual attention attacks.

SiamRPN++ (ResNet-50)	Precision (%)	Success Rate (%)
Original	91.4	69.6
Random Noise	85.0	64.9
Attack by $\mathcal{L}_1$	38.8	29.1
Attack by $\mathcal{L}_1^*$	37.1	27.6
Attack by $\mathcal{L}_2$	38.7	27.7
Attack by $\mathcal{L}_2^*$	34.3	25.6
Attack by $\mathcal{L}_1 + \mathcal{L}_2$	37.5	26.9
Attack by $\mathcal{L}_1^* + \mathcal{L}_2^*$	**33.7**	**25.2**

Generally, too few anchors make SiamFC unable to accurately estimate the target. Meanwhile, it reduces the risk of being attacked by adversarial samples. Moreover, we can see that our attack method has the best attack effect on SiamRPN, which reduces the precision on OTB100 by 59.8% and the success rate by 46.2%, respectively. Thus can be attributed to the excessive head parameters that make SiamRPN difficult to get fully trained. To a certain extent, this problem has been solved in siamRPN++ with the help of multi-stage learning and more efficient cross-correlation. As we can see, siamRPN++ has better robustness and more difficult to be attacked. Besides, our attacking method has the lowest degree of attack on SiamMask compared with other trackers. For example, the attack on SiamMask reduces the precision and success rate by 18.7% and 16.5% on OTB100 respectively, which can be attributed to the multi-task learning of SiamMask. Compared with SiamRPN and SiamRPN++, SiamMask adds a semantic segmentation branch and focuses on the tracked object with pixel-level, which makes the learned features more robust.

4.3. Ablation Study of Dual Attention Attack

We implement a series of experiments to analyze and evaluate the contribution of each component of our dual attention attacks. We choose the current state-of-the-art tracker SiamRPN++(R) as the representative and the tracking results on OTB100 are shown in Table 4.

Intuitively, we observe that random noises impact very little on tracking results, but our adversarial attacks cause a significant drop in tracking accuracy. Moreover, separately using the loss $\mathcal{L}_1$ and loss $\mathcal{L}_2$ in our experiments greatly reduce the accuracy of tracking and their damage to tracking is similar to each other in terms of the data. It thanks to our selection strategy for candidates of $\mathcal{L}_1$ and the global feature perturbation mechanism of $\mathcal{L}_2$. Second, we test the effectiveness of the distance-oriented confidence attention mechanism in $\mathcal{L}_1$ component, namely $\mathcal{L}_1^*$. Specifically, the $\mathcal{L}_1^*$ method further reduces the tracking accuracy by 1.7% and 1.5% for both precision and success rate metrics based on $\mathcal{L}_1$. At the same time, we validate the contribution of the activation-oriented feature attention mechanism in $\mathcal{L}_2$ component, namely $\mathcal{L}_2^*$, and reduce the tracking performance by 4.4% and 2.1% for success and precision, respectively.

Figure 4. Qualitative evaluation of one-shot adversarial attack in various trackers on video examples *Human7* and *Human2* from the OTB100 dataset. For each of the two subfigures, the first column represents the adversarial examples generated in the initial frame, except the clean example in the first row. The green, blue, and red rectangles represent the bounding boxes of ground-truth, tracking results before and after being attacked.

Moreover, through the experimental analysis, we can see that the feature attention mechanism brings more gain than the confidence attention mechanism. The potential reason is that the attention mechanism of $\mathcal{L}_1^*$ has narrowed the candidates to a more appropriate range, and all of these selected boxes will contribute to the attack. In addition, the feature attention mechanism can force the algorithm to mine channels that contribute more to the attack in the huge feature space, which effectively reduces the concerned scope belonging to $\mathcal{L}_2$ attack. Besides, the attacking strategy using two basic components $\mathcal{L}_1^*$ and $\mathcal{L}_2^*$ simultaneously achieves gain on each basis. Finally, the dual attention attack method obtains the best attack result by simultaneously employing two attention mechanisms.

4.4. Qualitative Evaluation

Fig. 4 shows examples of adversarial attacks against various trackers. We can see that the initial frame perturbation of the five trackers is so subtle that it is difficult to be observed by the human eye. Generally, adding adversarial attacks results in a large deviation of tracking results. Among them, the attacks on SiamFC and SiamRPN are stronger when the target scale changes greatly. In contrast, the impact on the results of SiamRPN++ is not obvious, which is partly attributed to the robust feature extraction using deeper models.

5. Conclusion

In this work, we highlight the adversarial perturbations against VOT to circumvent potential risks of the surveillance system. We focus on the adversarial attacks for free-model single object tracking and our attack target is a series of excellent trackers based on Siamese networks. We present a one-shot attack method that only perturbs slightly the pixel values of the initial frame image of a video, resulting in tracking failure in subsequent frames. Experimental results prove that our approaches can successfully attack the advanced Siamese network-based trackers. We hope that more researchers can pay attention to the adversarial attack and defense of the tracking algorithms in the future.

Acknowledgement This work is supported in part by the National Natural Science Foundation of China under Grant 61972188, 61771273, the National Key Research and Development Program of China under Grant 2018YFB1800204, the R&D Program of Shenzhen under Grant JCYJ201805-08152204044, the Science and Technology Planning Project of Shenzhen (No. JCYJ20180503182133411), and the research fund of PCL Future Regional Network Facilities for Large-scale Experiments and Applications (PCL2018KP001).

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision Workshops (ECCV Workshops)*, pages 850–865, 2016.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[4] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5374–5383, 2019.

[5] Heng Fan and Haibin Ling. Sanet: Structure-aware network for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 42–49, 2017.

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[8] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pages 1763–1771, 2017.

[9] Qing Guo, Xiaofei Xie, Lei Ma, Zhongguo Li, Wanli Xue, and Wei Feng. Spark: Spatial-aware online incremental attack against visual tracking. In *arxiv preprint arXiv:1910.0868*, 2019.

[10] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4834–4843, 2018.

[11] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *arXiv preprint arXiv:1810.11981*, 2018.

[12] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7066–7074, 2019.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Cehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[16] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4282–4291, 2019.

[17] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8971–8980, 2018.

[18] Peixia Li, Dong Wang, Lijun Wang, and Huchuan Lu. Deep visual tracking: Review and experimental comparison. *Pattern Recognition (PR)*, 76:323–338, 2018.

[19] Xiankai Lu, Chao Ma, Bingbing Ni, Xiaokang Yang, Ian Reid, and Ming-Hsuan Yang. Deep regression tracking with shrinkage loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 353–369, 2018.

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

[21] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.

[22] Nicolas Papernot, Patrick Mcdaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security & Privacy*, 2016.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[25] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1420–1429, 2016.

[26] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, 2019.

[27] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4854–4863, 2018.

[28] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338, 2019.

[29] Rey Reza Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[30] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(9):1834–1848, 2015.

[31] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1369–1378, 2017.

[32] Yunhua Zhang, Lijun Wang, Jinqing Qi, Dong Wang, Mengyang Feng, and Huchuan Lu. Structured siamese network for real-time visual tracking. In *Proceedings of the European Conference on Computer Cision (ECCV)*, pages 351–366, 2018.