Response to Reviewers of TCSVT-07024-2021: 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

Dear Editors:

We would like to express our heartfelt gratitude to you and the reviewers for the insightful and helpful comments. When we revised the paper, we carefully considered and followed all the comments and suggestions provided by you and the reviewers. To summarize, we have made the following revisions:

(1) We have added the advantages & limitations of the proposed method after experimental analysis and the future work to the conclusion.

(2) We have carefully considered and followed all the comments and suggestions related to the clarity of the writing, and made the new material a thoroughly revised manuscript.

(3) We have added 3 papers published in the IEEE Transactions on Circuits and Systems for Video Technology, which are most closely related to our manuscript, and analysed what is distinctive / new about our current manuscript related to these previously published papers.

Research topics related to adversarial attacks includes digital watermarking [1], 3D face presentation attacks [2] and adversarial defense [3]. Xiong et. al. [1] generate digital watermarking by slightly modifying the pixel values of videos frame to protect video

content from unauthorized access. Compared with [1], 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. [4] generate 3D face artifacts to fool the face recognition system. Compared with [4], our method directly modifies the input of the network instead of changing the input of cameras using 3D face artifacts. Wang et. al. [3] propose white-box attack/defense methods for image classifiers. Compared with [3], we focus on attacking the object tracking networks instead of the image classification networks.

Liu et. al. [5] propose a correlation filtering tracking algorithm based on the visual memory multi-template update strategy. Liu et. al. [6] construct a fuzzy aided detection module to boost the correlation filtering tracking system.

A comprehensive survey of the related trackers is beyond the scope of this paper, so we only briefly review two aspects that are most relevant to our work. Please refer to [7] for A thorough survey on visual object tracking methods.

We hope that our revised manuscript is now appropriate for publication in IEEE Transactions on Circuits and Systems for Video Technology. Specific responses to all the comments of each reviewer are included in the rest of this document and highlighted using bold font after the comments of each reviewer for the convenience of cross-reference. To make the changes easier to identify where necessary, we also have underlined most of the revised parts in the manuscript and provide an underlined version for the convenience of second review.

We are looking forward to your reply.

Yours sincerely,

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

Dr. Jin Gao (Contact author)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
Address: No. 95, Zhongguancun East Road, Haidian District,
Beijing 100190, P. R. China
Email: jin.gao@nlpr.ia.ac.cn

Response Letter to Reviewer #1

Dear Reviewer #1:

Thank you very much for your thorough review. Your insightful comments are very helpful for us to improve the quality of the paper. According to your comments and suggestions, we have carefully and extensively revised the manuscript. The main revised parts are highlighted by underlines in the underlined version for your convenience. You will find that all your comments and suggestions are considered and followed. We hope that our revised manuscript is now appropriate for publication in IEEE Transactions on Circuits and Systems for Video Technology. In addition, point-to-point responses to your comments are given below and highlighted using bold font in line with your comments in order to facilitate cross-referencing.

We are looking forward to your reply.

Yours sincerely,

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

Dr. Jin Gao (Contact author)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
Address: No. 95, Zhongguancun East Road, Haidian District,
Beijing 100190, P. R. China
Email: jin.gao@nlpr.ia.ac.cn

This paper addresses the task of attacking Siamese network-based trackers in a simple yet effective fashion. Unlike other methods that operate in the video-specific attacking regime (which resides on network inference for generating perturbations while tracking), this method is the first to perform universal targeted attacks for Siamese trackers utilizing both the translucent perturbation and the adversarial patch together. By adding the perturbation to the template and adding the patch to the search image while performing tracking, this work fools the Siamese trackers to the fake target region and thus makes them fail in tracking the real target object. Overall, this is an interesting paper, and it is well written and organized. As it is a resubmitted manuscript, I notice that the authors have made substantial changes to the previous manuscript, which are able to appropriately respond to the comments made by the previous reviewers. Although the template perturbation and adversarial patch are both easy to observe for human eyes as the previous reviewers have pointed out, and the SSIMs for them are also lower than the video-specific attacking method, e.g., FAN [8], this reviewer believes that this proposed new framework can be a new configuration of adversarial attack on visual tracking for its achieved balance between the attack efficiency and the perturbation perceptibility. This new configuration will attract increasing attention from the visual tracking attack community to study on more efficient attack methods.

