IEEE Transactions on Circuits and Systems for Video Technology

Decision Letter (TCSVT-07024-2021)

From: d.noel@ieee.org

To: jin.gao@nlpr.ia.ac.cn

CC: tcsvt.eic@gmail.com, fengwu@ustc.edu.cn, zhenbang.li@nlpr.ia.ac.cn, shiyaya@mail.ustc.edu.cn, jin.gao@nlpr.ia.ac.cn, shaoru.wang@nlpr.ia.ac.cn, bli@nlpr.ia.ac.cn, liangpcs@gmail.com, wmhu@nlpr.ia.ac.cn

Subject: Decision to undergo a MAJOR REVISION - TCSVT-07024-2021, A Simple and Strong Baseline for Universal Targeted Attacks on Siamese Visual Tracking

Body: 24-Aug-2021

Dr. Jin Gao
95 Zhongguancun East Road
Beijing
China
100190

Paper:TCSVT-07024-2021 A Simple and Strong Baseline for Universal Targeted Attacks on Siamese Visual Tracking

Dear Dr. Gao:

I am writing to you concerning the above referenced manuscript, which you submitted to the IEEE Transactions on Circuits and Systems for Video Technology.

Based on the enclosed set of reviews, I am recommending that the manuscript undergo a MAJOR REVISION and be resubmitted for consideration by the IEEE Transactions on Circuits and Systems for Video Technology.

We understand that the reason why you select the IEEE TCSVT for your manuscript is that your manuscript has a good match with this journal----Many related papers should have already been published in this journal. Therefore, before your new submission, you have to answer two questions clearly in your revised manuscript and responses: a) what are the 3-5 papers published in the IEEE Transactions on Circuits and Systems for Video Technology, which are most closely related to your manuscript; b) what is distinctive / new about your current manuscript related to these previously published papers.

Enclosed are the comments by the Associate Editor and reviewers of your paper. Please make sure to address ALL of the Associate Editor and reviewers' comments in your revised manuscript and to submit a document explaining in detail how these comments were addressed.

Your revised manuscript must be submitted back to ScholarOne Manuscripts https://mc.manuscriptcentral.com/tcsvt no later than FIVE (5) weeks from the date of this letter together with a reply to the Associate Editor and reviewers' comments to be further considered for publication in the IEEE Transactions on Circuits and Systems for Video Technology. If we do not receive your revised manuscript and reply within this specified time, your manuscript will be considered withdrawn.

You should submit a revised manuscript for consideration by the IEEE Transactions on Circuits and Systems for Video Technology only if you are confident that you can fully satisfy all the Associate Editor and reviewers' concerns. Please also note that, under the IEEE Transactions on Circuits and Systems for Video Technology editorial policy, revised papers that require a further revision will be rejected.

If the Associate Editor or reviewers have requested that the paper be revised to address problems with the use of the English language, you will also need to provide documentation of English editing (receipt from an editing service or letter from a colleague who has assisted with editing).

A professional editing service is available for authors looking to refine and polish the content of their papers for a fee: http://www.aje.com/en. However, the authors are free to select another professional editing service.

To revise your manuscript, log into https://mc.manuscriptcentral.com/tcsvt and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions" click on "Create a Revision." Your manuscript number has been appended to denote a revision.

Once the revised manuscript is prepared, you can upload it and submit it through your Author Center.

When submitting your revised manuscript, you will be able to respond to the comments made by the Associate Editor and reviewers in the space provided and/or upload a pdf document. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the Associate Editor and reviewers.

Please be sure to upload the revised manuscript together with the response to the reviewers' comments and any other documentation required in Dr. Gao’s account. That is the account that will have the number TCSVT-07024-2021 listed again.

ATTACHMENTS (COMMENTS) FROM THE REVIEWERS can be found by you, Dr. Gao, by going to the paper in your account. You will need to access the Author Center in Manuscript Central, scroll down until you see the information for this paper TCSVT-07024-2021, click on (View Letter) next to the word Decision under "Processing Status & Associate Editor," scroll to the very bottom of the decision letter, and click on the attachment(s) - if any, under "Files from Reviewer(s)."

IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant files before completing the submission.

If you have any questions regarding the review process or are experiencing technical difficulties, please contact Desiree Noel at d.noel@ieee.org.

Thank you for your submission to IEEE Transactions on Circuits and Systems for Video Technology. We look forward to receiving your revised manuscript.

Sincerely,

Feng Wu
Editor-in-Chief, IEEE Transactions on Circuits and Systems for Video Technology

AE Comments:

Associate Editor
Comments to the Author:
The reviewers are generally positive towards this paper, acknowleding this paper as being interesting. There are some commonly perceived limitations, though. For example, reviewers raised the issue of typos, which the authors are encouraged to address. Besides, more explanations and ablation experiments are suggeste by reviewers to further validate the proposed approach. I recommend this paper undergo a major revision.

