

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

This work is supported by National Key R&D Program of China (No. 2018AAA0102802, No. 2018AAA0102803, No. 2018AAA0102800), Natural Science Foundation of China (Grant No. 61972394, 62036011, 61721004), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040. Jin Gao and Bing Li are also supported in part by the Youth Innovation Promotion Association, CAS. (*Corresponding author: Jin Gao.*)

Z. Li, J. Gao, S. Wang and B. Li are with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190, China (e-mail: {zhenbang.li, jin.gao, bli}@nlpr.ia.ac.cn, wangshaoru2018@ia.ac.cn).

Y. Shi is with University of Science and Technology of China, Anhui, 230026, China (e-mail: shiyaya@mail.ustc.edu.cn).

P. Liang is with Zhengzhou University, Henan, 450001, China (e-mail: liangpc@gmail.com).

W. Hu is with CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190, China (e-mail: wmhu@nlpr.ia.ac.cn).

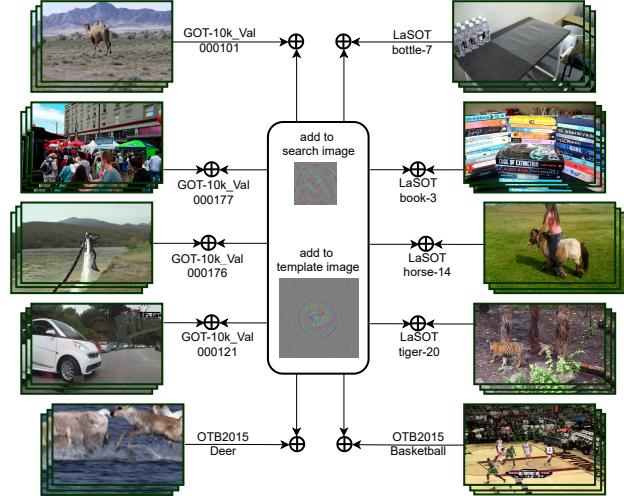


Fig. 1. Being universal, our generated perturbations can be conveniently exploited to perturb videos on-the-fly without extra intensive computations except the mere addition operations.

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 unseen data on-the-fly without extra intensive computations. However, no existing work has touched the topic of attacking the Siamese trackers using UAPs, because it is hard to apply existing UAPs to attack Siamese trackers directly. The main reason lies in the fact that, (a) most UAPs are designed for typical neural networks with one image as input while Siamese networks accept both the template and search images, and (b) the goal of existing UAP methods is to disturb unary or binary model outputs for single instance while we need to use universal perturbations to mislead Siamese trackers to follow a specified trajectory.

In this paper, we make the first attempt in finding the universal perturbations (see Fig. 1) that fool a state-of-the-

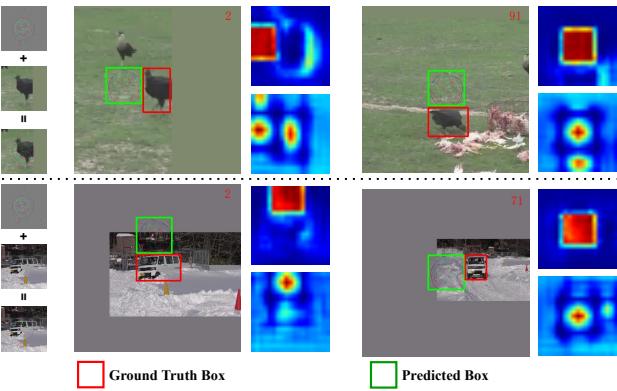


Fig. 2. An illustration of our attacks to SiamFC++ on some example tracking sequences from GOT-10k benchmark. Our approach generates the video-agnostic perturbations which can force SiamFC++ to follow a complicated trajectory at virtually no cost. This is realized by off-line training the perturbations so that the tracker mistakenly believes the *fake target* area contains the object to be tracked (see the top heatmap). Moreover, the quality assessment branch in SiamFC++ is also misled to confirm this result (see the bottom heatmap). The *fake target* size is gradually decreased.

art Siamese tracker, i.e., SiamFC++ [10], in a targeted attack manner, which comes at virtually no cost in the online-tracking phase. Specifically, we aim to attack the trackers by adding a universal perturbation to the template image and adding a *fake target*, i.e., a small universal adversarial patch, into the search image adhering to the predefined trajectory (as shown in Fig. 2), so that the tracker outputs the location and size of the *fake target* instead of the real target. Our generated video-agnostic perturbations allow 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 OTB2015 [11], GOT-10k [12], LaSOT [13], VOT2016 [14], VOT2018 [15] and VOT2019 [16] demonstrate the effectiveness and efficiency of our approach.

II. RELATED WORK

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 play an important role in correlation filter-based tracking, and the recent works (e.g., [17], [18], [19]) have largely advance 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 its success in the decomposition of classification and state estimation, 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 groundtruth 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 network. 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 scenarios 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 [35], 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 [36], [37], [38], [39], the adversary has limited or no knowledge about the model under attack [40]. In this paper, we focus on the white box attack for Siamese trackers. However, we surprisingly find that the trained 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 used for every datapoint of 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. However, RTAA only performs the untargeted attacks for 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 system without raising possible suspicion. SPARK [7] computes incremental perturbations by using information from the past frames to perform targeted attacks on 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. The recent real-time attacker TTP [5] exclusively uses the template image to generate temporally-transferable perturbation in a one-shot manner, and then adds it to every search image. However, this method still needs to generate perturbations 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 can not 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 $V = \{I_i\}_1^T$ denote the frames of a video sequence of length T . $B^{gt} = \{b_i^{gt}\}_1^T$ are used to represent the target's ground-truth positions in those frames. Visual object tracking aims to predict the positions $B^{pred} = \{b_i^{pred}\}_1^T$ 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_1^{gt} to get a template image \mathbf{z}_1 , and transforms the search frame I_i to get the search image \mathbf{x}_i centered at the position estimated in the previous frame. At each time-step, the template image \mathbf{z}_1 and the search image \mathbf{x}_i are first passed individually through a shared backbone network, i.e., Siamese network, and then fused using a channel-wise correlation operation:

$$\text{Feat}_j(\mathbf{z}_1, \mathbf{x}_i) = \psi_j(\phi(\mathbf{z}_1)) \star \psi_j(\phi(\mathbf{x}_i)), j \in \{\text{cls}, \text{reg}\} \quad (1)$$

where \star 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). ψ_{cls} and ψ_{reg} are both designed as two convolutional 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

