A Simple and Strong Baseline for Universal Targeted Attacks on Siamese Visual Tracking

Zhenbang Li, Yaya Shi, Jin Gao, Shaoru Wang, Bing Li, Pengpeng Liang, Weiming Hu, Senior Member, IEEE

Abstract—Siamese trackers are shown to be vulnerable to adversarial attacks recently. However, the existing attack methods craft the perturbations for each video independently, which comes at a non-negligible computational cost. In this paper, we show the existence of universal perturbations that can enable the targeted attack, e.g., forcing a tracker to follow the ground-truth trajectory with specified offsets, to be video-agnostic and free from inference in a network. Specifically, we attack a tracker by adding a universal translucent perturbation to the template image and adding a fake target, i.e., a small universal adversarial patch, into the search images adhering to the predefined trajectory, so that the tracker outputs the location and size of the fake target instead of the real target. Our approach allows perturbing a novel video to come at no additional cost except the mere addition operations — and not require gradient optimization or network inference. Experimental results on several datasets demonstrate that our approach can effectively fool the Siamese trackers in a targeted attack manner. We show that the proposed perturbations are not only universal across videos, but also generalize well across different trackers. Such perturbations are therefore doubly universal, both with respect to the data and the network architectures. We will make our code publicly available.

Index Terms—Visual tracking, adversarial attacks, Siamese networks.

I. INTRODUCTION

GIVEN an arbitrary detected or annotated object of interest in the initial video frame, visual object tracking is aimed at recognizing and localizing other instances of the same object in subsequent frames. This paradigm of tracking visual objects from a single initial exemplar in the online-tracking phase has been broadly cast as a Siamese network-based one-shot problem recently [1], [2], [3], [4], which is termed as Siamese visual tracking and recognised as highly effective and efficient for visual tracking.

Recently, the robustness of Siamese trackers has attracted much attention in the sense of testing their vulnerability to adversarial attacks, and the focus has been directed towards more efficient and low-cost attacks [5], [6], [7], [8]. However, these attack methods still need to craft the perturbations for each video independently based on either iterative optimization or adversarial network inference, for which the adequacy of additional computational resources may be not ensured by the computationally intensive tracking systems.

Universal adversarial perturbations (UAPs) proposed in [9] can fool most images from a data distribution in an image-agnostic manner. Being universal, UAPs can be conveniently exploited to perturb videos on-the-fly without extra intensive computations except the mere addition operations.

Fig. 1. Being universal, our generated perturbations can be conveniently exploited to perturb videos on-the-fly without extra intensive computations except the mere addition operations.
A. Siamese Visual Tracking

Siamese visual tracking is a fundamental research direction in template matching-based tracking besides the correlation filter-based methods. Both of them are aimed to “causally” estimate the positions of a template cropped from the initial video frame in the subsequent frames. Online learning has played an important role in correlation filter-based tracking, and recent works (e.g., [17], [18], [19]) have largely advanced the research in this area by exploring various online update strategies. However, most Siamese trackers do not exploit the online learning paradigm and innovatively formulate visual tracking as learning cross-correlation similarities between a target template and the candidates in the search region in an end-to-end convolution fashion. Tracking is then performed by locating the object in the search image region based on the highest visual similarity. This paradigm is formulated as a local one-shot detection task.

Recently, some Siamese trackers [10], [20], [21], [22], [23], [24] have demonstrated a significant performance improvement in visual tracking. In particular, SiamRPN [23] consists of one Siamese subnetwork for feature extraction and another region proposal subnetwork including the classification and regression branches separately. Based on the success of the decomposition of classification and state estimation in SiamRPN, SiamRPN++ [24] further breaks the restriction of strict translation invariance through a simple yet effective spatial aware sampling strategy and successfully trains a ResNet-driven Siamese tracker with significant performance gains. Apart from this kind of anchor-based methods, an anchor-free tracker SiamFC++ [10] is further designed by considering non-ambiguous scoring, prior target scale/ratio distribution, knowledge-free and estimation quality assessment guidelines. In this anchor-free paradigm, many recent works (e.g., [20], [21], [22]) are dedicated to more robust and efficient visual tracking based on various anchor-free settings. Ocean [20] directly predicts the position and scale of target objects in an anchor-free fashion; since each sample position in ground-truth boxes is well trained, the tracker is capable of rectifying inexact predictions of target objects during inference. FCOT [21] introduces an online regression model generator (RMG) based on the carefully designed anchor-free box regression branch, which enables the tracker to be more effective in handling target deformation during tracking procedure. SiamCAR [22] is both anchor and proposal free, and takes one unique response map to predict object location and bounding box directly; this setting significantly reduces the number of hyper-parameters, which keeps the tracker from complicated parameter tuning and makes the tracker significantly simpler, especially in training. In our experiments, we are focused on the anchor-free SiamFC++ tracker, whereas the transferability of our generated adversarial attacks to some other Siamese trackers is also studied. A comprehensive survey of the related trackers is beyond the scope of this paper; please refer to [25] for a thorough survey.

B. Adversarial Attacks

Adversarial attack [26] to image classification is first investigated in [27] with the aim of identifying the vulnerability of modern deep networks to imperceptible perturbations. Recent studies also emerge to investigate the adversarial attacks to other diverse types of tasks such as natural language processing [28], [29], [30], [31] and object detection [32]. Some recent research topics which are mostly related to adversarial attack also include digital watermarking [33], 3D face presentation attacks to the face recognition networks [34] and adversarial defense for image classifiers [26]. Xiong et al. [33] propose to generate the digital watermarking by slightly modifying the pixel values of video frames to protect video content from unauthorized access. Compared with [33], our purpose of modifying pixel values of video frames is attacking the tracker, instead of detecting the illegal distribution of a digital movie. Jia et al. [34] propose to generate 3D face artifacts to attack the face recognition networks. Compared with [34], our method directly modifies the input of the network instead of changing the input of cameras using 3D face artifacts. Wang et al. [26] propose the white-box attack/defense methods for image classifiers. Compared with [26], we focus on attacking the object tracking networks instead of the image classification networks. In the following, we mainly introduce some scenar-
ios of possible adversarial attacks which can be categorized along different dimensions.

White box attacks v.s. Black box attacks. In the white box attack setting [53], the adversary has full knowledge of the model including model type, model architecture and values of all parameters and trainable weights. In the black box setting [50], [57], [58], [59], the adversary has limited or no knowledge about the model under attack [40]. In this paper, we focus on the white box attacks to Siamese trackers. However, we surprisingly find that our perturbations based on the SiamFC++ tracker [10] can perform black box attacks to some other Siamese trackers such as SiamRPN [23], SiamRPN++ [24] and Ocean [20].

Non-universal attacks v.s. Universal attacks. In the setting of non-universal attacks [41], [42], [43], the adversary has to generate one perturbation for every new datapoint, which is time consuming. In the setting of universal attacks [44], [45], [46], [47], [48], only one perturbation is generated and then used for all datapoints in the dataset. In this paper, we focus on generating universal perturbations for Siamese trackers to perform efficient attacks.