Many thanks for your positive comments on the strength of our paper and the novelty of the proposed attack method.

In addition, I suggest the authors add more experiments to demonstrate the practicability of the attack method when the ground truth box information is missing in the training data. The experimental results show that it is effective to use the predicted boxes instead of ground truth boxes for training perturbations.

This question needs to be discussed with Jin Gao.

A small question is that it will be better if the authors can provide some pseudo code

Algorithm 1 Attack Process

Input: Imperceptible perturbation δ, adversarial patch 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: $B^{pred} = \{b_i^{pred}\}_1^T$

Generate the clean template image $\mathbf{z}_1$ according to I_1 and b_1^{gt}.
Generate the perturbed template image $\tilde{\mathbf{z}}_1 = \mathbf{z}_1 + \delta$.
Let $i = 2$.
while $i \leq T$ **do**
    Generate clean search image $\mathbf{x}_i$ according to I_i and b_{i-1}^{pred}.
    $b_i^{fake} = \{x_{0_i}, y_{0_i}, x_{1_i}, y_{1_i}\}$
    Generate the perturbed searach image $\tilde{\mathbf{x}} = A_{\text{add}}(\mathbf{x}, p_k, \{x_0, y_0, x_1, y_1\})$.
    C, R, Q $= f(\tilde{\mathbf{x}}_i, \tilde{\mathbf{z}}_1)$.
    Generate the predicted bounding box b_t^{pred} according to **C, R, Q**.
    $i = i + 1$.
end while
return δ_N, p_N.

for the untargeted attack and targeted attack processes in addition to the training process. This will facilitate the understanding of the attacking process while performing tracking.

Both the targeted attack and untargeted attack process follow the same steps as shown in Alg. 1. The difference between the target attack and the untargeted attack is the evaluation. For the targeted attack, we compare the AO between $B^{pred} = \{b_i^{pred}\}_1^T$ and $B^{fake} = \{b_i^{fake}\}_1^T$. For the untargeted attack, we compare the AO between $B^{pred} = \{b_i^{pred}\}_1^T$ and $B^{gt} = \{b_i^{gt}\}_1^T$.

In addition, the font size in Fig. 9 is too small to read on my computer, which needs to be improved.