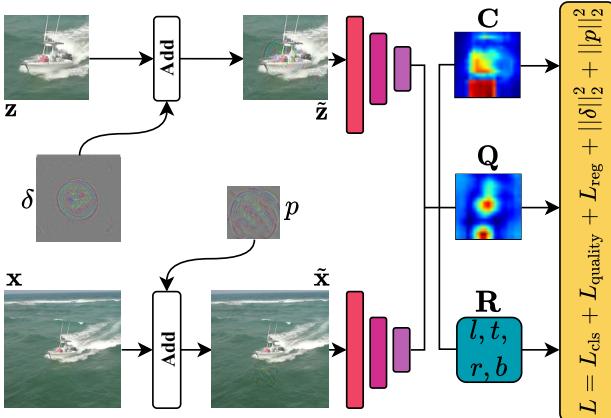


Fig. 3. The training pipeline of the proposed method. We aim to train a translucent perturbation δ for the template image \mathbf{z} , and an adversarial patch p for the search image \mathbf{x} . After adding δ to \mathbf{z} and adding p into \mathbf{x} , the tracker outputs the location and size of the *fake target* instead of the real target.

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 for Siamese trackers is directly using popular attack methods such as FGSM [55] and BIM [56]:

$$\begin{aligned} \min_{\delta_{\mathbf{x}_i}, \delta_{\mathbf{z}_i}} & L(y_i^{fake}, f(\mathbf{x}_i + \delta_{\mathbf{x}_i}, \mathbf{z}_i + \delta_{\mathbf{z}_i})) \\ \text{s.t. } & \|\delta_{\mathbf{x}_i}\|_p \leq \epsilon_1, \|\delta_{\mathbf{z}_i}\|_p \leq \epsilon_2, \end{aligned} \quad (2)$$

where L is the loss function of the SiamFC++ tracker. $(\mathbf{x}_i, \mathbf{z}_i)$ is the search-template pair of the video being tracked at frame i and $\mathbf{z}_i \equiv \mathbf{z}_1$. $y_i^{fake} = \{\mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*\}$ is derived from the *fake target* for the loss function and generated according to the fake groundtruth box b_i^{fake} that we want the tracker to output at frame i (see Sec. III-B). Additionally, the perturbation δ should be sufficiently small, which is commonly modeled through an upper bound ϵ 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 (\mathbf{x}, \mathbf{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], [57] are more practical for attacking Siamese trackers:

$$\begin{aligned} \min_{\delta_{\mathbf{x}}, \delta_{\mathbf{z}}} & \mathbb{E}_{(\mathbf{x}, \mathbf{z})} L(y^{fake}, f(\mathbf{x} + \delta_{\mathbf{x}}, \mathbf{z} + \delta_{\mathbf{z}})) \\ \text{s.t. } & \|\delta_{\mathbf{x}}\|_p \leq \epsilon_1, \|\delta_{\mathbf{z}}\|_p \leq \epsilon_2, \end{aligned} \quad (3)$$

where the input pair (\mathbf{x}, \mathbf{z}) is randomly picked from the training set. Note that universal adversarial perturbations are added to the whole image and y^{fake} is kept constant during the training process. As shown in [58], this vanilla UAP method is limited to untargeted attacks. To achieve universal targeted attacks, one feasible way is to paste 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 adversarial patch instead of the real target:

$$\min_{p_{\mathbf{x}}} \mathbb{E}_{(\mathbf{x}, \mathbf{z}, y^{fake})} L(y^{fake}, f(A_{paste}(\mathbf{x}, p_{\mathbf{x}}, b_{\mathbf{x}}^{fake}), \mathbf{z})), \quad (4)$$

where A_{paste} is a patch application operator [50] which pastes the adversarial patch $p_{\mathbf{x}}$ into the search image according to $b_{\mathbf{x}}^{fake}$. $b_{\mathbf{x}}^{fake} = \{x_0, y_0, x_1, y_1\}$ denotes the coordinates of the upper-left and lower-right corners of the *fake target* on the search image. y^{fake} is generated according to $b_{\mathbf{x}}^{fake}$ and both of them are variables during training. More specifically, A_{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 δ for the template image \mathbf{z} , and a translucent patch p for the search image \mathbf{x} . After adding δ to \mathbf{z} and adding the *fake target* patch p into \mathbf{x} , the tracker outputs the location and size of the adversarial patch instead of the real target (see Fig. 3). Both δ 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. This is achieved by

$$\begin{aligned} \min_{p_{\mathbf{x}}, \delta_{\mathbf{z}}} & \mathbb{E}_{(\mathbf{x}, \mathbf{z}, y^{fake})} L(y^{fake}, f(A_{add}(\mathbf{x}, p_{\mathbf{x}}, b_{\mathbf{x}}^{fake}), \mathbf{z} + \delta_{\mathbf{z}})) \\ \text{s.t. } & \|p_{\mathbf{x}}\|_p \leq \epsilon_1, \|\delta_{\mathbf{z}}\|_p \leq \epsilon_2, \end{aligned} \quad (5)$$

where A_{add} is a patch application operator which adds the patch into the search image according to $b_{\mathbf{x}}^{fake}$. The operator A_{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 $b_{\mathbf{x}}^{fake}$ area.

Our attacks on Siamese visual tracking with both the template and search images as input, enable us to exploit the advantages of both the perturbation generation method [55] 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 [5] 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

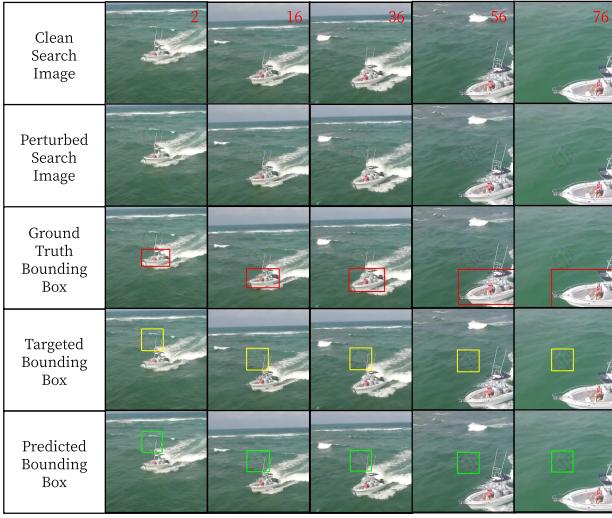


Fig. 4. Results of targeted attacks where the tracker is forced to follow a predefined *fake trajectory* $B^{fake} = \{b_i^{fake}\}_1^T$ (indicated by the yellow bounding boxes). The *fake trajectory* follows the real trajectory $B^{gt} = \{b_i^{gt}\}_1^T$ (indicated by the red bounding boxes) except that the adjacent boundaries of b_i^{fake} and b_i^{gt} are 2 pixels apart.