Imperceptible Perturbations v.s. Adversarial Patch. The imperceptible perturbations most commonly modify each pixel by a small amount and can be found using a number of optimization strategies such as Limited-memory BFGS [27] and PGD [49]. Different from the imperceptible perturbations, the adversarial patch is extremely salient to a neural network. The adversarial patch can be placed anywhere into the input image to cause the network to misbehave, and thus is commonly used for universal attacks [50]. Note that our adversarial patch works in the image domain instead of the network domain. In the network-domain case, the noise is allowed to take any value and is not restricted to the dynamic range of image value as in the image-domain case [51].

Untargeted Attacks v.s. Targeted Attacks. In the case of untargeted attacks, the adversary’s goal is to cause the network to predict any incorrect label and whatever the incorrect label is does not matter, e.g., pushing the object location estimation just outside the true search region in visual tracking. Targeted attacks, however, aim to change the network’s prediction to some specific target label. In visual tracking, the targeted attacks aim to intentionally drive trackers to output specified object locations following a predefined trajectory.

C. Adversarial Attacks in Visual Tracking

Recently, there are several explorations of the adversarial attacks to the visual tracking task. For example, PAT [52] generates physical adversarial textures via white-box attacks to steer the tracker to lock on the texture when a tracked object moves in front of it. However, PAT validates its method by attacking a light deep regression tracker GOTURN [53], which has low tracking accuracy on modern benchmarks. In this paper, we aim to attack state-of-the-art Siamese trackers. RTAA [54] takes temporal motion into consideration when generating lightweight perturbations over the estimated tracking results frame-by-frame. CSA [55] is a Siamese tracking attack method called cooling-shrinking attack. CSA can suppress the peak region of the heat map reflecting the target location, which is used to attack the tracker’s targeting ability. However, both of these two methods only perform the untargeted attacks to trackers, which is less challenging than the targeted attacks in this paper, as we aim to create arbitrary, complex trajectories at test time.

Targeted attacks to follow an erroneous path which looks realistic are crucial to deceive the real-world tracking systems without raising possible suspicion. SPARK [7] computes incremental perturbations by using information from the past frames to perform targeted attacks to Siamese trackers. However, SPARK needs to generate distinct adversarial examples for every search image through heavy iterative schemes, which is time-consuming to attack online tracking in real time. FAN [6] proposes a fast attack network for attacking SiamFC tracker. To perform untargeted attacks, FAN proposes a drift loss that shifts the tracker’s prediction of the target’s position. The tracking error accumulates over time until the tracker loses the target completely. To perform targeted attacks, FAN proposes embedding feature loss for increasing the similarity between the features of the adversarial sample and the regions specified by a particular trajectory. However, similar to CSA, FAN also requires running a generative network for each frame to obtain adversarial information, making it difficult to meet the demand for real-time tracking. The recent real-time attacker TTP [5] exclusively uses the template image to generate a single temporally-transferable perturbation in a one-shot manner, and then adds it to every search image of the video. However, this method still needs to generate the perturbation for each individual video, and its targeted attack setting requires diverse perturbations from several runs of network inference. It is ill-suited to attack a real-world online-tracking system when we cannot get access to the limited computational resources. In this paper, however, we propose video-agnostic perturbations which allow perturbing a novel video to come at no additional cost except the mere addition operations.

III. Method

In this section, we introduce our video-agnostic targeted attack framework for Siamese trackers. We aim to attack the tracker by adding a perturbation to the template image and adding a fake target, i.e., an adversarial patch, into the search images adhering to the predefined trajectory, so that the tracker outputs the location and size of the fake target instead of the real target. Below, we formalize our targeted attacks to SiamFC++ [10], and then introduce our perturbation strategy.

A. Problem Definition

Let $V = \{I_i\}_{i=1}^T$ denote the frames of a video sequence of length $T$. $B^{gt} = \{b^{gt}_i\}_{i=1}^T$ are used to represent the target’s ground-truth positions in those frames. Visual object tracking aims to predict the positions $B^{pred} = \{b^{pred}_i\}_{i=1}^T$ of the target in the subsequent frames given its initial state. In SiamFC++, the tracker first transforms the paired reference frame $I_1$ and annotation $b^{gt}_1$ to get a template image $z_1$, and transforms the search frame $I_i$ to get the search image $x_i$, centered at the position estimated in the previous frame. At each timestep, the template image $z_1$ and the search image $x_i$ are
first passed individually through a shared backbone network, i.e., Siamese network, and then fused using a channel-wise correlation operation:

\[
\text{Feat}_j(z_i, x_i) = \psi_j(\phi(z_i)) \ast \psi_j(\phi(x_i)), j \in \{\text{cls}, \text{reg}\}
\]  

(1)

where \( \ast \) denotes the channel-wise correlation operation, \( \phi(\cdot) \) denotes the feature extractor of the Siamese network, \( \psi_j(\cdot) \) denotes the layers specific to the classification or regression task, and \( j \) denotes the specific task type (cls denotes the classification task and reg denotes the regression task). \( \psi_{\text{cls}} \) and \( \psi_{\text{reg}} \) are both designed as two convolution layers for adapting generic features to the feature space specific to the classification/regression task. The fused features then act as input to a head network, which predicts a classification map \( C \), a bounding box regression map \( R \), and a quality assessment map \( Q \) in an anchor-free manner. In short, \( C \) encodes the probability of each spatial position to contain the target, \( R \) regresses the bounding box of the target, and \( Q \) predicts the target state estimation quality. The final bounding box is then generated according to \( C \), \( R \), and \( Q \).

A straight forward way to achieve targeted attacks to Siamese trackers is directly using popular attack methods such as FGSM \([56]\) and BIM \([57]\):

\[
\min_{\delta_x, \delta_z} L(y^{\text{fake}}, f(x_i + \delta_x, z_i + \delta_z))
\]

\[
\text{s.t. } ||\delta_x||_p \leq \epsilon_1, ||\delta_z||_p \leq \epsilon_2,
\]  

(2)

where \( L \) is the loss function of the SiamFC++ tracker. \((x_i, z_i)\) is the search-template pair of the video being tracked at frame \( i \) and \( z_i \equiv z_1 \). \( y^{\text{fake}} = \{C^*, R^*, Q^*\} \) is derived from the fake target for the loss function and generated according to the fake ground-truth box \( b_{x\text{fake}}^k \) that we want the tracker to output at frame \( i \) (see Sec. III-B). Additionally, the perturbation \( \delta \) should be sufficiently small, which is commonly modeled through an upper bound \( \epsilon \) on the \( l_p \)-norm, commonly denoted as \( || \cdot ||_p \).