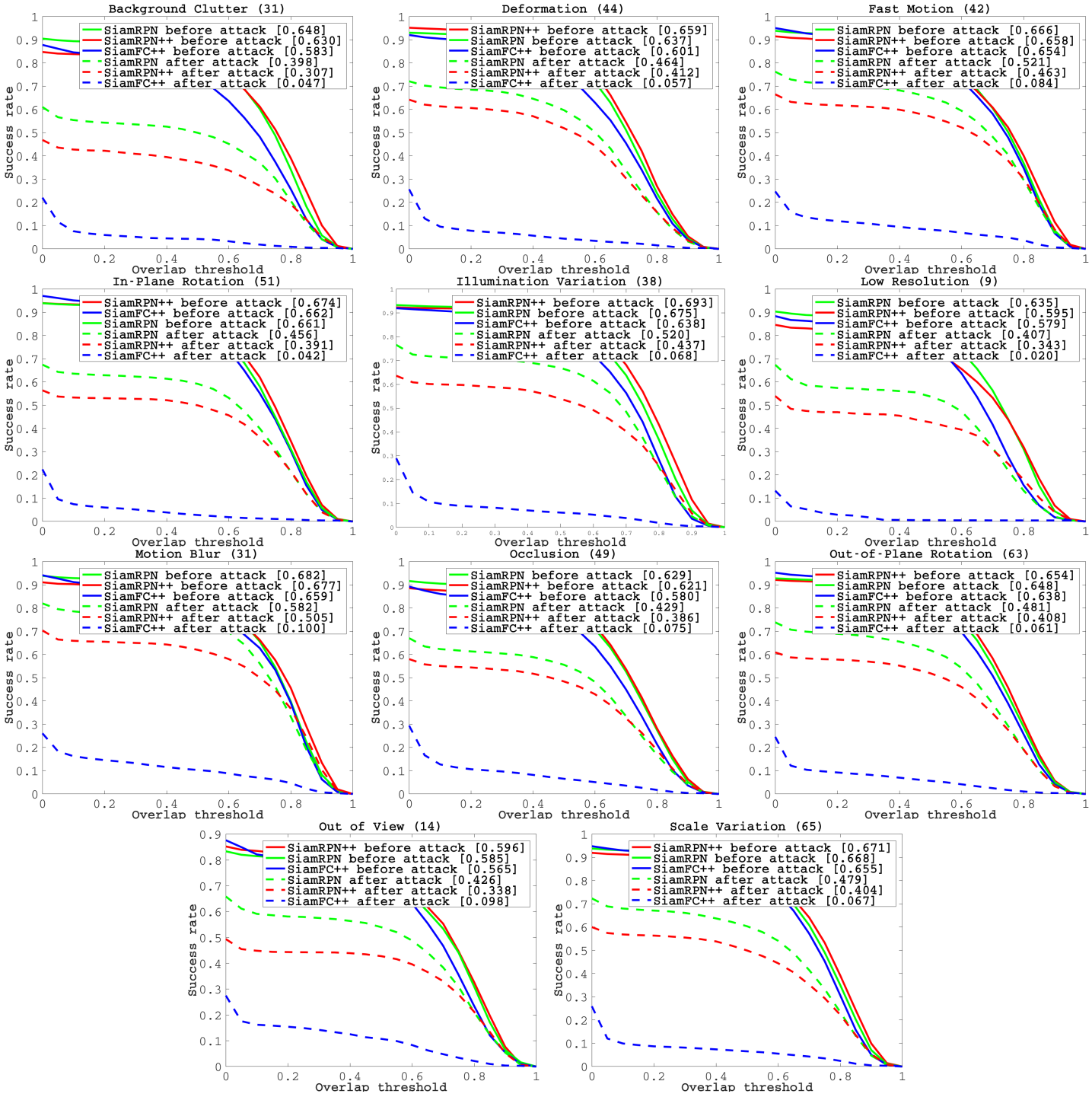

Figure 9: Untargeted attack results of the different trackers under the 11 attributes: out-of-view (OV), occlusion (OCC), illumination variation (IV), out-of-plane rotation (OPR), scale variation (SV),deformation (DEF), low resolution (LR), fast motion (FM), background clutters (BC), motion blur (MB), and in-plane rotation (IPR). The results show good transferability of our attacks to different tracking architectures, even if the generated perturbations are applied to anchor-based trackers.

Also, some recent papers are valued to be referred (2020-2021), to ehahnce the quality.

...

Research topics related to adversarial attacks includes digital watermarking [1], 3D face presentation attacks [2] and adversarial defense [3]. Xiong et. al. [1] generate digital watermarking by slightly modifying the pixel values of videos frame to protect video content from unauthorized access. Compared with [1], 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. [4] generate 3D face artifacts to fool the face recognition system. Compared with [4], our method directly modifies the input of the network instead of changing the input of cameras using 3D face artifacts. Wang et. al. [3] propose white-box attack/defense methods for image classifiers. Compared with [3], we focus on attacking the object tracking networks instead of the image classification networks.

Response Letter to Reviewer #2

Dear Reviewer #2:

Thank you very much for your thorough review. Your insightful comments are very helpful for us to improve the quality of the paper. According to your comments and suggestions, we have carefully and extensively revised the manuscript. The main revised parts are highlighted by underlines in the underlined version for your convenience. You will find that all your comments and suggestions are considered and followed. We hope that our revised manuscript is now appropriate for publication in IEEE Transactions on Circuits and Systems for Video Technology. In addition, point-to-point responses to your comments are given below and highlighted using bold font in line with your comments in order to facilitate cross-referencing.

We are looking forward to your reply.