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 of the region where the real target exists on the heatmaps are always high. Thus it is necessary to perturb the template image to cooling down hot regions where the real target exists and increasing 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 $V = \{I_i\}_1^T$ 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 . We first randomly pick paired frames I_t, I_s from V . The clean template image $\mathbf{z} \in \mathbb{R}^{127 \times 127 \times 3}$ is generated according to I_t and b_t^{gt} , and the perturbed template image is:

$$\tilde{\mathbf{z}} = \mathbf{z} + \delta_k. \quad (6)$$

Similarly, the clean search image $\mathbf{x} \in \mathbb{R}^{303 \times 303 \times 3}$ 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:

$$\tilde{\mathbf{x}} = A_{add}(\mathbf{x}, p_k, b_k^{fake}), \quad (7)$$

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, A_{add} adds the patch into \mathbf{x} at location $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$. Subsequently, the SiamFC++ tracker takes $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{z}}$ as input and predicts \mathbf{C} , \mathbf{R} , \mathbf{Q} in an anchor-free manner.

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

For classification, a location (x, y) in the feature map ψ_{cls} is considered as a positive sample (i.e., $\mathbf{C}_{x,y}^* = 1$) if its corresponding location $(\lfloor \frac{s}{2} \rfloor + xs, \lfloor \frac{s}{2} \rfloor + ys)$ in the input image falls into the fake bounding box, and considered as a negative sample (i.e., $\mathbf{C}_{x,y}^* = 0$) 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 $(\lfloor \frac{s}{2} \rfloor + xs, \lfloor \frac{s}{2} \rfloor + ys)$ to the four edges of the fake bounding box, denoted as the four-dimensional vector $\mathbf{t}^* = (l^*, t^*, r^*, b^*)$. Thus, the fake regression label \mathbf{R}^* at location (x, y) can be expressed as:

$$\begin{aligned} l^* &= (\lfloor \frac{s}{2} \rfloor + xs) - x_0, & t^* &= (\lfloor \frac{s}{2} \rfloor + ys) - y_0 \\ r^* &= x_1 - (\lfloor \frac{s}{2} \rfloor + xs), & b^* &= y_1 - (\lfloor \frac{s}{2} \rfloor + ys) \end{aligned} \quad (8)$$

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 1×1 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 \mathbf{Q}^* at location (x, y) can be expressed as:

$$\text{IoU}^* = \frac{\text{Intersection}(b_k^{pred}, b_k^{fake})}{\text{Union}(b_k^{pred}, b_k^{fake})} \quad (9)$$

Training Objective. The loss function is calculated as follows:

$$\begin{aligned} L &= \frac{\alpha}{N_{pos}} \sum_{x,y} L_{cls}(\mathbf{C}_{x,y}, \mathbf{C}_{x,y}^*) \\ &+ \frac{\beta}{N_{pos}} \sum_{x,y} \mathbf{1}_{\{\mathbf{C}_{x,y}^* > 0\}} L_{quality}(\mathbf{Q}_{x,y}, \mathbf{Q}_{x,y}^*) \\ &+ \frac{\gamma}{N_{pos}} \sum_{x,y} \mathbf{1}_{\{\mathbf{C}_{x,y}^* > 0\}} L_{reg}(\mathbf{R}_{x,y}, \mathbf{R}_{x,y}^*) \\ &+ \eta_1 \cdot \|\delta_k\|_2^2 + \eta_2 \cdot \|p_k\|_2^2, \end{aligned} \quad (10)$$

where $\mathbf{C}_{x,y}, \mathbf{R}_{x,y}, \mathbf{Q}_{x,y}$ represent the values of $\mathbf{C}, \mathbf{R}, \mathbf{Q}$ at location (x, y) respectively, and $\mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*$ are the fake labels generated according to the position and size of the *fake target*. $\mathbf{1}$ is the indicator function that takes 1 if the condition in subscribe holds and takes 0 if not. N_{pos} denotes the number of positive samples in the training phase, L_{cls} denotes the focal loss [59] for classification result, $L_{quality}$ denotes the binary cross entropy (BCE) loss for quality assessment, and L_{reg} denotes the IoU loss [60] 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., [55], [56]). One of the simplest methods to generate adversarial image I^{adv} is FGSM [55] and works by linearizing the loss function around the network weights and obtaining an optimal max-norm constrained perturbation for generating the adversarial image:

$$I^{adv} = I + \epsilon \cdot \text{sign}(\nabla_I J(I, y_{true})), \quad (11)$$

Algorithm 1 Training Process

Input: Training dataset \mathcal{V} , Siamese tracker f , and max iteration number N .
Output: δ, p .

- 1: Let $k = 0$.
- 2: **while** $k < N$ **do**
- 3: Randomly pick a video $V \in \mathcal{V}$. The corresponding ground truth is $B^{gt} = \{b_i^{gt}\}_1^T$.
- 4: Randomly pick paired frames I_t, I_s from V .
- 5: Generate template image \mathbf{z} according to I_t and b_t^{gt} .
- 6: $\tilde{\mathbf{z}} = \mathbf{z} + \delta_k$.
- 7: Generate search image \mathbf{x} according to I_s and b_s^{gt} .
- 8: Calculate the *fake target* position $\{x_0, y_0, x_1, y_1\}$ with respect to the search image.
- 9: $\tilde{\mathbf{x}} = A_{add}(\mathbf{x}, p_k, \{x_0, y_0, x_1, y_1\})$.
- 10: $\mathbf{C}, \mathbf{R}, \mathbf{Q} = f(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$.
- 11: Generate fake labels $\mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*$ using $\{x_0, y_0, x_1, y_1\}$.
- 12: Calculate loss $L(\mathbf{C}, \mathbf{R}, \mathbf{Q}, \mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*)$ using Eqn. 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** δ_N, p_N .