A popular choice is to set \( p = \infty \). However, these methods require perturbing the input image pair \((x, z)\) frame by frame, which comes at a non-negligible cost, considering that visual object tracking is a real-time task. As a consequence, universal adversarial perturbations (UAPs) \([9], [58]\) are more practical for attacking Siamese trackers:

\[
\min_{p_k, \delta_{x^k}, \delta_{z^k}} \mathbb{E}_{(x, z, y^{\text{fake}})} L(y^{\text{fake}}, f(A_{\text{paste}}(x, p_k, b_{x\text{fake}}^k), z))
\]

\[
\text{s.t. } ||p_k||_p \leq \epsilon_1, ||\delta_{x^k}||_p \leq \epsilon_2, ||\delta_{z^k}||_p \leq \epsilon_2,
\]  

(3)

where \( A_{\text{paste}} \) is a patch application operator which pastes the adversarial patch \( p_k \) into the template image \( b_{x\text{fake}}^k \) and \( b_{x\text{fake}}^k \) denotes the fake target in the search image. \( y^{\text{fake}} \) is generated according to \( b_{x\text{fake}}^k \) and both of them are variables during training. More specifically, \( A_{\text{paste}} \) means that, in the region where the perturbation is pasted, the pixel values of the original image are replaced with the pixel values of the perturbation. Note that this vanilla adversarial patch method uses the clean template image during training. However, CNN attacks are usually expected to be imperceptible whereas the above method has to paste an obviously noticeable fake target patch into tracking frames, which raises the risk of being suspected.

To overcome the aforementioned shortcomings, we propose to train a translucent perturbation \( \delta \) for the template image \( z \), and a translucent patch \( p \) for the search image \( x \). After adding \( \delta \) to \( z \) and adding the fake target patch \( p \) into \( x \), the tracker outputs the location and size of the adversarial patch instead of the real target:

\[
\min_{p_k, \delta_{x^k}, \delta_{z^k}} \mathbb{E}_{(x, z, y^{\text{fake}})} L(y^{\text{fake}}, f(A_{\text{paste}}(x, p_k, b_{x\text{fake}}^k), z))
\]

\[
\text{s.t. } ||p_k||_p \leq \epsilon_1, ||\delta_{x^k}||_p \leq \epsilon_2, ||\delta_{z^k}||_p \leq \epsilon_2,
\]  

(4)

where \( A_{\text{paste}} \) is a patch application operator which adds a small patch into the search image according to \( b_{x\text{fake}}^k \). This also departs from Eq. (3) in that the universal adversarial perturbation \( \delta_n \) is added to the whole search image \( x \) and limited to untargeted attacks. More specifically, the operator \( A_{\text{add}} \) refers to the injection of our translucent adversarial patch into the search image, i.e., the values of our translucent adversarial patch are added to the pixel values of the original image in the \( b_{x\text{fake}}^k \) area.

Our attacks to Siamese visual tracking with both the template and search images as input, enable us to exploit the advantages of both the perturbation generation method \([56]\) and adversarial patch generation method \([50]\) simultaneously. However, applying both of the above methods to trackers is...
non-trivial. In this paper, we demonstrate that the perturbation and the adversarial patch are both indispensable for attacking trackers and can be jointly optimized in an end-to-end training manner. First, the small patch injected into the search image works as a fake target. Since it is crucial to design a strategy for the tracker to follow a predefined trajectory, the TTP method achieves this goal by pre-computing perturbations corresponding to diverse directions and using them to force the tracker to follow the predefined trajectory. However, this method is sub-optimal because the attacked tracker only predicts the approximation of the specified trajectory. As shown in our experiments, the targeted attack performance of TTP is not as good as ours. Second, the adversarial patch is usually added to the background region of the search image and does not change the pixel values in the foreground region where the real target exists. Therefore if we do not add additional perturbation to the template image, the response values in the region where the real target exists on the heatmaps are always high. Thus it is necessary to perturb the template image to cool down hot regions where the real target exists and increase the responses at the position of the fake target.

B. Generating Video-Agnostic Perturbations

In this subsection, we show how to train our video-agnostic perturbations (δ, p) for Siamese trackers. At beginning, each element in δ and p is initialized to 0. During the \( k \)-th iteration of training, a video \( V = \{I_t\}_{t=1}^T \) is randomly selected from the training dataset \( \mathcal{V} \). Assuming the template perturbation at the \( k \)-th iteration is \( \delta_k \in \mathbb{R}^{127 \times 127 \times 3} \), and the adversarial patch is \( p_k \).

First we randomly pick paired frames \( I_s, I_t \) from \( V \). The clean template image \( z \in \mathbb{R}^{127 \times 127 \times 3} \) is generated according to \( I_t \) and \( b_t^{gt} \), and the perturbed template image is:

\[
\tilde{z} = z + \delta_k. \tag{6}
\]

Similarly, the clean search image \( x \in \mathbb{R}^{303 \times 303 \times 3} \) is generated according to \( I_s \) and \( b_s^{gt} \). As mentioned before, the patch is regarded as a fake target and added into the search images. We force the center position of the fake target to near the center position of the real target within a shift range of 64 pixels, where the shift is defined as the maximum range of translation generated from a uniform distribution. The perturbed search image is generated as follows:

\[
\tilde{x} = A_{add}(x, p_k, b_k^{fake}), \tag{7}
\]

where \( b_k^{fake} = \{x_0, y_0, x_1, y_1\} \) denotes the coordinates of the upper-left and lower-right corners of the fake target in the search image, and \( A_{add} \) adds the patch into \( x \) at location \((x_0 + x_1, y_0 + y_1)\). Subsequently, the SiamFC++ tracker takes \( \tilde{x} \) and \( \tilde{z} \) as input and predicts \( C, R, Q \) in an anchor-free manner.

Generating Fake Labels. The fake labels are composed of three parts: fake classification label \( C^* \), fake regression label \( R^* \) and fake quality estimation label \( Q^* \).

For classification, a location \((x, y)\) in the feature map \( \psi_{cls} \) is considered as a positive sample (i.e., \( C_{x,y}^* = 1 \)) if its corresponding location \((\frac{x}{2}, x, \frac{y}{2} + y)\) in the input image falls into the fake bounding box, and considered as a negative sample (i.e., \( C_{x,y} = 0 \)) vice versa. \( s = 8 \) denotes the total step length of the feature extraction network.

In the regression branch, the last convolution layer predicts the distance of a input image position \((\frac{x}{2} + x, \frac{y}{2} + y)\) to the four edges of the fake bounding box, denoted as the four-dimensional vector \( t^* = (l^*, t^*, r^*, b^*) \). Thus, the fake regression label \( R^* \) at location \((x, y)\) can be expressed as:

\[
\begin{align*}
    l^* &= (\frac{x}{2} + x) - x_0, & t^* &= (\frac{x}{2} + y) - y_0, \\
    r^* &= x_1 - (\frac{x}{2} + y), & b^* &= y_1 - (\frac{y}{2} + y)
\end{align*} \tag{8}
\]

SiamFC++ assumes that feature locations near the center of the target will have much more importance than other locations. Following the design of SiamFC++, we use a \( 1 \times 1 \) convolution layer for quality assessment, i.e., to learn the intersection over union (IoU) score of the predicted box \( b_k^{pred} \) and \( b_k^{fake} \). Thus, the fake quality estimation label \( Q^* \) at location \((x, y)\) can be expressed as:

\[
\text{IoU}^* = \frac{\text{Intersection}(b_k^{pred}, b_k^{fake})}{\text{Union}(b_k^{pred}, b_k^{fake})} \tag{9}
\]