Yours sincerely,
Zhenbang Li, Yaya Shi, Jin Gao, Shaoru Wang, Bing Li, Pengpeng Liang, Weiming Hu

Dr. Jin Gao (Contact author)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
Address: No. 95, Zhongguancun East Road, Haidian District,
Beijing 100190, P. R. China
Email: jin.gao@nlpr.ia.ac.cn

In this paper, the authors train a universal adversarial patch to add on both template and search regions of a Siamese based tracker to deteriorate its original performance. The proposed perturbations are video-agnostic, leading to a low computational cost during attack. The experiment validations show that the proposed method achieves favorable attack results on OTB2015, GOT-10k, LaSOT, UAV123, VOT2016, VOT2018 and VOT2019. In addition, the generated perturbations transfer well on other Siamese trackers as well. The idea of this paper is interesting and the experiments are thorough.

Many thanks for your positive comments on the strength of our paper and the novelty of the proposed attack method.

However, there are some concerns over the implementation, performance and writing. 1. The authors state that training with Ep. 4 leads to an obvious patch on the images while using Eq.5 into the training process results in a less obvious patch. The reviewer considers that giving a constraint (e.g. l_∞) on the p_x in Eq.4 can make the perturbation imperceptible intuitively. Please give more analysis on this setting. Besides, the reviewer hopes to know the reason why give an extra perturbation on the template region. The perturbations on template and search regions look similar, while the authors say that they are different. Please state the difference between the patch application operator on search examples and the operator on template examples. In addition, the denotations of A_{paste} in Eq.4 and A_{add} in Eq.5 seem like the same one.

Thanks for the good comment. Before analysis if giving constraint (e.g. L_∞) on the p_x in Eq.4 is really lead to imperceptible or not, let us first clarify the concept of the A_{paste} and A_{add}, which is not tell good in the previous manuscript.

A_{paste} means that, in the region where the perturbation is pasted, the pixel values of the original image is replaced with the pixel values of the perturbation. Thus, even if we add constraints to the patch so that the perturbation values are small, when the patch is pasted to the image, it will look like an almost black region, which is obvious rather than imperceptible.

The operator A_{add} means to apply additive noise, where the injecting small perturbations into the image. Often need constraint. the injected noise are small and sometimes invisible. If we add a patch onto the image, the final pixel value is the original value plus a new value.

The operator $A_{\mathbf{add}}(\mathbf{x}, p_{\mathbf{x}}, b_{\mathbf{x}}^{fake})$ and $\mathbf{z} + \delta_{\mathbf{z}}$ is similar. The only difference is that, $A_{\mathbf{add}}$ means add the perturbation to a specific region because the patch size is less than the image, while $+$ means add the perturbation to the full image because the size is the same.

For your convenience of cross checking, the new added text is shown as follows:

A_{paste} means that, in the region where the perturbation is pasted, the pixel values of the original image is replaced with the pixel values of the perturbation. The operator

Table 1: rewrq

	Untargeted Attack		Targeted Attack	
	AO	SR	AO	SR
track2	0.1747	0.1442	0.8448	0.8972

A_{add} means to apply additive noise, where the injecting small perturbations into the input samples. Often need constraint. the injected noise are small and sometimes invisible.

2. I agree with reviewer 3, the ground truth boxes are inaccessible to trackers during the inference. It seems that the authors use the ground truth boxes to generate the fake trajectory in lines 51-57 on page 7. Please clarify it.

We generate new trajectory without the test GT: we let the fake box run straight to border of the image, speed and derection is random. For your convenience, the new text is added as following:

Let us now turn to targeted attacks. In this context, we consider two scenarios: 1. The attacker forces the tracker to follow a fixed direction. We illustrate this with 4 different directions, consisting of shifting the box by (3, 3) pixels in each consecutive frames, corresponding to the four directions 45, -45, 135, -135. 2. The attacker seeks for the tracker to follow a more complicated trajectory. ...

3. For the experiments, the authors should conduct the ablation study on only adding perturbations on the template images or the search regions to show the impact of p and δ.

Thanks for your advice and we have conducted the ablation study on only adding perturbations on the template images or the search regions to show the impact of p and δ. As we can see, only use one of them is not useful, because they are trained together. For your convenience of cross review, the new text is given as follows.

To see the impact of p and δ, we add the ablation study on only adding perturbations on the template images or the search regions. The result is shown in table 2. Experiment results shows that only train in one branch is not useful. For the untargeted attack, the only template result is 0.50, the only search result is 0.71. While both attack is 0.**. For the targeted attack, the only template result is 0.16, the only search attack is 0.16, while both attack is 0.**.