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 ϵ 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 [56]:

$$I_0^{adv} = I, \\ I_{N+1}^{adv} = Clip_{I, \epsilon} \left\{ I_N^{adv} + \epsilon \cdot \text{sign}(\nabla_I J(I_N^{adv}, y_{true})) \right\}, \quad (12)$$

where the pixel values of intermediate results are clipped after each step to ensure that they are in an ϵ -neighbourhood of the original image. The BIM method can be easily made into an attacker for a specific desired target class, called Iterative Target Class Method [56]:

$$I_0^{adv} = I, \\ I_{N+1}^{adv} = Clip_{I, \epsilon} \left\{ I_N^{adv} - \epsilon \cdot \text{sign}(\nabla_I J(I_N^{adv}, y_{target})) \right\}. \quad (13)$$

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:

$$\delta_{k+1} = \delta_k - \epsilon_1 \cdot \text{sign}(\nabla_{\delta_k} L) \quad (14)$$

$$p_{k+1} = p_k - \epsilon_2 \cdot \text{sign}(\nabla_{p_k} L), \quad (15)$$

where ϵ_1 and ϵ_2 are to ensure that the perturbation added to the template/search image is small. During training, we only optimize the values of (δ, 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 (δ, p) are trained, we can use them to perturb the template and search images of any novel video

Algorithm 2 Attack Process

Input: The trained perturbations δ and p , Siamese tracker f , video $V = \{I_i\}_1^T$. b_1^{gt} is the position of the real target in the first frame. $B^{fake} = \{b_i^{fake}\}_1^T$ is the trajectory we hope the tracker to output.
Output: $B^{pred} = \{b_i^{pred}\}_1^T$

- 1: Generate the clean template image \mathbf{z}_1 according to I_1 and b_1^{gt} .
- 2: Generate the perturbed template image $\tilde{\mathbf{z}}_1 = \mathbf{z}_1 + \delta$.
- 3: Let $i = 2$.
- 4: **while** $i \leq T$ **do**
- 5: Generate clean search image \mathbf{x}_i according to I_i and b_{i-1}^{pred} .
- 6: $b_i^{fake} = \{x_{0i}, y_{0i}, x_{1i}, y_{1i}\}$
- 7: Generate the perturbed search image $\tilde{\mathbf{x}}_i = A_{add}(\mathbf{x}_i, p, \{x_{0i}, y_{0i}, x_{1i}, y_{1i}\})$.
- 8: $\mathbf{C}, \mathbf{R}, \mathbf{Q} = f(\tilde{\mathbf{x}}_i, \tilde{\mathbf{z}}_1)$.
- 9: Generate the predicted bounding box b_i^{pred} according to $\mathbf{C}, \mathbf{R}, \mathbf{Q}$.
- 10: $i = i + 1$.
- 11: **end while**
- 12: **return** B^{pred}

for attacking. Both δ 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 $B^{fake} = \{b_i^{fake}\}_1^T$ is the trajectory we hope the tracker to output. During tracking the i -th frame of the video $V = \{I_i\}_1^T$, we need to add p into \mathbf{x}_i according to $b_i^{fake} = \{x_{0i}, y_{0i}, x_{1i}, y_{1i}\}$:

$$\mathbf{x}_i = A_{add}(\mathbf{x}_i, p, \{x_{0i}, y_{0i}, x_{1i}, y_{1i}\}). \quad (16)$$

The tracker then takes $\tilde{\mathbf{z}}_1 = \mathbf{z}_1 + \delta$ and $\tilde{\mathbf{x}}_i$ 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 groundtruth 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

TABLE I
CHARACTERISTICS OF THE DATASETS USED TO TRAIN AND EVALUATE THE PROPOSED ATTACK METHOD.

	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
Test set	GOT-10k validation split	180	21K	10 fps	150	6
	LaSOT test split	280	690K	30 fps	70	14
	OTB-15	100	59K	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

the proportion of frames for which the center of the predicted rectangle is within 20 pixels of the groundtruth center. Since the precision metric is sensitive to the resolution of the images and the size of the bounding boxes, a metric of normalized precision over the size of 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 to evaluate the performance of a tracker in a relatively short duration. Different from other datasets, the VOT dataset has a reinitialization module. When the tracker loses the target (i.e., the overlap is zero between the predicted result and the annotation), the tracker will be reinitialized for the remaining frames based on the groundtruth annotation. Three metrics are used to evaluate the performance of a tracker in VOT: (1) accuracy, (2) robustness and (3) EAO (expected average overlap). The accuracy measures how well the bounding box predicted by the tracker overlaps with the groundtruth bounding box. The robustness measures how many times the tracker loses the target (fails) during tracking. EAO combines accuracy and robustness to evaluate the overall performance of the tracker. Characteristics of these datasets are summarized in Table I.

Generating the Fake Trajectory. We need to predefine a specific trajectory $B^{fake} = \{b_i^{fake}\}_1^T$ for each video to achieve targeted attack in the online-tracking phase, which we call the *fake trajectory*. We denote the real trajectory as $B^{gt} = \{b_i^{gt}\}_1^T$ with the groundtruth 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_i^{fake} and b_i^{gt} 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 pre-defined in the academic study of object tracking, while in practical application, the target position of the first frame is often obtained by running an object detector on this frame. In this scenario, once the target position of the first frame is obtained, we can put our adversarial patch near the target to mislead the tracker. In the remaining frames, we can specify an arbitrary *fake trajectory* to place the adversarial patch.

Image Quality Assessment. We use structural similarity (SSIM) [61] to evaluate the quality and perceptibility of our

generated perturbations δ 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 [62]. We implement our approach in Pytorch and train our perturbations using three GTX 1080Ti GPUs. We adopt COCO [63], ILSVRC-VID [64] 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 ϵ_1 and ϵ_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 phase, the off-the-shelf SiamFC++ tracking network model¹ is fixed and used for the whole evaluation, the spatial size of the template image is set to 127×127 , and the search image is 303×303 . 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. 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

¹It 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/1BevcIEZr_kgyFjhxayOFw08DFI2u5Zi7/view

TABLE II
OVERALL ATTACK RESULTS ON VOT2016, VOT2018, VOT2019.