Training Objective. The loss function is calculated as follows:

\[
L = \frac{\alpha}{N_{pos}} \sum_{x,y} L_{cls}(C_{x,y}, C_{x,y}^*) + \frac{\beta}{N_{pos}} \sum_{x,y} I_{c_{x,y}>0} L_{quality}(Q_{x,y}, Q_{x,y}^*) + \frac{\gamma}{N_{pos}} \sum_{x,y} I_{c_{x,y}>0} L_{reg}(R_{x,y}, R_{x,y}^*) + \eta_1 \cdot ||\delta_k||_2^2 + \eta_2 \cdot ||p_k||_2^2, \tag{10}
\]

where \( C_{x,y}, R_{x,y}, Q_{x,y} \) represent the values of \( C, R, Q \) at location \((x, y)\) respectively, and \( C^*, R^*, Q^* \) are the fake labels generated according to the position and size of the fake target. \( I \) is the indicator function that takes 1 if the condition in subscribe holds and takes 0 if not. \( N_{pos} \) denotes the number
Algorithm 1 Training Process

Input: Training dataset \( V \), Siamese tracker \( f \), and max iteration number \( N \).

Output: \( \delta, p \).

1. Let \( k = 0 \).
2. while \( k < N \) do
3.     Randomly pick a video \( V \in V \). The corresponding ground-truth is \( B^{gt} = \{b^{gt}_i\}_I \).
4.     Randomly pick paired frames \( I_t, I_s \) from \( V \).
5.     Generate template image \( z \) according to \( I_t \) and \( b^{gt}_t \).
6.     \( \hat{z} = z + \delta \).
7.     Generate search image \( x \) according to \( I_s \) and \( b^{gt}_s \).
8.     Calculate the fake target position \( \{x_0, y_0, x_1, y_1\} \) with respect to the search image.
9.     \( \hat{x} = A_{add}(x, p, \{x_0, y_0, x_1, y_1\}) \).
10. \( C, R, Q = f(\hat{x}, \hat{z}) \).
11. Generate fake labels \( C^*, R^*, Q^* \) using \( \{x_0, y_0, x_1, y_1\} \).
12. Calculate loss \( L(C, R, Q, C^*, R^*, Q^*) \) using Eq. (10).
13. \( \delta_{k+1} = \delta_k - \epsilon_1 \cdot \text{sign}(\nabla_{\delta_k} L) \).
14. \( p_{k+1} = p_k - \epsilon_2 \cdot \text{sign}(\nabla_{p_k} L) \).
15. \( k = k + 1 \).
16. end while
17. return \( \delta_N, p_N \).

Algorithm 2 Attack Process

Input: The trained perturbations \( \delta \) and \( p \), Siamese tracker \( f \), video \( V = \{I_t\}_I \). \( b^{gt}_t \) is the position of the real target in the first frame. \( B^{fake} = \{b^{fake}_i\}_I \) is the trajectory we hope the tracker to output.

Output: \( B^{pred} = \{b^{pred}_i\}_I \).

1. Generate the clean template image \( I_t \) according to \( I_t \) and \( b^{gt}_t \).
2. Generate the perturbed template image \( \hat{z}_1 = z_1 + \delta \).
3. Let \( i = 2 \).
4. while \( i \leq T \) do
5.     Generate clean search image \( x_t \) according to \( I_t \) and \( b^{pred}_t \).
6.     \( b^{fake} = \{x_0, y_0, x_1, y_1\} \).
7.     Generate the perturbed search image \( \hat{x}_i = A_{add}(x_t, p, \{x_0, y_0, x_1, y_1\}) \).
8. \( C, R, Q = f(\hat{x}_i, \hat{z}_i) \).
9. \( \hat{x}_i = A_{add}(x_t, p, \{x_0, y_0, x_1, y_1\}) \).
10. \( i = i + 1 \).
11. end while
12. return \( B^{pred} \).

of positive samples in the training phase, \( L_{cls} \) denotes the focal loss [60] for classification result, \( L_{quality} \) denotes the binary cross entropy (BCE) loss for quality assessment, and \( L_{reg} \) denotes the IoU loss [61] for bounding box regression.

Optimization. Before introducing our perturbation update process in offline training, we first revisit the popular adversarial example generation methods (e.g., [56, 57]). One of the simplest methods to generate adversarial image \( I_{adv} \) is FGSM [56] and works by linearizing the loss function around the network weights and obtaining an optimal max-norm constrained perturbation for generating the adversarial image:

\[
I_{adv} = I + \epsilon \cdot \text{sign}(\nabla_I J(I, y_{true})),
\]

where \( I \) is the input clean image, and the values of its pixels are integer numbers in the range \([0, 255]\). \( y_{true} \) is the true label for the image \( I \). \( J(I, y_{true}) \) is the cost function for training the neural network, and \( \epsilon \) is a hyper-parameter to be chosen. A straightforward way to extend the above method is applying it multiple times with small step size. This leads to Basic Iterative Method (BIM) introduced in [57]:

\[
I_{adv}^0 = I, \quad I_{adv}^{k+1} = C_{clip}(I_{adv}^{k} + \epsilon \cdot \text{sign}(\nabla_I J(I_{adv}^{k}, y_{true}))),
\]

where the pixel values of intermediate results are clipped after each step to ensure that they are in an \( \epsilon \)-neighbourhood of the original image. The BIM method can be easily adapted into an attacker for a specific desired target class, called Iterative Target Class Method [57]:

\[
I_{adv}^0 = I, \quad I_{adv}^{k+1} = C_{clip}(I_{adv}^{k} - \epsilon \cdot \text{sign}(\nabla_I J(I_{adv}^{k}, y_{target}))),
\]

We utilize this Iterative Target Class Method to update our perturbation values during the training process. To achieve a balance between the attack efficiency and the perturbation perceptibility, we constrain the perturbation values in the loss function of Eq. (10) instead of using the clip operation. At each training step, our perturbations are updated as follows:

\[
\delta_{k+1} = \delta_k - \epsilon_1 \cdot \text{sign}(\nabla_{\delta_k} L)
\]

\[
p_{k+1} = p_k - \epsilon_2 \cdot \text{sign}(\nabla_{p_k} L),
\]

where \( \epsilon_1 \) and \( \epsilon_2 \) are to ensure that the perturbation added to the template/search image is small. During training, we only optimize the values of \( (\delta, p) \) and the network weights remain intact. We outline this training procedure in Algorithm 1.

C. Attacking the Tracker at Inference Time

Once the perturbations \( (\delta, p) \) are trained, we can use them to perturb the template and search images of any novel video for attacking. Both \( \delta \) and \( p \) are universal (i.e., video-agnostic), which means perturbing a novel video only involves the mere addition of the perturbations to the template and search images – and does not require gradient optimization or network inference. Assume \( B^{fake} = \{b^{fake}_i\}_I \) is the trajectory we hope the tracker to output. During tracking the \( i \)-th frame of the video \( V = \{I_t\}_I \), we need to add \( p \) into \( x_t \) according to \( b^{fake} = \{x_0, y_0, x_1, y_1\} \):

\[
\hat{x}_i = A_{add}(x_t, p, \{x_0, y_0, x_1, y_1\}).
\]

The tracker then takes \( \hat{z}_1 = z_1 + \delta \) and \( \hat{x}_i \) as input, and the subsequent tracking procedure remains the same as SiamFC++. We outline this procedure in Algorithm 2.

