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 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 relavant to this paper includes digital watermarking [1], 3D face presentation attacks [2] and adversarial defense [3]. A digital watermark is a kind of marker covertly embedded in image data. Both digital watermarking methods [1] and adversarial attack methods (slightly) modify the pixel values of images. However, they share different goals. The purpose of digital watermarking is to determine whether the video is a genuine video by detecting the presence of scrambling and thus copyright protection, and we reduce the tracking effect by adding scrambling. 3D face presentation attacks [2] aim to fool the face recognition system, where an imposter tries to be authenticated as a genuine client by presenting fake biometric traits to system sensors. Both the face presentation attack methods and our attack method all aim to fool the computer vision algorithms. Most 3D face presentation attack methods do not directly modify pixels values, the way of generating face artifacts includs wearing wearable facial masks [3], building 3D facial models [4], and through facial makeup to change the The adversary is able to place face artifacts in the physical environment seen by the camera, thereby indirectly changing the algorithm input to perform the attack. While in this paper, the adversary has direct access to the actual data fed into the model. In other words, the adversary can choose specific float32 values as input for the model. Defenses against adversarial attacks are aimed to build a robust classifier so that it correctly identifies adversarial images. For example, Wang et. al. [3] builds a robust classification system that can be viewed as a structural black box. After adding a buffer to the classification system, attackers can be efficiently deceived. The real evaluation results of the generated adversarial examples are often contrary to what the attacker thinks. However, this paper focus on the image classification task, while we focus on attacking the visual object tracking task, which is more difficult.

[4] proposes a correlation filtering algorithm based on the visual memory multi-template update strategy.

[5] constructs a fuzzy aided detection module to boost the tracking system.

for detail information about tracking, please refer to [6].

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 [7], 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 thenovelty 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 z_1 according to I_1 and b_1^{gt}.
Generate the perturbed template image z̃_1 = z_1 + δ.
Let i = 2.
while i ≤ T do
    Generate clean search image 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 x̃ = A_add(x, p_k, {x_0, y_0, x_1, y_1}).
    C, R, Q = f(x̃_i, z̃_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.

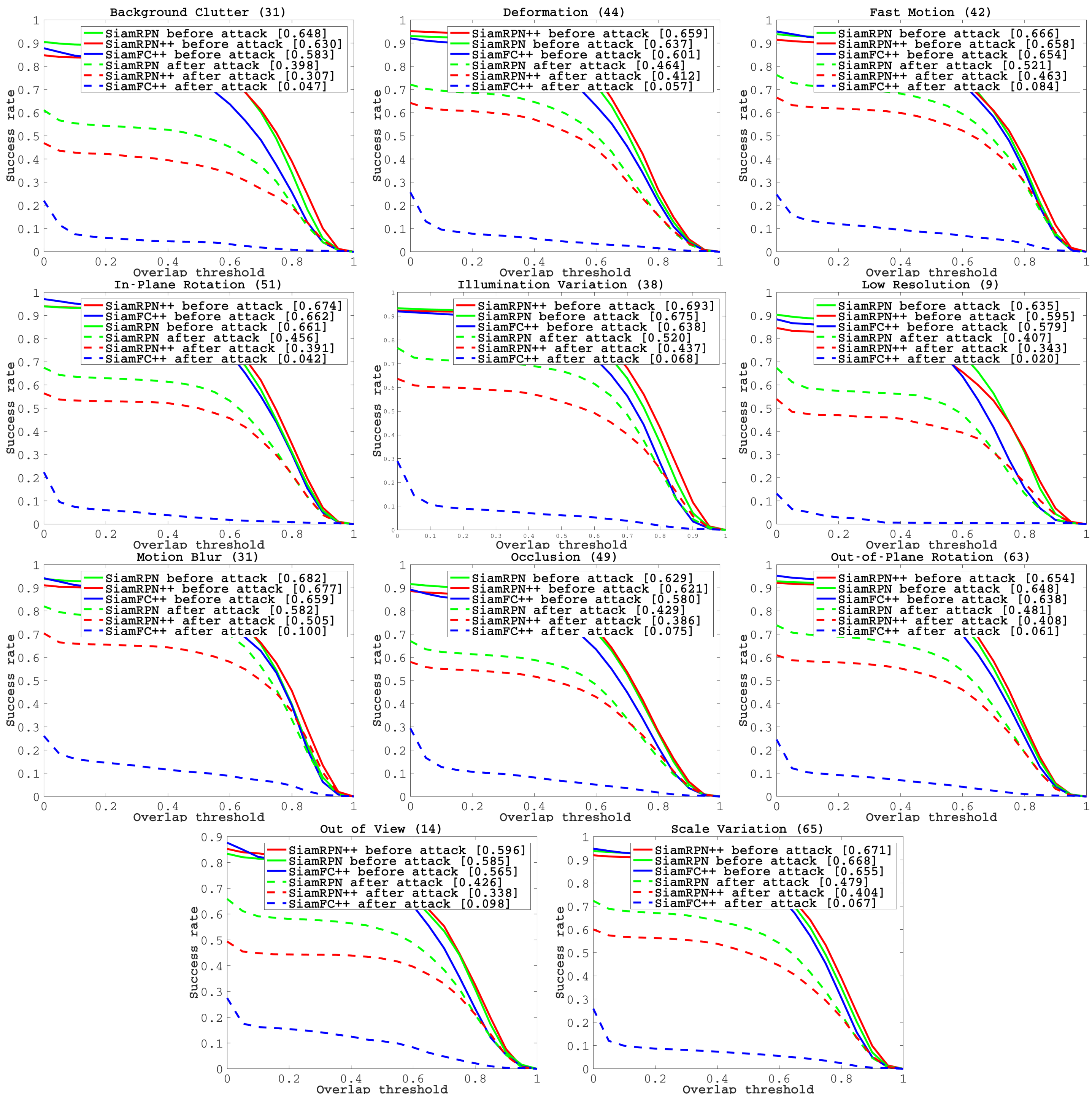

Figure 9: Untargeted attack results of the different trackers under the 11 attributes: out-of-view (OV), occlusion (OCC), illumination variation (IV), out-of-plane rotation (OPR), scale variation (SV),deformation (DEF), low resolution (LR), fast motion (FM), background clutters (BC), motion blur (MB), 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.
...

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_inf) on the p_x in Eq.4 can make the perturbation imperceptible intuitively. Please give more analysis on this setting.

Sorry that we did not clarify the meaning of paste. If we paste the perturbation, we means we use a patch to replace the original image, which means the patch and the original image is discrete (even if we give a constraint).

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.

The operator $A_{\textbf{add}}(\mathbf{x}, p_{\mathbf{x}}, b_{\mathbf{x}}^{fake})$ and $\mathbf{z} + \delta_{\mathbf{z}})$ is similar. The only difference is that, $A_{\textbf{add}}$ means add the perterbation 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.

In addition, the denotations of A_paste in Eq.4 and A_add in Eq.5 seem like the same one.

Sorry that we make the wrong citation and not clarify the meaning of add and paste. If we paste the perturbation, we means we use a patch to replace the original image, which means the patch and the original image is discrete (even if we give a constraint). The final pixel value is changed from the original value to a new value. If we add a patch onto the image, the final pixel value is the original value plus a new value.

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.

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

	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

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.

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 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 your advice and we have changed this errors.

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.

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 [8]:

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

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 [?]:

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

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.

...

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

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.

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 [8]:

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

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 [?]:

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

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.

...

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 [9] 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".

...

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. 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.