Benchmarks	Metrics	Before Attack	Untargeted Attack
VOT2016	Accuracy	0.626	0.393
	Robustness	0.144	9.061
	EAO	0.460	0.007
VOT2018	Accuracy	0.587	0.342
	Robustness	0.183	8.981
	EAO	0.426	0.007
VOT2019	Accuracy	0.556	0.345
	Robustness	0.537	8.824
	EAO	0.243	0.010

TABLE III
OVERALL ATTACK RESULTS ON OTB2015, GOT-VAL AND LASOT.

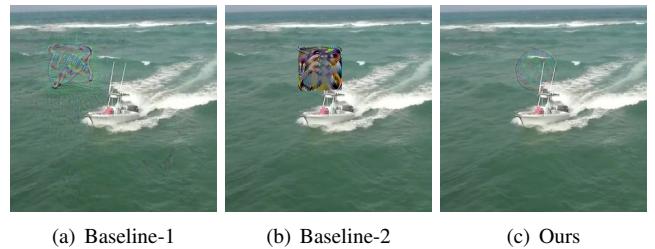
Benchmarks	Metrics	Before Attack	Untargeted Attack	Targeted Attack
OTB2015	AO	0.642	0.063	0.759
	Precision	0.861	0.092	0.795
GOT-Val	SR	0.897	0.123	0.890
	AO	0.760	0.153	0.840
LaSOT	Precision	0.514	0.046	0.605
	Norm. Prec.	0.551	0.048	0.702
	AO	0.525	0.069	0.691
FPS		58	58	58

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 perturbations added to the whole search image and cannot perform targeted attacks, while our method adds a *small* universal patch to the search image 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 an adversarial patch into the search image without the l_p -norm constraint as ours; (2) Baseline-2 uses the clean template image while we add translucent perturbation to the template image. As a consequence, Baseline-2 generates an obviously noticeable patch while ours is as translucent as in Baseline-1 (see Fig. 5). Moreover, our attack performance is also superior to Baseline-2 as shown in Table IV

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 V 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 8,000 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

TABLE IV
COMPARISON OF ATTACK PERFORMANCE WITH 2 BASELINE METHODS ON GOT-VAL IN TERMS OF AO. BASELINE-1 PERFORMS UNTARGETED ATTACKS BASED ON THE UAP [9] METHOD. BASELINE-2 PERFORMS TARGETED ATTACKS BASED ON THE ADVERSARIAL PATCH [50] METHOD.

Methods	Untargeted Attack	Targeted Attack
Baseline-1 [9]	0.09	-
Baseline-2 [50]	0.23	0.78
Ours	0.15	0.84



(a) Baseline-1 (b) Baseline-2 (c) Ours

Fig. 5. Visualization of perturbations on search images. Baseline-1 generates perturbations added to the whole search image and cannot perform targeted attacks. Baseline-2 pastes an obviously noticeable patch into the search image, which obviously raises the risk of being suspected. Our method adds a *small* universal patch to the search image to perform targeted attacks, which is as translucent as in Baseline-1.

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.*

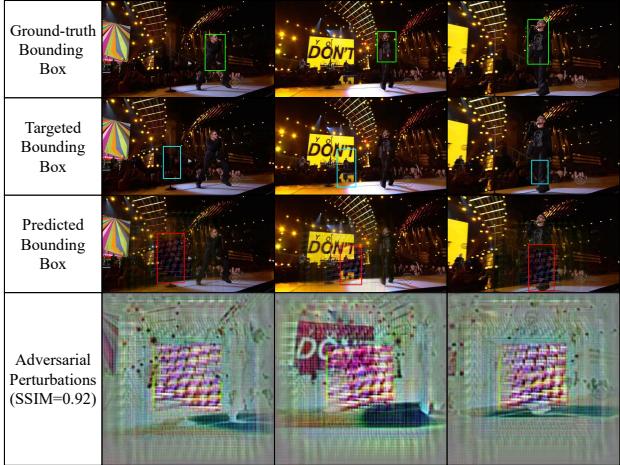
We further compare our method with FAN [6] to examine the perceptibility. FAN generates independent perturbation for each frame, while our perturbations are universal. FAN calculates the average SSIM value of all search images for each dataset, and *the SSIM for their perturbed search image is calculated in the whole area of the search image.* As shown in Fig. 6, both FAN and our method can perform targeted attack. Although FAN has its average SSIM value for OTB2015 dataset to only decrease to 0.92, its perturbation is also a little noticeable for some video sequences as shown in Fig. 6(a). Note that Fig. 6(a) is directly borrowed from their paper as the original perturbation results are not available. Moreover, our universal perturbations achieve better targeted attack performance than FAN on OTB2015, i.e., we achieve a precision score of 0.795, compared to 0.420 for FAN.

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-by-64, 32-by-32 and 16-by-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 Figure 7, the 32-by-32 patch can make the AO score decrease to 0.146 dramatically and its area only accounts for 1% of the search image area.

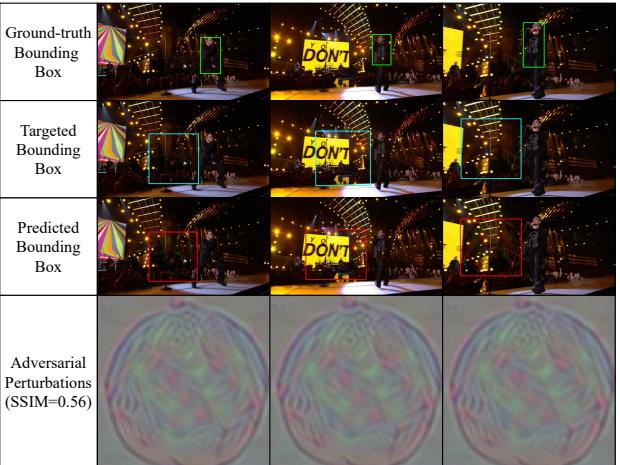
Since we can only achieve a balance between the attack efficiency and the perturbation perceptibility, our universal

TABLE V
INFLUENCE OF THE NUMBER OF TRAINING ITERATIONS ON GOT-VAL.