IV. EXPERIMENTS

A. Experimental Setup

Evaluation Benchmarks. We evaluate our video-agnostic perturbation method for targeted attacks on several tracking benchmarks, i.e., OTB2015 [11], GOT-10k [12], LaSOT [13], VOT2016 [14], VOT2018 [15] and VOT2019 [16]. Generally speaking, OTB2015 is a typical tracking benchmark which is widely used for evaluation for several years. GOT-10k has the advantage of magnitudes wider coverage of object classes. LaSOT has much longer video sequences with an average duration of 84 seconds. OTB2015, GOT-10k and LaSOT all...
follow the One-Pass Evaluation (OPE) protocol and their evaluation methodologies are similar as the measurement is mostly based on the success and precision of the trackers over the test videos. For instance, they all measure the success based on the fraction of frames in a sequence where the intersection-over-union (IoU) overlap of the predicted and ground-truth rectangles exceeds a given threshold, and then the trackers are ranked using the area-under-the-curve (AUC) criterion. Since the average of IoU overlaps (AO) over all the test video frames is recently proved to be equivalent to the AUC criterion, we thus denote the success measurement as AO in the following. Besides AO, a success rate (SR) metric is also directly used to measure the percentage of successfully tracked frames given a threshold as in GOT-10k. As for the precision, it encodes the proportion of frames for which the ground-truth bounding box is proposed and the trackers are then ranked using the AUC for normalized precision between 0 and 0.5. VOT [14], [15], [16] introduces a series of tracking competitions with up to 60 sequences in each of them, aiming at the study of object tracking, while in practical applications, the target position of the first frame is predefined in the academic setting.

### Table I

| Dataset          | Videos | Total frames | Frame rate | Object classes | Num. of attributes |
|------------------|--------|--------------|------------|----------------|-------------------|
| **Training set** |        |              |            |                |                   |
| GOT-10k training split | 9.34K  | 1.4M         | 10 fps     | 480            | 6                 |
| LaSOT training split | 1.12K  | 2.83M        | 30 fps     | 70             | 14                |
| COCO2017         | n/a    | 118K         | n/a        | 80             | n/a               |
| ILSVRC-VID       | 5.4K   | 1.6M         | 30 fps     | 30             | n/a               |
| **GOT-10k validation split** | 180    | 21K          | 10 fps     | 150            | 6                 |
| LaSOT test split | 280    | 690K         | 30 fps     | 70             | 14                |
| OTB-15           | 100    | 50K          | 30 fps     | 22             | 11                |
| VOT2016          | 60     | 21K          | 30 fps     | 16             | 6                 |
| VOT2018          | 60     | 21K          | 30 fps     | 24             | 6                 |
| VOT2019          | 60     | 19K          | 30 fps     | 30             | 6                 |

### Generating the Fake Trajectory

We need to predefine a specific trajectory $B^{fake}_{1} = \{b^{fake}_{i}\}_{i=1}^T$ for each video to achieve targeted attack in the online-tracking phase, which we call the fake trajectory. We denote the real trajectory as $B^{gt} = \{b^{gt}_{i}\}_{i=1}^T$ with the ground-truth bounding boxes. It is possible to manually label arbitrary $B^{fake}$ for each video, however, it will be time-consuming in our experimental evaluation. So we generate $B^{fake}$ based on $B^{gt}$. Specifically, the fake trajectory follows the real trajectory except that the adjacent boundaries of $b^{fake}_{i}$ and $b^{gt}_{i}$ are 2 pixels apart (Fig. 4). Because the annotations of GOT-10k’s test split are kept private, we choose to use its validation set for our targeted attack evaluation and denote it as GOT-Val. Note that the target position of the first frame is predefined in the academic setting.

### Image Quality Assessment

We use structural similarity (SSIM) [62] to evaluate the quality and perceptibility of our generated perturbations $\delta$ and $p$. It is difficult to perceive the perturbation when SSIM is close to 1 (see Table V).

### B. Implementation Details

In our evaluation, the backbone Siamese network of our base tracker SiamFC++ [10] adopts GoogLeNet [63]. We implement our approach in Pytorch and train our perturbations using three GTX 1080Ti GPUs. We adopt COCO [64], ILSVRC-VID [65] and the training splits of GOT-10k [12] and LaSOT [13] as our training set. We train the perturbations for 8192 iterations with a mini-batch of 96 images (32 images per GPU). Both the hyper-parameters $\epsilon_1$ and $\epsilon_2$ for the template perturbation and the adversarial patch are set to 0.1. We generate training samples following the practice in SiamFC++. During both the training and online-tracking phases, the offline-shelf SiamFC++ tracking network model is fixed and used for the whole evaluation, the spatial size of the template image is set to $127 \times 127$, and the search image is $303 \times 303$. In Eq. (10), we set $\alpha = 1, \beta = 1, \gamma = 1, \eta_1 = 0.005$, and $\eta_2 = 0.005$.

### C. Overall Attack Results on the Evaluation Benchmarks

We test the performance of our targeted attack method on the evaluation benchmarks and gather the overall results in Table II and Table III. We evaluate the untargeted and

1. It is exclusively trained on the training split of GOT-10k by the authors of SiamFC++ and can be downloaded from https://drive.google.com/file/d/1BevcfI2s_kgyFjhxayOFw08DF12a5Z7i/view
targeted attack performance using the following metrics on specific datasets: (1) Accuracy, Robustness and EAO on VOT, (2) AO and Precision on OTB2015, (3) SR and AO on GOT-Val, and (4) Precision, Normalized Precision and AO on LaSOT. It is shown that the base tracker SiamFC++ can achieve state-of-the-art performance on all the evaluation benchmarks and run in real time (at about 58 fps on a GTX 1080Ti GPU). However this real-time performance requires the computationally intensive tracking system to occupy most of the computational resources, and thus it is appealing to develop a virtually costless attack method to fool the tracking system without scrambling for the resources. As shown in Tables II and III, our method can satisfy this appealing demand and fool the SiamFC++ tracker effectively by misleading the tracker to follow a predefined fake trajectory. Moreover, the high AO and Precision performance calculated by aligning with the fake trajectory indicates a more effective targeted attack without raising possible suspicion (see last column of Table III).

**D. Analyses of the Perceptibility of Our Perturbations**

We firstly compare our attack method with two baseline methods to show the motivation behind our proposed approach. As introduced in Sec. III-A, our goal is deteriorating the tracking performance of Siamese trackers at a low computational cost as well as making the perturbations less obvious. Baseline-1 performs untargeted attacks based on the UAP [9] method. The difference between Baseline-1 and our method is that, Baseline-1 generates one perturbation added to the whole search images and cannot perform targeted attacks, while our method adds a small universal patch to the search images to perform targeted attacks (see Fig. 5).

