TrackPGD: Efficient Adversarial Attack using Object Binary Masks against Robust Transformer Trackers

Fatemeh Nourilenjan Nokabadi
Université Laval, IID, Mila
fatemeh.nourilenjan-nokabadi.1@ulaval.ca

Yann Batiste Pequignot
Université Laval, IID
yann.pequignot@iid.ulaval.ca

Jean-François Lalonde
Université Laval, IID
jean-francois.lalonde@gel.ulaval.ca

Christian Gagné
U. Laval, IID, Mila, Canada-CIFAR AI Chair
christian.gagne@gel.ulaval.ca

Abstract

Adversarial perturbations can deceive neural networks by adding small, imperceptible noise to the input. Recent object trackers with transformer backbones have shown strong performance on tracking datasets, but their adversarial robustness has not been thoroughly evaluated. While transformer trackers are resilient to black-box attacks, existing white-box adversarial attacks are not universally applicable against these new transformer trackers due to differences in backbone architecture. In this work, we introduce TrackPGD, a novel white-box attack that utilizes predicted object binary masks to target robust transformer trackers. Built upon the powerful segmentation attack SegPGD, our proposed TrackPGD effectively influences the decisions of transformer-based trackers. Our method addresses two primary challenges in adapting a segmentation attack for trackers: limited class numbers and extreme pixel class imbalance. TrackPGD uses the same number of iterations as other attack methods for tracker networks and produces competitive adversarial examples that mislead transformer and non-transformer trackers such as MixFormerM, OSTrackSTS, TransT-SEG, and RTS on datasets including VOT2022STS, DAVIS2016, UAV123, and GOT-10k.

1 Introduction

Adversarial attacks (Guo et al., 2020; Jia et al., 2021, 2020; Yan et al., 2020) can significantly harm neural networks performances by adding carefully crafted imperceptible noise to the input. For instance, classification networks are highly susceptible to such adversarial perturbations (Shao et al., 2022; Mahmood et al., 2021). For object segmentation and object tracking, some recent attacks (Gu et al., 2022; Guo et al., 2020; Jia et al., 2021) have been developed to disrupt the object masks and bounding boxes output, respectively. Since 2020, the Visual Object Tracking (VOT) challenge (Kristan et al., 2020) introduced a new sub-challenge that requires object trackers to present binary masks as output and be evaluated based on their mask generation ability. In addition, novel trackers with transformer backbones have emerged (Cui et al., 2022; Ye et al., 2022; Chen et al., 2023; Bhat et al., 2020), whose approaches infer the object mask, bounding box, and confidence score as a triple output. Given that objects with complex shapes are rarely aligned with the frame axis, the corresponding axis-aligned bounding boxes contain more background pixels than object pixels (Paul et al., 2022). Therefore, object bounding boxes are not the best option to describe an object's appearance in comparison to object masks. As a result, object binary masks have become increasingly important both as a valid output of trackers and a valid criterion for evaluation in

AdvML-Frontiers'24: The 3nd Workshop on New Frontiers in Adversarial Machine Learning@NeurIPS'24, Vancouver, CA.

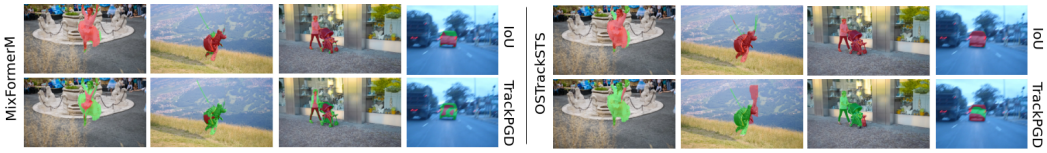

Figure 1: The binary masks generated by MixFormerM (Cui et al., 2022) and OSTrackSTS (Ye et al., 2022) attacked by our TrackPGD vs. IoU (Jia et al., 2021) which is the most recent and applicable attack for these two trackers up to now. The green/red masks are the before/after the attack outputs.

Table 1: Characterization of attack methods– IoU (Jia et al., 2021), CSA (Yan et al., 2020), SPARK (Guo et al., 2020), RTAA (Jia et al., 2020) and TrackPGD (ours)– on robust object trackers including MixFormerM (Cui et al., 2022), OSTrackSTS (Ye et al., 2022), TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022). TrackPGD is the only applicable White-box attack against MixFormerM, OSTrackSTS and RTS, using a new proxy (i.e. binary mask), while the two other methods that may be applied to these trackers, IoU and CSA, are Black-box methods manipulating bounding box.

Attack Setting	Method	Attack Proxy	MixFormerM	OSTrackSTS	TransT-SEG	RTS
Black-box	IoU	Object bbox	✓	✓	✓	✓
	CSA	Object bbox, heat-maps	✓	✓	✓	×
White-box	SPARK	Regression and classification labels	×	×	✓	×
	RTAA	Regression and classification labels	×	×	✓	×
	TrackPGD	**Object binary mask**	✓	✓	✓	✓

tracking benchmarks. Visual tracking can be compromised by adversarial attacks in different ways, by manipulating elements such as object bounding boxes (Jia et al., 2021), classification labels (Jia et al., 2020; Yan et al., 2020; Guo et al., 2020), and regression labels (Jia et al., 2020; Guo et al., 2020). In order to apply white-box attacks on new transformer trackers (Cui et al., 2022; Ye et al., 2022; Chen et al., 2023; Bhat et al., 2020), the adversarial loss should be computed using the tracker's elements and outputs. While some white-box attacks, such as SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020), are not applicable to new robust object trackers, such as RTS (Paul et al., 2022) and MixFormer (Cui et al., 2022), the black-box attacks are not effective enough to challenge these trackers. In this work, we propose TrackPGD, a novel and efficient white-box attack that uses the object binary mask as a proxy. Figure 1 illustrates TrackPGD's effectiveness in disrupting binary masks of MixFormerM (Cui et al., 2022) and OSTrackSTS (Ye et al., 2022), with green/red masks representing predictions before/after the attack. TrackPGD caused more damage than the Intersection over Union (IoU) method (Jia et al., 2021), which only slightly degraded the bounding box while leaving the binary mask mostly unchanged. The IoU method is the most recent applicable attack against MixFormerM (Cui et al., 2022), OSTrackSTS (Ye et al., 2022) and RTS (Paul et al., 2022).

