We thank the reviewers for their generous comments on the manuscript and we will explain their concerns point by point.

Response to Reviewer #1

1. The authors mentioned that "a universal imperceptible perturbation is added to the template image". However, according to the description in the paper, the perturbation added to template image is not imperceptible.

We admit that the description "imperceptible" is not precise because the perturbation can be found by human eyes. The reason is that the universal attack on the object tracking task is more difficult than on the image classification task. Specifically, the goal of existing universal attack methods is to disturb unary or binary model outputs for single instance while we need to use universal perturbations to mislead Siamese trackers to follow a specified trajectory. We have replaced the phase "imperceptible perturbation" with "translucent perturbation" [1] in the manuscript to avoid ambiguity. What's more, making the perturbation looks like the background is a very meaningful future work.

2. In table VII, the cost for CSA method is 4720ms, which is different from the cost time reported in the original CSA method. In CSA, attacking the template and search region require 3ms and 9ms, respectively. Why is there such a big difference?

In table VII, the attack cost is calculated on the video level instead of the image level. CSA needs 9ms to process a frame and needs an average of 4720ms to attack a video. We have changed "attack cost" to "attack cost per frame" in the manuscript to avoid ambiguity.

Response to Reviewer #2

1. Adversarial patch makes the attack very obvious for the human eye. I think the attack should not only for the deep learning model but also for the human eyes.

This is the same as Reviewer #1's first question.

2. Your baseline approaches are not state-of-the-art algorithms. Why not compare to SPARK and TTP. "However, RTAA only performs the untargeted attacks for trackers, which is less challenging than the targeted attacks in this paper, as we aim to create arbitrary, complex trajectories at test time." If that is the case, it should be easy to achieve better performance than their approach. I think the experiments should including the methods of untargeted attack.

Thanks for your advice and we have added experiments as you suggested. Specifically, we compare with state-of-the-art attack methods including SPARK, TTP, CSA and FAN in both targeted and untargeted attack settings in Sec. IV. F, as shown in Table 1 and Table 2:

Table 1: Untargeted attack: Precision score on OTB2015.

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

Table 2: Targeted attack: Precision score on OTB2015.

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

3. "However, SPARK needs to generate distinct adversarial examples for every search image through heavy iterative schemes, which is time-consuming to attack online-tracking in real-time." But you are not real-time as well. Why do you mention this part? You also didn't show the response time of your system.

Our perturbations are trained off-line and can perturb a novel video to come at no additional cost except the mere addition operations, not requiring gradient optimization or network inference, so we can attack online-tracking in real-time. The attack cost per frame is show in Table 1.

4. There are some datasets adopted in [2, 3] that you should include in your experiments such as VOT2018, VOT2019, VOT2016, OTB2015 and UAV123. Otherwise, it is very difficult to justify the performance of the proposed method.

Thanks for your advice and we have added experiments on VOT2018, VOT2016, OTB2015 and UAV123 in Sec. IV. D, as shown in Table 3 and Table 4:

Table 3: Overall attack results on VOT2016, VOT2018, VOT2019 and UAV123.

Benchmarks	Metrics	Before Attack	Untargeted Attack
VOT2016	Accuracy	0.626	0.393
	Robustness	0.144	9.061
	EAO	0.460	0.007
VOT2018	Accuracy	0.587	0.342
	Robustness	0.183	8.981
	EAO	0.426	0.007
VOT2019	Accuracy	0.556	0.345
	Robustness	0.537	8.824
	EAO	0.243	0.010
UAV123	AO	0.623	0.064
	Precision	0.781	0.187

Table 4: Overall attack results on OTB-15, GOT-Val and LaSOT.