Baseline-2 performs targeted attacks based on the adversarial patch [50] method. The differences between Baseline-2 and our method include: (1) Baseline-2 pastes the adversarial patch into the search images while Baseline-1 generates one perturbation added to the whole search images and cannot perform targeted attacks. Baseline-2 pastes the obviously noticeable patch into the search images, which obviously raises the risk of being suspected. Our method adds a small universal patch to the search images to perform targeted attacks, which is as translucent as in Baseline-1.

We also examine the influence of the number of training iterations on the perceptibility of our perturbations. We use SSIM value to indicate the perceptibility, which ranges from 0 to 1. If the SSIM value is close to 1, the difference between the perturbed image and the original image is small, which means it is imperceptible. It can be observed in Table IV that as the number of iterations increases, the AO score with respect to the fake trajectory increases significantly, while it decreases significantly with respect to the real trajectory. After about 8000 training iterations, the resulting perturbations prevent the tracker from tracking most of the targets in GOT-Val, and the AO with respect to the real trajectory decreases from 0.760 to 0.153. Note that the AO decreases significantly faster at the beginning (training iterations less than 2048). This demonstrates the fast convergence capability of our method. However, the SSIM values of our perturbed images gradually decrease as the training proceeds. At the iteration number of 8192, the SSIM value of our perturbed template image decreases to 0.79, and 0.56 for the perturbed search image. Note that the SSIM for our perturbed search image is calculated in the sub-region where the patch is placed.

We further compare our method with FAN [6] to examine the perceptibility. FAN independently generates different perturbations for each frame, while our perturbations are

---

**TABLE II**

| Benchmarks | Metrics | Before Attack | Untargeted Attack | Targeted Attack |
|------------|---------|---------------|-------------------|-----------------|
| VOT2016    | Accuracy | 0.626         | 0.393             |                 |
| VOT2018    | Accuracy | 0.587         | 0.342             |                 |
| VOT2019    | Accuracy | 0.556         | 0.345             |                 |

**TABLE III**

| Benchmarks | Metrics | Before Attack | Untargeted Attack | Targeted Attack |
|------------|---------|---------------|-------------------|-----------------|
| OTB2015    | AO      | 0.642         | 0.063             | 0.759           |
|            | Precision| 0.861         | 0.092             | 0.795           |
| GOT-Val    | SR      | 0.897         | 0.123             | 0.890           |
|            | AO      | 0.760         | 0.153             | 0.840           |
| LaSOT      | Precision| 0.514         | 0.046             | 0.605           |
|            | Norm. Prec. | 0.551        | 0.048             | 0.702           |
| FPS        | 58      | 58            | 58                |                 |

**TABLE IV**

| Methods   | Untargeted Attack | Targeted Attack |
|-----------|-------------------|-----------------|
| Baseline-1 | 0.09              |                 |
| Baseline-2 | 0.23              | 0.78            |
| Ours      | 0.15              | 0.84            |
Table V

| Iterations | Targeted Attack AO | Targeted Attack SR | Untargeted Attack AO | Untargeted Attack SR |
|------------|--------------------|--------------------|----------------------|----------------------|
| 1          | 0.14               | 0.1                | 0.76                 | 0.89                 |
| 2          | 0.14               | 0.1                | 0.77                 | 0.89                 |
| 4          | 0.14               | 0.1                | 0.76                 | 0.89                 |
| 8          | 0.14               | 0.1                | 0.76                 | 0.89                 |
| 16         | 0.15               | 0.11               | 0.75                 | 0.88                 |
| 32         | 0.18               | 0.11               | 0.73                 | 0.86                 |
| 64         | 0.47               | 0.11               | 0.48                 | 0.53                 |
| 128        | 0.73               | 0.11               | 0.27                 | 0.25                 |
| 256        | 0.78               | 0.11               | 0.17                 | 0.18                 |
| 512        | 0.82               | 0.11               | 0.15                 | 0.12                 |
| 1024       | 0.84               | 0.11               | 0.15                 | 0.12                 |
| 2048       | 0.84               | 0.11               | 0.15                 | 0.12                 |
| 4096       | 0.84               | 0.11               | 0.15                 | 0.12                 |
| 8192       | 0.84               | 0.11               | 0.15                 | 0.12                 |

Fig. 6. The results under targeted attacks compared with FAN [6]. FAN crafts the perturbations for each video independently, which comes at a non-negligible computational cost. To achieve video-agnostic universal attacks, we relax the constraint on the value of the perturbations, achieving a balance between the attack efficiency and the perturbation perceptibility.

Fig. 7. AO with respect to the real/fake trajectory on the perturbed GOT-Val dataset using three different patch sizes. The 32×32 patch can make the AO score decrease to 0.146 dramatically.

We note that it is important to make the size of the adversarial patch get small to avoid possible suspicion. So we experimentally examine the influence of the adversarial patch size in Fig. 7. We compare three different sizes of the adversarial patch, i.e., 64×64, 32×32 and 16×16, to conduct this experiment. We evaluate the AO score of SiamFC++ on the perturbed GOT-Val dataset using these three different patch sizes. As shown in Fig. 7 the 32×32 patch can make the AO score decrease to 0.146 dramatically and its area only accounts for 1% of the search image area.

Since we can only achieve a balance between the attack efficiency and the perturbation perceptibility, our universal perturbation method may be a double-edged sword as it may result in suspicious attacks. Prior works commonly train a network to prevent not only suspicious attacks but also modifying every pixel. So we also examine a new strategy to perturb the search images so as to reduce the possible suspicion. Specifically, we first convert one search image into YCbCr color space, and then add a perturbation to the entire search image in the Y channel and a different perturbation to a very small region of 64×64 in the CbCr channels. Different from RGB color space, YCbCr color space encodes a color image similar to human eyes’ retina, which separates the RGB components into a luminance component (Y) and two chrominance components (Cb as blue projection and Cr as red projection). We choose YCbCr color space since its color channels are less correlated than RGB [66]. Moreover, the available. Moreover, our universal perturbations achieve better targeted attack performance than FAN on OTB2015 as shown in Sec. IV-G.
CbCr channels are less sensitive to the human vision system than the Y channel [66], which means that the patch added in the CbCr channels has better performance of transparency. For the template image, we also convert it into YCbCr color space, and then add the perturbation to the entire template image in all YCbCr channels. The perturbed template and search images are finally converted into RGB color space and fed into the tracking network. The other steps of the training process are the same to the attack method in Sec. III. As shown in Table VI, the performance of attacking the YCbCr color space is slightly degraded compared to the RGB color space. However, attacking the YCbCr color space brings better imperceptibility (see Fig. 8). Compared with the RGB space, the SSIM value for the search images increases from 0.56 to 0.78. Note that the perturbation for the search images in the Y channel results in SSIM value of 0.95 outside the added patch area, which indicates good imperceptibility.

E. Other Analyses

Use Other Similar Random Patterns. To further illustrate the effectiveness of our proposed method, we manually generate similar random patterns and add them on the template and search regions (see Fig. 9) to show their attack performance. Specifically, we design two kinds of random patterns: (1) the random patterns similar to our trained perturbations, and (2) the random patterns generated using zero-mean Gaussian noise with standard deviation of 50.0. We replace the trained perturbations with the above random patterns to attack SiamFC++_GoogleNet, and experimentally evaluate them on GOT-Val. As shown in Table VII, these random patterns cannot effectively attack the tracker.