4. As reviewer 1 and reviewer 2 say, the perturbations added to template regions and research regions are not imperceptible, which may be helpful to misguide the tracker. The reviewer considers that adding a similar random pattern on the template and search regions to further illustrate the effectiveness of the proposed method.

To illustrate the effectiveness of the proposed method, we add a random pattern on the template and search regions: the mean value and standard deviation

Table 2: q2eq23

	Untargeted Attack		Targeted Attack	
	AO	SR	AO	SR
Only Template	0.5097	0.5669	0.1555	0.1064
Only Search	0.7137	0.8414	0.1599	0.1320
Both				

is the same as the template/search image. Specifically, the mean value of the template perturbation is *, the standard deviation of the template image is ***. The mean value of the search patch is ***, the standard deviation of the search image is ***. We create the gaussian noise. 'FGT-AO', 'FGT-SR-50', 'GT-AO', 'GT-SR-50' 0.14250828600526147, 0.10118303356737521, 0.7604903420033434, 0.8957391555256324 mean z=0.05680246651172638, std z=6.4726996421813965, mean x=-0.12454431504011154, std x=9.011011123657227**

5. There are some minor problems, grammar errors and typos in this paper. The reviewer hopes the authors polish this paper again.

- On page 2, '1016' → '2016' in line 56. There is a same one in line 47 on page 6.

- The denotation of B_x^{fake} in Eq.4 is not clear enough, even though it can be inferred by the later part.

- On page 4, 'imperceptible' → 'imperceptibly' in line 57.

- The reinitialization of VOT-toolkit should be mentioned in the part of 'experimental setup'.

- On page 7, '.(see Table I)' is a typo.

Thanks for the good comment. We have carefully checked through the whole text and corrected the grammar mistakes and typos. Specifically, We have replaced "Experiment results on OTB2015 [9], GOT-10k [10], LaSOT [10], UAV123 [11], VOT1016 [12], VOT2018 [13] and VOT2019 [14]. benchmarks demonstrate the effectiveness and efficiency of our approach." with "Experiment results on OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT2016 [12], VOT2018 [13] and VOT2019 [14]. benchmarks demonstrate the effectiveness and efficiency of our approach." **in Section I of the revised manuscript.**

We have replaced "In SiamFC++, the tracker first transforms the paired reference frame I_1 and annotation b_1^{gt} to get an 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." with "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." **in Section III.A of the revised manuscript.**

We have replaced "However, CNN attacks are usually expected imperceptible but the above method has to paste an obviously noticeable fake target patch to tracking frames, which raises the risk of being suspected." with "However, CNN attacks are usually expected imperceptibly but the above method has to paste an obviously noticeable fake target patch to tracking frames, which raises the risk of being suspected." **in Section III.A of the revised manuscript.**

We have replaced "A_add **adds the patch into the search image x at location** $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$**." with** "$A_{add}$ adds the patch into the search image **x** at location $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$." **in Section III.B of the revised manuscript.**

We have replaced "We evaluate our video-agnostic perturbations for targeted attacks on several tracking benchmarks, i.e., OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT1016 [12], VOT2018 [13] and VOT2019 [14]." with "We evaluate our video-agnostic perturbations for targeted attacks on several tracking benchmarks, i.e., OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT2016 [12], VOT2018 [13] and VOT2019 [14]." **in Section IV.A of the revised manuscript.**

We have replaced "We adopt COCO [16], ILSVRC-VID [17] and the training splits of GOT-10k [10] and LaSOT [15] as our training set. (see Table ??)" with "We adopt COCO [16], ILSVRC-VID [17] and the training splits of GOT-10k [10] and LaSOT [15] as our training set." **in Section IV.B of the revised manuscript.**

Response Letter to Reviewer #3

Dear Reviewer #3:

Thank you very much for your thorough review. Your insightful comments are very helpful for us to improve the quality of the paper. According to your comments and suggestions, we have carefully and extensively revised the manuscript. The main revised parts are highlighted by underlines in the underlined version for your convenience. You will find that all your comments and suggestions are considered and followed. We hope that our revised manuscript is now appropriate for publication in IEEE Transactions on Circuits and Systems for Video Technology. In addition, point-to-point responses to your comments are given below and highlighted using bold font in line with your comments in order to facilitate cross-referencing.