	Iterations	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192
Targeted Attack	AO	0.14	0.14	0.14	0.14	0.14	0.14	0.15	0.18	0.47	0.73	0.78	0.82	0.84	0.84
	SR	0.1	0.1	0.1	0.1	0.1	0.1	0.11	0.15	0.49	0.78	0.84	0.88	0.89	0.89
Untargeted Attack	AO	0.76	0.77	0.76	0.76	0.76	0.76	0.75	0.73	0.48	0.27	0.22	0.17	0.15	0.15
	SR	0.89	0.9	0.89	0.89	0.9	0.89	0.88	0.86	0.53	0.25	0.18	0.14	0.12	0.12
SSIM of δ		1	1	1	1	0.99	0.99	0.97	0.94	0.88	0.84	0.82	0.81	0.8	0.79
SSIM of p		0.98	0.98	0.98	0.98	0.98	0.98	0.93	0.78	0.56	0.50	0.51	0.52	0.53	0.56



(a) FAN



(b) Ours

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.

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 image so as to reduce the possible suspicion. Specifically, we first convert the 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 channel.

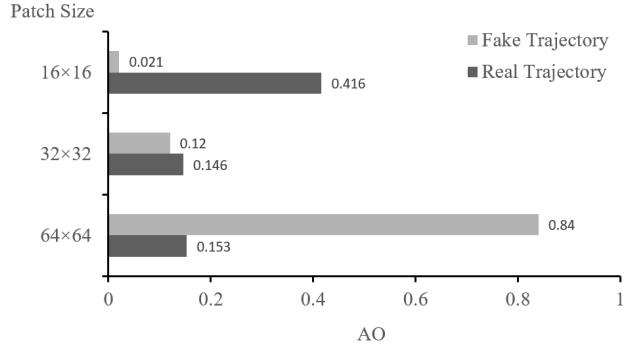


Fig. 7. AO with respect to the real/fake trajectory on the perturbed GOT-Val dataset using three different patch sizes. The 32-by-32 patch can make the AO score decrease to 0.146 dramatically.

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 [65]. Moreover, the CbCr channels are less sensitive to the human vision system than the Y channel [65], 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 image increases from 0.56 to 0.78. Note that the perturbation of the search image in the Y channel results in SSIM value of 0.95 outside the added patch area, which slightly deteriorate the 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 pattern similar to our trained

TABLE VI

COMPARISON BETWEEN ATTACKING THE DIFFERENT RGB AND YCbCr COLOR SPACES. ATTACKING THE YCbCr COLOR SPACE MEANS WE FIRST CONVERT THE 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 CHANNEL. THE ATTACK PERFORMANCE IS EVALUATED ON GOT-VAL.

Different Color Space Attack	Untargeted Attack		Targeted Attack		SSIM	
	AO	SR	AO	SR	δ	p
RGB Attack	0.153	0.123	0.840	0.890	0.56	0.79
YCbCr Attack	0.246	0.227	0.682	0.756	0.78	0.79

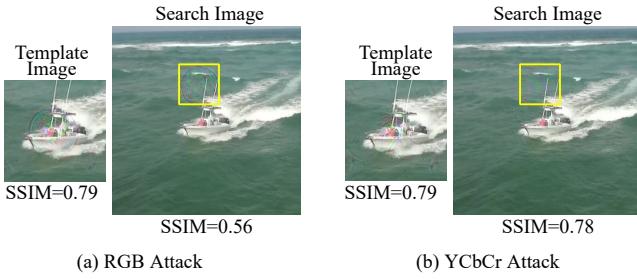


Fig. 8. Visualization of the perturbations in attacking the different RGB and YCbCr color spaces. Note that the perturbation of the search image in the Y channel results in SSIM value of 0.95 outside the added patch area, which slightly deteriorate the imperceptibility.

perturbations, and (2) the random pattern generated using zero-mean Gaussian noise with standard deviation 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 δ 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 image leads to the AO of 0.153. For the targeted attack, only perturbing the template/search image leads to the AO of 0.156/0.160, while perturbing both the template and search image 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 0.747 when adversarial information is generated using only quality assessment loss, indicating that quality assessment loss can cause a slight degradation in the performance of the tracker. However, the tracker's AO score with respect to the *fake trajectory* is only 0.148, indicating that the perturbations generated using quality assessment loss alone can barely cause the tracker to follow the specified trajectory. When adversarial information is generated using only classification loss, the

TABLE VII
ATTACK PERFORMANCE COMPARISON USING OTHER RANDOM PATTERNS.
IT IS EVALUATED ON GOT-VAL.

Perturbations used to perform attack	Untargeted Attack		Targeted Attack	
	AO	SR	AO	SR
Trained Perturbations	0.153	0.123	0.840	0.890
Similar Pattern	0.736	0.871	0.153	0.118
Gaussian Noise	0.740	0.875	0.144	0.101

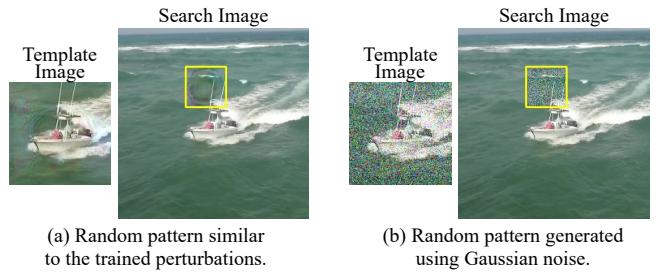


Fig. 9. Visualization of the manually generated similar random patterns.

tracker's AO score decreases from 0.760 to 0.282 with respect to the real trajectory and the AO with respect to the *fake trajectory* is up to 0.643. This is because the perturbation information generated using classification loss causes the tracker to localize to the fake target instead of the real target's location. In conclusion, all loss terms are beneficial, and the classification/regression term is more important than the quality assessment term.

Attack Performance w/o Groundtruth Information. Our perturbations are trained using datasets with groundtruth box information. However, we may only know the video content while the groundtruth box information is not available in practice. In case of this scenario, we can first run the tracker that needs to be attacked on these videos and generate one prediction result for each frame. The predicted boxes can be considered as groundtruth information for training the perturbations. To verify this, we run the off-the-shelf SiamFC++ tracker on videos in GOT-10k training set. The predicted bounding boxes are used to train our perturbations. We evaluate the untargeted and targeted attack performance using the following metrics on specific datasets: (1) AO and Precision on OTB2015, (2) SR and AO on GOT-Val, and (3) Precision, Normalized Precision and AO on LaSOT. As shown in Table X, it is effective to use the predicted boxes instead of groundtruth boxes for training our perturbations, though the targeted attack performance is affected due to some ambiguous prediction results used for training.