Only Perturb the Template or Search Image. To analyze the impact of $p$ and $\delta$ in our perturbations, we evaluate the attack performance when only adding perturbations on the template images or the search regions on GOT-Val. The result is shown in Table VIII. For the untargeted attack, only perturbing the template/search image leads to the AO of 0.510/0.714, while perturbing both the template and search images leads to the AO of 0.153. For the targeted attack, only perturbing the template/search image leads to the AO of 0.156/0.132, while perturbing both the template and search images leads to the AO of 0.840. Experimental results show that perturbing both of them can achieve better attack performance than only perturbing the template/search image.

Influence of Different Training Loss Components. We implement a series of experiments to analyze the contribution of each loss component. In Table IX we report the attack results on the GOT-Val dataset. The AO score of the tracker with respect to the real trajectory decreases from 0.760 to...
the tracker to follow a fixed direction in each video. For each video, we assign a random direction from 4 different directions, each of which consists of shifting the box away by (+3, ±3) pixels for each consecutive frame, corresponding to one of the four directions 45°, −45°, 135°, −135°. The attack performance is evaluated on GOT-Val. As shown in Table XI, our attack method achieves effective attacks under both of the two different fake trajectory generation ways.

\section*{F. Transferability Analyses}

In this part, we analyze the transferability of our proposed attack method. Specifically, we directly apply our perturbations trained on SiamFC++_GoogleNet to other tracking networks including SiamFC++_ShuffleNet, SiamFC++_AlexNet, SiamRPN, SiamRPN++ and Ocean.

Transferability to Different Backbones. We evaluate the transferability of our attacks when applying the perturbations to two more different backbones of SiamFC++, i.e., ShuffleNet \cite{26} and AlexNet \cite{27}. The experimental results are shown in Table XII. In the case of SiamFC++, AlexNet, the AO with respect to the real trajectory decreases from 0.72 to 0.496. Our perturbations also generalize well to SiamFC++_ShuffleNet, despite the customized components such as pointwise group convolution and channel shuffle operation in ShuffleNet.

Transferability to Different Tracking Architectures. We also evaluate the transferability of our attacks when applying the perturbations to three more state-of-the-art trackers: SiamRPN \cite{28}, SiamRPN++ \cite{29} and Ocean \cite{20}. SiamRPN and SiamRPN++ are anchor-based trackers, and Ocean is an anchor-free tracker. The experimental results are shown in Table XIII. In the case of SiamRPN, the AO with respect to the real trajectory decreases from 0.666 to 0.483 and the performance of SiamRPN++ is decreased from 0.676 to 0.418. In the case of Ocean, the AO with respect to the real trajectory decreases from 0.902 to 0.282. The results show good transferability of our attacks to different tracking architectures, even if the generated perturbations are applied to anchor-based trackers.
G. Comparison with Other Attack Methods

We firstly compare our attack method with the recent state-of-the-art attack methods on OTB2015, including CSA [55], RTAA [54], SPARK [7], FAN [6] and TTP [5]. Please refer to Sec. II-C for more details about these methods. The untargeted attack results in Table XIV show that our targeted attack method can also achieve superior untargeted attack performance as well as the RTAA, SPARK and TTP attack methods and even better results than CSA and FAN. What’s more, the targeted attack results in Table XV also show that our achieved precision score with respect to the fake trajectory is up to 0.795, which is significantly better than FAN and TTP. This further demonstrates our superior targeted attack performance when using the adversarial patch for attacks.

We also compare our untargeted attack results on VOT2016 and VOT2018 with CSA [55], RTAA [54], FAN [6] and TTP [5] in Table XV. In summary, compared with other attack methods, our method has two key advantages. First, the proposed attack method only needs to perform the addition operations to deploy the adversarial information for any novel video without gradient optimization or network inference, making it possible to attack a real-world online-tracking system when we can not get access to the limited computational resources. Second, the proposed perturbations show good transferability to other anchor-free or anchor-based trackers. The main limitation of our work is that the translucent perturbations may result in suspicious attacks.

V. CONCLUSION

In this paper, we propose a video-agnostic targeted attack method for Siamese trackers. We aim to attack the tracker by adding a perturbation to the template image and adding a fake target, i.e., a small adversarial patch, into the search image adhering to the predefined trajectory, so that the tracker outputs the location and size of the fake target instead of the real target. Being universal, the generated perturbations can be conveniently exploited to perturb videos on-the-fly without extra computations. Extensive experiments on several popular datasets show that our method can effectively fool the Siamese trackers in a targeted attack manner. In the future work, we expect that it will be possible to further reduce the possible suspicion by increasing the perturbation’s imperceptibility while maintaining the attack efficiency.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments and suggestions.

REFERENCES

[1] Y. Shan, X. Zhou, S. Liu, Y. Zhang, and K. Huang, “SiamFPN: A deep learning method for accurate and real-time maritime ship tracking,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 1, pp. 315–325, 2021.
[2] J. Fan, H. Song, K. Zhang, K. Yang, and Q. Liu, “Feature alignment and aggregation Siamese networks for fast visual tracking,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 4, pp. 1296–1307, 2021.
[3] D. Li, F. Porikli, G. Wen, and Y. Kuai, “When correlation filters meet Siamese networks for real-time complementary tracking,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 2, pp. 509–519, 2020.
[4] M. Jiang, Y. Zhao, and J. Kong, “Mutual learning and feature fusion Siamese networks for visual object tracking,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 8, pp. 3154–3167, 2021.
[5] K. K. Nakka and M. Salzmann, “Temporally-transferable perturbations: Efficient, one-shot adversarial attacks for online visual object trackers,” arXiv preprint arXiv:2012.15183, 2020.
[6] S. Liang, X. Wei, S. Yao, and X. Cao, “Efficient adversarial attacks for visual object tracking,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 34–50.
[7] J. Gao, X. Xie, F. Juefei-Xu, L. Ma, Z. Li, W. Xue, W. Feng, and Y. Liu, “SPARK: Spatial-aware online incremental attack against visual tracking,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 202–219.
B. Wang, M. Zhao, W. Wang, F. Wei, Z. Qin, and K. Ren, “Are you W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial attacks B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, “AutoTrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 11 920–11 929.

S. Liu, S. Wang, C.-T. Lin, and Z. Lv, “Fuzzy detection aided real-time and robust visual tracking under complex environments,” IEEE Trans. on Fuzzy Syst., vol. 29, no. 1, pp. 90–102, 2021.

Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, “Ocean: Object-aware anchor-free tracking,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 771–787.

Y. Cui, C. Jiang, L. Wang, and G. Wu, “Fully convolutional online tracking,” arXiv preprint arXiv:2004.07109, 2020.

D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, “SiamCAR: Siamese fully convolutional classification and regression for visual tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6268–6276.

B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with Siamese region proposal network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8971–8980.

B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “SiamRPN++: Evolution of Siamese visual tracking with very deep networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4282–4291.