We are looking forward to your reply.

Yours sincerely,
Zhenbang Li, Yaya Shi, Jin Gao, Shaoru Wang, Bing Li, Pengpeng Liang, Weiming Hu

Dr. Jin Gao (Contact author)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
Address: No. 95, Zhongguancun East Road, Haidian District,
Beijing 100190, P. R. China
Email: jin.gao@nlpr.ia.ac.cn

(a) SSIM=0.85 (b) SSIM=0.56 (c) SSIM=0.79 (d) SSIM=0.79

Figure 2: YCbCr Attack.

This paper employs the universal perturbation attacks on Siamese visual trackers. There are still the following concerns about the proposed method. 1. As stated in the paper, the proposed method " does not require gradient optimization or network inference". However, this is a double-edged sword since it resulted in suspicious attacks. Prior works commonly train a network to prevent not only suspicious attacks but also modifying every pixel. I think it's a major problem with this work. I suggest considering a proper strategy to remove/reduce it.

Thanks for your advices. Our patch is obvious (SSIM=0.56). We think the main reason is that we have to learn useful information in such small regions 64*64 as a fake target. While the information of the real target is not changed. So it may be top heat map on the real object, so the network will work hard to create a fake target, which lead to high perturbation values. But, if we can break the target information, it will be easy to attack. Our solution is attack the whole image. Y channel, CbCr channel. As we can see, the SSIM of the search image is changed from 0.56 to 0.85. The SSIM of the template image is changed from 0.79 to 0.79.

2. The proposed method and offline training phase should be explained more clearly. For instance, the termination of offline training or offline optimization is missed affecting perturbation values.

Thanks for the good comment. For your convenience in cross-checking, the new text is given as follows.

Before introducing the proposed method, we first revisit the popular adversarial example generation methods. One of the simplest methods to generate adversarial images I^{adv} works by linearizing loss function in L_∞ neighbourhood of a clean image and finds exact maximum of linearized function using following closed-form equation [18]:

$$I^{adv} = I + \epsilon \, sign\big(\nabla_I L(I, y_{true})\big), \tag{1}$$

where I is the input image, and the values of the pixels are integer numbers in the range [0, 255]. y_{true} is the true label for the image I. $L(I, y)$ is the cost function of the neural network for the attack purpose, given image I and label y. 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, and clipping pixel values of intermediate results after each step to ensure that they are in an ϵ-neighbourhood of the original image. This leads to the Basic Iterative Method (BIM) introduced in [**?**]:*

$$\begin{aligned} I_0^{adv} &= I, \\ I_{N+1}^{adv} &= Clip_{I,\epsilon}\{I_N^{adv} + \alpha\, sign(\nabla_I L(I_N^{adv}, y_{true}))\}, \end{aligned} \tag{2}$$

where $Clip_{I,\epsilon}\{I'\}$ is the function which performs per-pixel clipping of the image I', so that the result will be in L_∞ ϵ-neighbourhood of the source image I.

3. The descriptions of figures and tables are not self-explanatory.

...

4. I suggest adding the advantages & limitations of the proposed method after experimental analysis and future works to the conclusion.

Compared with other attack method, our method has two key advantages. First, because our video-agnostic feature, we are fast, no need to network interferes or gradient calculation, 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 the perturbation values are not as small as ohter non-universal perturbations due to the task difficulty. In future work, we expect that it will be possible to demonstrate the existence of a single perturbation to attack multi vision tasks, including image classification, object detection and video object tracking/segementation.

5. Some of the experiments require more explanations. For example, transferability has been investigated by different backbones & architectures. It's needed to mention these experiments are with/without the training phase or not.

Thanks for point out this. These experiments do not need the training phase.

6. In experiments, I suggest considering two scenarios of different directions and trajectories.

We generate new trajectory without the test GT: we let the fake box run straight to border of the image, speed and derection is random.