An Alternative Way to Generate Fake Trajectory. Besides the strategy to generate the *fake trajectory* following the real trajectory in Sec. IV-A, we also consider an alternative way to generate the *fake trajectory*. Specifically, the attacker forces the tracker to follow a fixed direction in each video. For different videos, we assign a random direction from 4 different directions for each of them, each of which consists of shifting the box away by $(\pm 3, \pm 3)$ pixels for each consecutive frame, corresponding to one of the four directions $45^\circ, -45^\circ, 135^\circ, -135^\circ$. The attack performance

TABLE VIII

ATTACK PERFORMANCE COMPARISON WHEN ONLY ADDING THE PERTURBATION ON THE SEARCH/TEMPLATE IMAGE. IT IS EVALUATED ON GOT-VAL.

Template	Search	Untargeted Attack		Targeted Attack	
		AO	SR	AO	SR
✓		0.510	0.567	0.156	0.106
	✓	0.714	0.841	0.160	0.132
✓	✓	0.153	0.123	0.840	0.890

TABLE IX

ANALYSIS OF THE IMPACT OF EACH LOSS COMPONENT ON GOT-VAL.

L_{cls}	$L_{quality}$	L_{reg}	Tragetted Attack		Untargeted Attack	
			AO	SR	AO	SR
✓			0.643	0.726	0.282	0.274
	✓		0.148	0.110	0.747	0.882
		✓	0.726	0.762	0.243	0.238
✓	✓	✓	0.840	0.890	0.153	0.123

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.

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, SiamPRN, 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 [66] and AlexNet [67]. 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 [23], SiamRPN++ [24] and Ocean [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, including CSA [68], RTAA [54], SPARK [7], FAN [6] and TTP [5], with respect to the untargeted attack result (Table XIV) and targeted attack

TABLE X

ATTACK RESULTS ON OTB2015, GOT-VAL AND LASOT WITH OR WITHOUT THE GROUNDDRUTH INFORMATION.

Benchmarks	Metrics	Untargeted Attack		Targeted Attack	
		w/ GT	w/o GT	w/ GT	w/o GT
OTB2015	Precision	0.063	0.056	0.759	0.752
	SR	0.092	0.080	0.795	0.794
	AO	0.123	0.121	0.890	0.893
GOT-Val	Precision	0.153	0.160	0.840	0.833
	Norm. Prec.	0.046	0.043	0.605	0.531
	AO	0.048	0.044	0.702	0.660
LaSOT	Precision	0.069	0.063	0.691	0.646
	Norm. Prec.	0.153	0.123	0.840	0.890
	AO	0.069	0.063	0.691	0.646

TABLE XI

INFLUENCE OF TWO DIFFERENT WAYS TO GENERATE THE FAKE TRAJECTORY: FOLLOWING A FIXED DIRECTION IN EACH VIDEO AND FOLLOWING THE REAL TRAJECTORY. IT IS EVALUATED ON GOT-VAL.

Type of the Fake Trajectory	Untargeted Attack		Targeted Attack	
	AO	SR	AO	SR
Fixed direction	0.175	0.144	0.845	0.897
Follow the real trajectory	0.153	0.123	0.840	0.890

result (Table XV) on OTB2015. SPARK [7] reduces the precision score with respect to the real trajectory from 0.851 to 0.064. 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. RTAA [54] reduces the precision score with respect to the real trajectory from 0.880 to 0.050. However, RTAA only performs the untargeted attacks for trackers, which is less challenging than the targeted attacks in this paper, as we aim to create arbitrary, complex trajectories at test time. CSA [68] 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, this method requires running a generative network for each frame to obtain adversarial information, making it difficult to meet the demand for real-time tracking. CSA reduces the precision score with respect to the real trajectory from 0.851 to 0.458. Recently, Liang *et al.* [6] propose a fast attack network (FAN) 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 features of regions specified by a particular trajectory. FAN can reduce the precision score with respect to the real trajectory from 0.720 to 0.180. However, the precision score with respect to the *fake trajectory* is only 0.420. Similar to CSA, FAN also requires running a generative network for each frame to obtain adversarial information. Nakka *et al.* [5] propose the temporally transferable perturbations (TTP) for attacking the SiamRPN++ tracker. TTP generates a single adversarial perturbation from the template image and adds this perturbation to each search image of the video. TTP

TABLE XII

TRANSFERABILITY TO DIFFERENT BACKBONES OF OUR PERTURBATIONS TRAINED ON SIAMFC++_GOOGLENET. IT IS EVALUATED ON GOT-VAL.

Backbones	Before Attack		Untargeted Attack		Targeted Attack	
	AO	SR	AO	SR	AO	SR
GoogLeNet	0.760	0.897	0.153	0.123	0.840	0.890
AlexNet	0.720	0.850	0.496	0.577	0.327	0.336
ShuffleNet	0.766	0.888	0.496	0.557	0.409	0.426

TABLE XIII

TRANSFERABILITY TO DIFFERENT TRACKING ARCHITECTURES OF OUR PERTURBATIONS TRAINED ON SIAMFC++_GOOGLENET. IT IS EVALUATED ON OTB2015.

Trackers	Before Attack		Untargeted Attack	
	AO	Precision	AO	Precision
SiamRPN++ [24]	0.676	0.879	0.418	0.556
SiamRPN [23]	0.666	0.876	0.483	0.643
Ocean [20]	0.672	0.902	0.237	0.282

reduces the precision score with respect to the real trajectory from 0.910 to 0.080, which is effective for untargeted attacks. However, the precision score with respect to the *fake trajectory* is only 0.692, which demonstrates that TTP's performance of targeted attacks is limited. In addition, this method needs to run a generative network for each video to obtain adversarial information, and thus requires the computational and storage resources of the real-world object tracking platform, which makes it difficult to deploy the method to resource-constrained platforms. Our method can reduce the precision score with respect to the real trajectory from 0.861 to 0.092. The precision score with respect to the *fake trajectory* is up to 0.795, which is significantly better than other methods.