Reviewer Comments:

Reviewer: 1

Comments to the Author
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 [6], 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.

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.

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

Reviewer: 2

Comments to the Author
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. 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 \delta.

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.

Taking all the factors into account, the reviewer suggests a major revision.

Reviewer: 3

Comments to the Author
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.

Reviewer: 4

Comments to the Author
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_{a}dd should be A_{add}.
 (b) In page 2, difference datasets GOT-10K[12], LaSOT[12] 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”.

Missing Key References Question (Optional) (List important references missing from the paper):

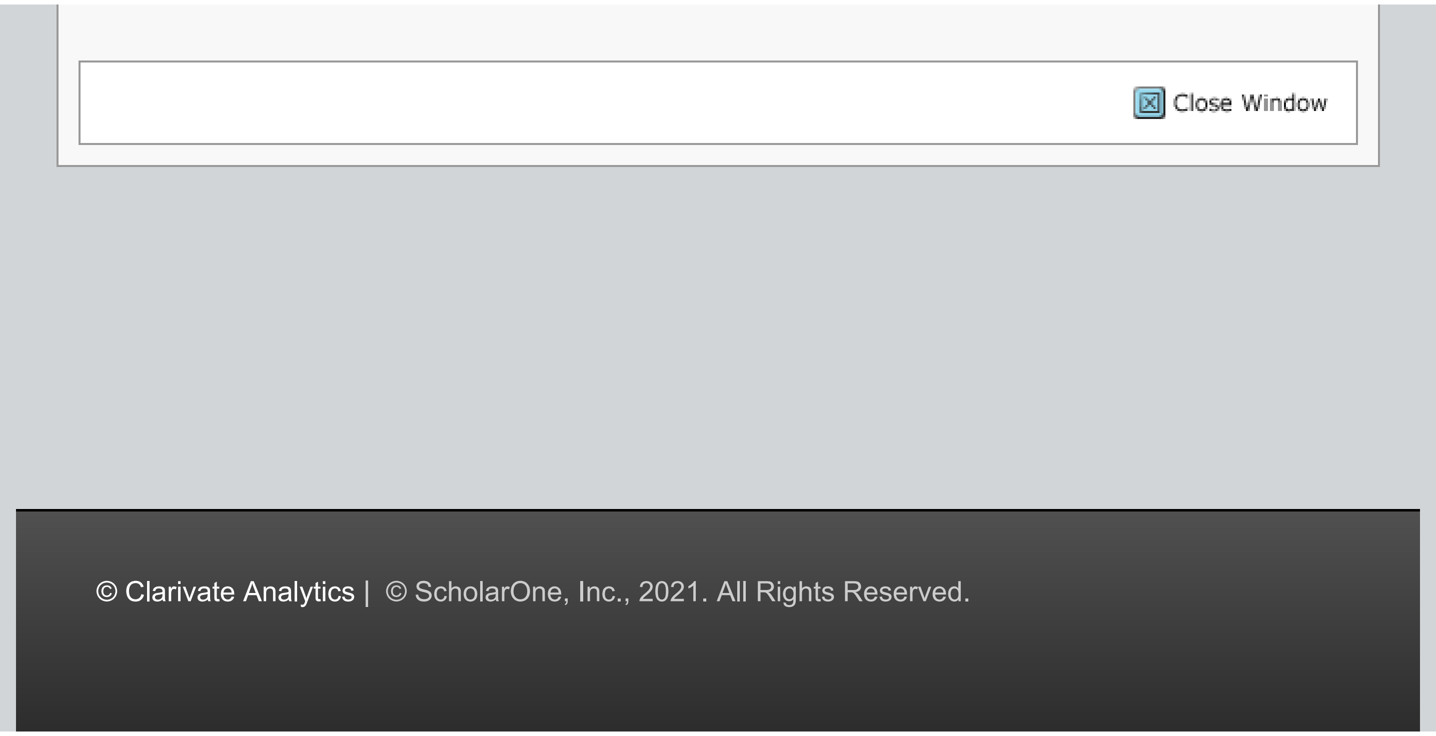
Reviewer: 1
Missing key ref :

Reviewer: 2
Missing key ref :

Reviewer: 3
Missing key ref : "Deep Learning for Visual Tracking: A Comprehensive Survey," in IEEE Transactions on Intelligent Transportation Systems, 2021.

Reviewer: 4
Missing key ref :

Date Sent: 24-Aug-2021



☒ Close Window