7. There are still some typo and grammar mistakes in the paper.

Thanks for the good comment. We have carefully checked through the whole text and corrected the grammar mistakes and typos. Specifically, We have replaced "Experiment results on OTB2015 [9], GOT-10k [10], LaSOT [10], UAV123 [11], VOT1016 [12], VOT2018 [13] and VOT2019 [14]. benchmarks demonstrate the effectiveness and efficiency of our approach." with "Experiment results on OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT2016 [12], VOT2018 [13] and VOT2019 [14].

benchmarks demonstrate the effectiveness and efficiency of our approach." **in Section I of the revised manuscript.**

We have replaced "In SiamFC++, the tracker first transforms the paired reference frame I_1 and annotation b_1^{gt} to get an 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." with "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." **in Section III.A of the revised manuscript.**

We have replaced "However, CNN attacks are usually expected imperceptible but the above method has to paste an obviously noticeable fake target patch to tracking frames, which raises the risk of being suspected." with "However, CNN attacks are usually expected imperceptibly but the above method has to paste an obviously noticeable fake target patch to tracking frames, which raises the risk of being suspected." **in Section III.A of the revised manuscript.**

We have replaced "A_add adds the patch into the search image x at location $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$." with "$A_{add}$ adds the patch into the search image $\mathbf{x}$ at location $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$." **in Section III.B of the revised manuscript.**

We have replaced "We evaluate our video-agnostic perturbations for targeted attacks on several tracking benchmarks, i.e., OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT1016 [12], VOT2018 [13] and VOT2019 [14]." with "We evaluate our video-agnostic perturbations for targeted attacks on several tracking benchmarks, i.e., OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT2016 [12], VOT2018 [13] and VOT2019 [14]." **in Section IV.A of the revised manuscript.**

We have replaced "We adopt COCO [16], ILSVRC-VID [17] and the training splits of GOT-10k [10] and LaSOT [15] as our training set. (see Table ??)" with "We adopt COCO [16], ILSVRC-VID [17] and the training splits of GOT-10k [10] and LaSOT [15] as our training set." **in Section IV.B of the revised manuscript.**

8. Missing key ref : "Deep Learning for Visual Tracking: A Comprehensive Survey," in IEEE Transactions on Intelligent Transportation Systems, 2021.

Thanks for the good comment.

A comprehensive survey of the related trackers is beyond the scope of this paper, so we only briefly review two aspects that are most relevant to our work. Please refer to [7] for A thorough survey on visual object tracking methods.

Response Letter to Reviewer #4

Dear Reviewer #4:

Thank you very much for your thorough review. Your insightful comments are very helpful for us to improve the quality of the paper. According to your comments and suggestions, we have carefully and extensively revised the manuscript. The main revised parts are highlighted by underlines in the underlined version for your convenience. You will find that all your comments and suggestions are considered and followed. We hope that our revised manuscript is now appropriate for publication in IEEE Transactions on Circuits and Systems for Video Technology. In addition, point-to-point responses to your comments are given below and highlighted using bold font in line with your comments in order to facilitate cross-referencing.

We are looking forward to your reply.

Yours sincerely,
Zhenbang Li, Yaya Shi, Jin Gao, Shaoru Wang, Bing Li, Pengpeng Liang, Weiming Hu

Dr. Jin Gao (Contact author)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
Address: No. 95, Zhongguancun East Road, Haidian District,
Beijing 100190, P. R. China
Email: jin.gao@nlpr.ia.ac.cn

This paper proposes a universal targeted attacks method on Siamese visual tracking task. It seems that the method is feasible and somewhat novel.

Many thanks for your positive comments on the strength of our paper and the novelty of the proposed attack method.

My major concern is that the writing and organization need to be carefully modified and optimized. Besides, 1. Authors are focused on the anchor-free tracker in the experiments, but the anchor-free Siamese trackers are not well mentioned. I suggest the authors to include the discussion of state-of-the-art of other similar approaches (e.g., FCOT, SiamCAR, OCEAN, et. al.).

Thanks for your advice and we have test the transferability to the other anchor free tracker OCEAN.

2. The model is trained based on Eq.11 and Eq.12. It is not clear why the sign function is introduced. It is also suggested to describe the derivation of these two equations in detail.

Thanks for the good comment. For your convenience in cross-checking, the new text is given as follows.