We also compare our untargeted attack results on VOT2016 and VOT2018 with CSA [68], RTAA [54], FAN [6] and TTP [5] in Table XVI. 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

TABLE XIV

STATE-OF-THE-ART COMPARISON OF UNTARGETED ATTACK PERFORMANCE ON OTB2015 IN TERMS OF PRECISION SCORE.

Method	Tracker	Attack Cost per Frame(ms)	Before Attack	Untargeted Attack
RTAA	DaSiamRPN	-	0.880	0.050
SPARK	SiamRPN	41.4	0.851	0.064
CSA	SiamRPN	9	0.851	0.458
FAN	SiamFC	10	0.720	0.180
TTP	SiamRPN++	8	0.910	0.080
Ours	SiamFC++	~ 0	0.861	0.092

TABLE XV

STATE-OF-THE-ART COMPARISON OF TARGETED ATTACK PERFORMANCE ON OTB2015 IN TERMS OF PRECISION SCORE.

Method	Tracker	Targeted Attack
FAN	SiamFC	0.420
TTP	SiamRPN++	0.692
Ours	SiamFC++	0.795

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.*, pp. 1–1, 2020.
- [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] Q. Guo, 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.
- [8] X. Chen, X. Yan, F. Zheng, Y. Jiang, S.-T. Xia, Y. Zhao, and R. Ji, "One-shot adversarial attacks on visual tracking with dual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10176–10185.
- [9] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1765–1773.
- [10] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proc. Assoc. Adv. Artif. Intell.*, 2020, pp. 12549–12556.
- [11] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2411–2418.

TABLE XVI
STATE-OF-THE-ART COMPARISON OF UNTARGETED ATTACK PERFORMANCE ON VOT2016/2018 IN TERMS OF ACCURACY, ROBUSTNESS AND EXPECTED AVERAGE OVERLAP (EAO).

Dataset	Method	Tracker	Before Attack			Untargeted Attack		
			Accuracy	Robustness	EAO	Accuracy	Robustness	EAO
VOT2016	RTAA	DaSiamRPN	0.625	0.224	0.439	0.521	1.613	0.078
	Ours	SiamFC++	0.626	0.144	0.460	0.393	9.061	0.007
	RTAA	DaSiamRPN	0.585	0.272	0.380	0.536	1.447	0.097
	FAN	SiamFC	0.503	0.585	0.188	0.420	-	-
	CSA	SiamRPN	0.570	0.440	0.261	0.430	1.900	0.076
	TPP	SiamRPN++	0.600	0.320	0.340	0.520	7.820	0.014
	Ours	SiamFC++	0.587	0.183	0.426	0.342	8.981	0.007

- [12] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *arXiv preprint arXiv:1810.11981*, 2018.
- [13] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5374–5383.
- [14] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. P. Pflugfelder, L. Čehovin, T. Vojir, G. Häger, A. Lukezic, G. Fernández *et al.*, "The visual object tracking VOT2016 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 777–823.
- [15] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Čehovin Zajc, T. Vojir, G. Bhat, A. Lukezic, A. Eldesokey *et al.*, "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 0–0.
- [16] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Čehovin Zajc, O. Drbohlav, A. Lukezic, A. Berg *et al.*, "The seventh visual object tracking VOT2019 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 0–0.
- [17] 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. 11920–11929.
- [18] S. Liu, S. Wang, X. Liu, A. H. Gandomi, M. Daneshmand, K. Muhammad, and V. H. C. De Albuquerque, "Human memory update strategy: A multi-layer template update mechanism for remote visual monitoring," *IEEE Trans. Multimedia*, vol. 23, pp. 2188–2198, 2021.
- [19] S. Liu, S. Wang, X. Liu, 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.
- [20] 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.
- [21] Y. Cui, C. Jiang, L. Wang, and G. Wu, "Fully convolutional online tracking," *arXiv preprint arXiv:2004.07109*, 2020.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–26, 2021.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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. 3, pp. 1–41, 2020.
- [30] 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.
- [31] 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.*, vol. 34, no. 05, 2020, pp. 8018–8025.
- [32] 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.
- [33] 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.
- [34] 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.*, pp. 1–1, 2020.
- [35] 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.
- [36] 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.
- [37] 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.
- [38] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Security*, 2017, pp. 506–519.
- [39] J. Li, R. Ji, H. Liu, J. Liu, 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.
- [40] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie *et al.*, "Adversarial attacks and defences competition," *arXiv preprint arXiv:1804.00097*, 2018.
- [41] 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.
- [42] B. Li, C. Chen, W. Wang, and L. Carin, "Second-order adversarial attack and certifiable robustness," *arXiv preprint arXiv:1809.03113*, 2018.
- [43] 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.
- [44] V. Khrulkov and I. Oseledets, "Art of singular vectors and universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8562–8570.
- [45] 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.
- [46] 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. 14521–14530.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017.
- [51] D. Karmon, D. Zoran, and Y. Goldberg, “LaVAN: Localized and visible adversarial noise,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2507–2515.
- [52] 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.
- [53] 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.
- [54] S. Jia, C. Ma, Y. Song, and X. Yang, “Robust tracking against adversarial attacks,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 69–84.
- [55] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [56] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Proc. Int. Conf. Learn. Representation*, 2017.
- [57] A. Shafahi, M. Najibi, Z. Xu, J. Dickerson, L. S. Davis, and T. Goldstein, “Universal adversarial training,” in *Proc. Assoc. Adv. Artif. Intell.*, vol. 34, no. 04, 2020, pp. 5636–5643.
- [58] H. Hirano and K. Takemoto, “Simple iterative method for generating targeted universal adversarial perturbations,” *Algorithms*, vol. 13, no. 11, p. 268, 2020.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] 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.
- [64] 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.
- [65] 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.
- [66] 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.
- [67] 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.
- [68] 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.



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.



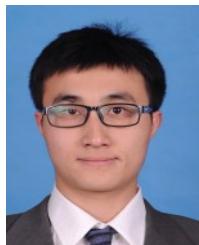
Bing Li received the Ph.D. degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, Beijing, China, in 2009. From 2009 to 2011, he worked as a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing. 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 the School of Information Engineering. 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.



Zhenbang 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.