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, Stephen J. Maybank

Dear Editors:

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 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, Stephen J. Maybank

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, Stephen J. Maybank

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 [1], 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 Targeted Attack Process

Input: Imperceptible perturbation δ, and adversarial patch p, and Siamese tracker f.
Output: 123

Let $k = 0$.
while $k < N$ **do**
  Randomly pick a video $V \in \mathcal{V}$. The corresponding ground truth is $B^{gt} = \{b_i^{gt}\}_1^T$.
  Randomly pick paired frames I_t, I_s from V.
  Generate template image $\mathbf{z}$ according to I_t and b_t^{gt}.
  $\tilde{\mathbf{z}} = \mathbf{z} + \delta_k$.
  Generate search image $\mathbf{x}$ according to I_s and b_s^{gt}.
  Calculate the *fake target* position $\{x_0, y_0, x_1, y_1\}$ with respect to the search image.
  $\tilde{\mathbf{x}} = A_{\text{add}}(\mathbf{x}, p_k, \{x_0, y_0, x_1, y_1\})$.
  C, R, Q $= f(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$.
  Generate fake labels $\mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*$ using $\{x_0, y_0, x_1, y_1\}$.
  Calculate loss $L(\mathbf{C, R, Q}, \mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*)$ using Equ. **??**.
  $\delta_{k+1} = \delta_k - \epsilon_1 \cdot \text{sign}(\nabla_{\delta_k} L)$.
  $p_{k+1} = p_k - \epsilon_2 \cdot \text{sign}(\nabla_{p_k} L)$.
  $k = k + 1$.
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.

...

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

...

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

...

Algorithm 2 Untargeted Attack Process

Input: Imperceptible perturbation δ, and adversarial patch p, and Siamese tracker f.
Output: 123

Let $k = 0$.
while $k < N$ **do**
  Randomly pick a video $V \in \mathcal{V}$. The corresponding ground truth is $B^{gt} = \{b_i^{gt}\}_1^T$.
  Randomly pick paired frames I_t, I_s from V.
  Generate template image $\mathbf{z}$ according to I_t and b_t^{gt}.
  $\tilde{\mathbf{z}} = \mathbf{z} + \delta_k$.
  Generate search image $\mathbf{x}$ according to I_s and b_s^{gt}.
  Calculate the *fake target* position $\{x_0, y_0, x_1, y_1\}$ with respect to the search image.
  $\tilde{\mathbf{x}} = A_{\text{add}}(\mathbf{x}, p_k, \{x_0, y_0, x_1, y_1\})$.
  C, R, Q $= f(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$.
  Generate fake labels $\mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*$ using $\{x_0, y_0, x_1, y_1\}$.
  Calculate loss $L(\mathbf{C, R, Q}, \mathbf{C}^*, \mathbf{R}^*, \mathbf{Q}^*)$ using Equ. **??**.
  $\delta_{k+1} = \delta_k - \epsilon_1 \cdot \text{sign}(\nabla_{\delta_k} L)$.
  $p_{k+1} = p_k - \epsilon_2 \cdot \text{sign}(\nabla_{p_k} L)$.
  $k = k + 1$.
end while
return δ_N, p_N.

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, Stephen J. Maybank

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.

...

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.

...

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.

...

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.

...

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.

...

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.

...

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, Stephen J. Maybank

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 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.

...

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.

...

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.

...

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.

...

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

...

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, Stephen J. Maybank

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.

...

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.).

...

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.

...

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 [2] 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] 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.

[2] 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.