S. M. Marvasti-Zadeh, L. Cheng, H. Ghanefar-Yakhdan, and S. Kasaei, “Deep learning for visual tracking: A comprehensive survey,” IEEE Trans. Intell. Transp. Syst., pp. 1–26, 2021.

B. Wang, M. Zhao, W. Wang, F. Wei, Z. Qin, and K. Ren, “Are you confident that you have successfully generated adversarial examples?” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 6, pp. 2089–2099, 2021.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

S. Ren, Y. Deng, K. He, and W. Che, “Generating natural language adversarial examples through probability weighted word saliency,” in Proc. Assoc. Comput. Ling., 2019, pp. 1085–1097.

W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial attacks on deep-learning models in natural language processing: A survey,” ACM Trans. Intell. Syst. Technol., vol. 11, no. 4, pp. 6268–6276, 2020.

J. X. Morris, E. Lifland, J. Y. Yoo, and Y. Qi, “TextAttack: A framework for adversarial attacks in natural language processing,” arXiv preprint arXiv:2005.05909, 2020.

D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, “Is BERT really robust? A strong baseline for natural language attack on text classification and entailment,” in Proc. Assoc. Adv. Artif. Intell., 2020, pp. 8018–8025.

X. Wei, S. Liang, N. Chen, and X. Cao, “Transferable adversarial attacks for image and video object detection,” in Proc. Int. Joint Conf. Artif. Intell., 2019, pp. 954–960.

L. Xiong, X. Han, C.-N. Yang, and Y.-Q. Shi, “Robust reversible watermarking in encrypted image with secure multi-party based on lightweight cryptography,” IEEE Trans. Circuits Syst. Video Technol., pp. 1–1, 2021.

S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, “3D face anti-spoofing with factorized bilinear coding,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 5031–5045, 2021.

L. Meng, C.-T. Lin, T.-P. Jung, and D. Wu, “White-box target attack for EEG-based BCI regression problems,” in Proc. Int. Conf. Neural Inf. Process., 2019, pp. 476–488.

M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, “Query-efficient hard-label black-box attack: An optimization-based approach,” arXiv preprint arXiv:1807.04457, 2018.

Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, “NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 3866–3876.

N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in Proc. Conf. Comput. Commun. Security, 2017, pp. 506–519.

J. Li, R. Ji, H. Liu, J. Li, B. Zhong, C. Deng, and Q. Tian, “Projection & probability-driven black-box attack,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 362–371.

A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, H. Xu, C. Xie et al., “Adversarial attacks and defences competition,” arXiv preprint arXiv:1804.00697, 2018.

H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, “Adversarial attack on graph-structured data,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 1115–1124.

B. Li, C. Chen, W. Wang, and L. Carin, “Second-order adversarial attack and certifiable robustness,” arXiv preprint arXiv:1809.03113, 2018.

Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, “Tactics of adversarial attack on deep reinforcement learning agents,” arXiv preprint arXiv:1703.06748, 2017.

V. Khurulkov and I. Oseledets, “Art of singular vectors and universal adversarial perturbations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8562–8570.

K. R. Mopuri, U. Ojha, U. Garg, and R. V. Babu, “NAG: Network for adversary generation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 742–751.

C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon, “Understanding adversarial examples from the mutual influence of images and perturbations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14 530.

K. R. Mopuri, A. Ganeshan, and R. V. Babu, “Generalizable data-free objective for crafting universal adversarial perturbations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 10, pp. 2452–2465, 2018.

S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, “ShapeShifter: Robust physical adversarial attack on Faster R-CNN object detector,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases, 2018, pp. 52–68.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017.

T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” arXiv preprint arXiv:1712.09665, 2017.

D. Karmon, D. Zoran, and Y. Goldberg, “LaVAN: Localized and visible adversarial noise,” in Proc. Int. Conf. Mach. Learn., 2018, pp. 2507–2515.

R. R. Wiyatno and A. Xu, “Physical adversarial textures that fool visual object tracking,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 4822–4831.

D. Held, S. Thrun, and S. Savarese, “Learning to track at 100 FPS with deep regression networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 749–765.

S. Jia, C. Ma, Y. Song, and X. Yang, “Robust tracking against adversarial attacks in learning,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 69–84.

B. Yan, D. Wang, H. Lu, and X. Yang, “Cooling-Shrinking Attack: Blinding the tracker with imperceptible noises,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 990–999.
[56] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.
[57] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in Proc. Int. Conf. Learn. Representation, 2017.
[58] A. Shafahi, M. Najibi, Z. Xu, J. Dickerson, L. S. Davis, and T. Goldstein, “Universal adversarial training,” in Proc. Assoc. Adv. Artif. Intell., 2020, pp. 5636–5643.
[59] H. Hirano and K. Takemoto, “Simple iterative method for generating targeted universal adversarial perturbations,” Algorithms, vol. 13, no. 11, p. 268, 2020.
[60] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2980–2988.
[61] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “UnitBox: An advanced object detection network,” in Proc. ACM Multimedia, 2016, pp. 516–520.
[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, 2004.
[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.
[64] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.
[65] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.
[66] Y. Tan, J. Qin, X. Xiang, W. Ma, W. Pan, and N. N. Xiong, “A robust watermarking scheme in YCbCr color space based on channel coding,” IEEE Access, vol. 7, pp. 25 026–25 036, 2019.
[67] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6848–6856.
[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

Zhenghang Li received the B.S. degree in computer science and technology from Beijing Institute of Technology, Beijing, China, in 2016. Currently, he is a Ph.D. student with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include object tracking and deep learning.

Yaya Shi received the B.S. degree from the Hefei University of Technology, Anhui, China in 2018. Currently, She is a Ph.D. student with the University of Science and Technology of China, Anhui, China. Her research interests include video captioning and deep learning.

Jin Gao received the B.S. degree from the Beihang University, Beijing, China, in 2010, and the Ph.D. degree from the University of Chinese Academy of Sciences (UCAS), in 2015. Now he is an associate professor with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests include visual tracking, autonomous vehicles, and service robots.

Shaoru Wang received the B.S. degree from the Department of School of Information and Communication Engineering in Beijing University of Posts and Telecommunications, Beijing, China in 2018. He is currently pursuing his Ph.D. degree at the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His current research interests include computer vision and deep learning.

Bing Li received the Ph.D. degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, Beijing, China, in 2009. From 2009 to 2011, he worked as a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently a Professor with CASIA. His current research interests include computer vision, color constancy, visual saliency detection, multi-instance learning, and data mining.

Pengpeng Liang received the B.S. and M.S. degrees in computer science from Zhengzhou University, China in 2008 and 2011, respectively and the Ph.D. degree from Temple University in 2016. He worked at Amazon from 2016 to 2017. Then, he joined Zhengzhou University as an assistant professor in 2018. Currently, he is pursuing his Ph.D. degree at the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include computer vision and deep learning.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University, Zhejiang, China. Since 1998, he has been with the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, where he is currently a Professor. He has published more than 200 papers on peer reviewed international conferences and journals. His current research interests include visual motion analysis and recognition of harmful Internet multimedia.