Before introducing the proposed method, we first revisit the popular adversarial example generation methods. One of the simplest methods to generate adversarial images I^{adv} works by linearizing loss function in L_∞ neighbourhood of a clean image and finds exact maximum of linearized function using following closed-form equation [18]:

$$I^{adv} = I + \epsilon \, sign\big(\nabla_I L(I, y_{true})\big), \tag{3}$$

*where I is the input image, and the values of the pixels are integer numbers in the range [0, 255]. y_{true} is the true label for the image I. $L(I, y)$ is the cost function of the neural network for the attack purpose, given image I and label y. 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, and clipping pixel values of intermediate results after each step to ensure that they are in an ϵ-neighbourhood of the original image. This leads to the Basic Iterative Method (BIM) introduced in [**?**]:*

$$\begin{gathered} I_0^{adv} = I, \\ I_{N+1}^{adv} = Clip_{I,\epsilon}\{I_N^{adv} + \alpha \, sign(\nabla_I L(I_N^{adv}, y_{true}))\}, \end{gathered} \tag{4}$$

where $Clip_{I,\epsilon}\{I'\}$ is the function which performs per-pixel clipping of the image I', so that the result will be in L_∞ ϵ-neighbourhood of the source image I.

3. The format of all the Tables is suggested to be unified.

Thanks for the good comment. We have unified the format of all the tables.

4. There are many typos in the paper, including but not limited to the following errors:

(a) In the third line below Eq.7, A_add should be A_{add}.

(b) In page 2, difference datasets GOT-10K[12], LaSOT [15] cite the same reference.

(c) In page 3, Subsection A of Section 3, "to get an template image" should be "to get a template image".

Thanks for the good comment. We have carefully checked through the whole text and corrected the grammar mistakes and typos. Specifically, We have replaced "Experiment results on OTB2015 [9], GOT-10k [10], LaSOT [10], UAV123 [11], VOT1016 [12], VOT2018 [13] and VOT2019 [14]. benchmarks demonstrate the effectiveness and efficiency of our approach." with "Experiment results on OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT2016 [12], VOT2018 [13] and VOT2019 [14]. benchmarks demonstrate the effectiveness and efficiency of our approach." **in Section I of the revised manuscript.**

We have replaced "In SiamFC++, the tracker first transforms the paired reference frame I_1 and annotation b_1^{gt} to get an 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." with "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." **in Section III.A of the revised manuscript.**

We have replaced "However, CNN attacks are usually expected imperceptible but the above method has to paste an obviously noticeable fake target patch to tracking frames, which raises the risk of being suspected." with "However, CNN attacks are usually expected imperceptibly but the above method has to paste an obviously noticeable fake target patch to tracking frames, which raises the risk of being suspected." **in Section III.A of the revised manuscript.**

We have replaced "A_add adds the patch into the search image x at location $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$." with "$A_{add}$ adds the patch into the search image $\mathbf{x}$ at location $(\frac{x_0+x_1}{2}, \frac{y_0+y_1}{2})$." **in Section III.B of the revised manuscript.**

We have replaced "We evaluate our video-agnostic perturbations for targeted attacks on several tracking benchmarks, i.e., OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT1016 [12], VOT2018 [13] and VOT2019 [14]." with "We evaluate our video-agnostic perturbations for targeted attacks on several tracking benchmarks, i.e., OTB2015 [9], GOT-10k [10], LaSOT [15], UAV123 [11], VOT2016 [12], VOT2018 [13] and VOT2019 [14]." **in Section IV.A of the revised manuscript.**

We have replaced "We adopt COCO [16], ILSVRC-VID [17] and the training splits of GOT-10k [10] and LaSOT [15] as our training set. (see Table ??)" with "We adopt COCO [16], ILSVRC-VID [17] and the training splits of GOT-10k [10] and LaSOT [15] as our training set." **in Section IV.B of the revised manuscript.**

References

[1] 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.

[2] S. R. Arashloo, "Unseen face presentation attack detection using sparse multiple kernel fisher null-space," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 445–461.

[12] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. P. Pflugfelder, L. Cehovin, T. Vojír, 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.

[13] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. ˇCehovin 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.

[14] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. ˇCehovin 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.

[15] 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.

[16] 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.

[17] 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.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.