Benchmarks	Metrics	Clean Videos	Perturbed Videos	
		Real Traj.	Real Traj.	Fake Traj.
OTB-15	AO	0.642	0.063	0.759
	Precision	0.861	0.092	0.795
GOT-Val	SR	0.897	0.123	0.890
	AO	0.760	0.153	0.840
LaSOT	Precision	0.514	0.046	0.605
	Norm. Prec.	0.551	0.048	0.702
	AO	0.525	0.069	0.691
FPS		58	58	58

Response to Reviewer #3

1. The authors use the concept of video agnostic, however, for the experiment, I do not see the experiments to demonstrate the video agnostic property of the proposed method. From my understanding, video agnostic property is that the adversarial examples generated from one dataset can be applied to another dataset. However, I do not see the experimental results to demonstrate this and the transferability across different datasets are required to demonstrate the performance of the proposed framework.

We aggree that the adversarial examples generated from one dataset should be applied to another dataset. So we add experiments to verify this. Specifically, we adopt COCO, ILSVRC-VID and the training splits of GOT-10k and LaSOT as our training set. The test set includes VOT2016, VOT2018, VOT2019, UAV123, and OTB2015 (see Table 3 and Table 4). Experimental results demonstrate the transferability of the proposed method across different datasets

2. The paper attacks the video tracking assumed that the training data is available which in practice may not be available. Also, even though we know the video content only but without the ground truth box information, then how can we use the proposed approach to generate effective approach. If not, then the practicability of the approach is quite limited.

Our perturbations generate well on different datasets. Once trained on public datasets (i.e., COCO, ILSVRC-VID and the training splits of GOT-10k and LaSOT), the perturbations can effectively attack videos on VOT2016, VOT2018, VOT2019, UAV123, and OTB2015 (see Table 3 and Table 4). Thus, we can deploy perturbations trained on public datasets directly to real-world scenarios without fine-tuning using the private training set.

3. The technical novelty of this paper is not high as it directly borrows the approach from the traditional adversarial attack and use two additional path to train the neural networks. Can the authors provide theoretical analysis to explain the effectiveness based on such small change?

To the best of our knowledge, our work is the first attempt to generate video-agnostic perturbations to attack siamese trackers. Besides FGSM and Brown et al.' work, other adversarial example generation methods such as C&W [4] and PGD [5] can also be integrated into our attacking system to further improve the attack effect. In short, we focus on proposing a video-agnostic attacking system for siamese trackers instead of prosing a specific adversarial example generation method. By studying the attack method of Siamese trackers, we argue that the disadvantage of Siamese trackers is that template matching-based tracking mechanism is vulnerable to attacks, i.e., if both the template and the search image are perturbed at the same time, then the tracking algorithm can easily be misled. The implication for improving the tracking robustness is to make the tracker aware of the semantic information of the target being tracked, rather than relying on templates alone, so that the semantic information can be relied on to ensure tracking robustness.

4. I suggest the authors to add the table to summarize the characteristics of the used datasets.

Thanks for your advice and we have added the table to summarize the characteristics of the used datasets in Sec. IV. A, as shown in Table 5:

Table 5: Characteristics of the datasets used to train and evaluate the proposed attack method.

	Dataset	Videos	Total frames	Frame rate	Object classes	Num. of attributes
Training set	GOT-10k training split	9.34K	1.4M	10 fps	480	6
	LaSOT training split	1.12K	2.83M	30 fps	70	14
	COCO2017	n/a	118K	n/a	80	n/a
	ILSVRC-VID	5.4K	1.6M	30 fps	30	n/a
Test set	GOT-10k test split	420	56K	10 fps	84	6
	UAV123	123	113K	30 fps	9	12
	LaSOT test split	280	690K	30 fps	70	14
	OTB-15	100	59K	30 fps	22	11
	VOT2016	60	21K	30 fps	16	6
	VOT2018	60	21K	30 fps	24	6
	VOT2019	60	19K	30 fps	30	6

5. The paper contains some grammar mistakes and typos, the authors need to improve it by double check.

Thanks for your advice and we changed the errors in the paper.

References

[1] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 232–15 241.

[2] Q. Guo, X. Xie, F. Juefei-Xu, L. Ma, Z. Li, W. Xue, W. Feng, and Y. Liu, "Spark: Spatial-aware online incremental attack against visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 202–219.

[3] S. Jia, C. Ma, Y. Song, and X. Yang, "Robust tracking against adversarial attacks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 69–84.

[4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.