Targeting object binary masks in adversarial attacks is essential due to their increasing use in tracking backbones (Cui et al., 2022; Ye et al., 2022; Yan et al., 2021; Paul et al., 2022) and datasets (Kristan et al., 2023). In Table 1, several attack methods and their attack proxy are listed along with their applicability to some new transformer trackers, MixFormerM (Cui et al., 2022), OSTrackSTS (Ye et al., 2022) and TransT-SEG (Chen et al., 2023). Using a novel attack proxy, the object binary mask, TrackPGD is developed to generate the adversarial frame patches from the predicted binary mask. In addition, the experimental results on object bounding box evaluation demonstrate that the TrackPGD is an effective attack over neural network-based object trackers. Table 1 lists various attacks, the proxies, and applicability to new trackers. For example, MixFormerM (Cui et al., 2022) does not output object candidate labels, making label-based attacks such as RTAA (Jia et al., 2020) and SPARK (Guo et al., 2020) unsuitable as white-box attacks. In a nutshell, our contributions can be summarized as follows:

- We propose a novel white-box attack, TrackPGD, which builds the adversarial noise from the binary mask to attack transformer trackers.
- A new loss term – built upon SegPGD attack – is proposed to compute the TrackPGD loss with the aim of misleading visual trackers in providing an accurate binary mask.

- We demonstrate that MixFormerM suffers significant accuracy (-75%), average overlap (-97%) and robustness (-91%) reductions after applying TrackPGD, evaluated on the VOT2022STS dataset.
- Experimental results also demonstrate that the perturbations generated by TrackPGD have a significant influence on bounding box predictions in tracking benchmarks.

2 Related works

Robust Object Trackers Tracker robustness was initially defined as the frequency of failures, marked by the overlap between prediction and ground truth dropping to zero. This definition is employed in datasets like VOT2016 (Kristan et al., 2016) and VOT2018 (Kristan et al., 2018). A newer protocol introduced in VOT2020 (Kristan et al., 2020) includes multiple anchors in each test sequence spaced approximately 50 frames apart to facilitate anchor-based short-term tracking (Kristan et al., 2020, 2023). In contrast, other tracking datasets, like LaSOT benchmark (Fan et al., 2019), conduct a one-pass evaluation. The MixFormerM (Cui et al., 2022) was one of the three most robust trackers on the VOT2022STS (Kristan et al., 2023). The MixFormer (Cui et al., 2022) tracker proposed a Mixed Attention Module (MAM) to jointly extract features of search and template regions. MixFormerM (Cui et al., 2022; Kristan et al., 2023) contains two networks, one tracker network and the other is AlphaRefine (Yan et al., 2021) for object segmentation. But TransT-SEG (Chen et al., 2023) jointly predicts the object bounding box and generates the object binary mask via a single network that uses a light cross-attention mechanism to fuse deep features. Another transformer tracker, OSTrackSTS (Ye et al., 2022), gained the third rank on the VOT-STS2022 (Kristan et al., 2023) sequestered dataset. The OSTrackSTS (Ye et al., 2022) is the One-Stream Track method (Ye et al., 2022) coupled with the AlphaRefine (Yan et al., 2021) to provide object binary masks. The Robust Tracking by Segmentation (Paul et al., 2022) is a new robust discrimintative tracker to infer the object binary mask from a whole video frame. Using a segmentation branch, and an instance localization branch, RTS achieved a robust tracking performance by predicting the target mask in the tracking process (Kristan et al., 2023).

Adversarial attacks for object trackers The attack proxy can be classification labels (Jia et al., 2020; Yan et al., 2020; Guo et al., 2020), regression output (Jia et al., 2020) or even the motion features of the object (Jia et al., 2021, 2020). In the Robust Tracking against Adversarial Attacks (RTAA) method (Jia et al., 2020), specific labels for classification and regression are modified. An online incremental attack called SPARK (Guo et al., 2020) has been developed for object trackers that use the previous frame perturbations and manipulate the object classification and regression labels to generate the adversarial frame. The Cooling-Shrinking Attack (CSA) method, as described in (Yan et al., 2020), involves training two Generative Adversarial Networks (GANs) to create adversarial images for the SiamRPN++ tracker (Li et al., 2019). The cooling term cools down the heated pixels in the heatmap of SiamRPN++, while the shrinking term shrinks the predicted bounding box. The IoU attack is a black-box approach that uses the predicted bounding box to inject two types of noises (Jia et al., 2021) into frames. The directional and orthogonal perturbations are injected in a given frame to degrade the predicted bounding box from the previous prediction.

3 Preliminaries: SegPGD

The SegPGD (Gu et al., 2022) loss for a segmentation network (L_{SegPGD}) is defined to steer the neural network toward providing an inaccurate segmentation map. This loss is computed between the mask annotation Y and the adversarial segmentation map $f_{\text{seg}}(X^{\text{adv}})$ generated by the segmentation network from the adversarial image X^{adv}:

$$L_{\text{SegPGD}}(f_{\text{seg}}(X^{\text{adv}}), Y) = \frac{1-\lambda}{HW} L_{\text{pos}} + \frac{\lambda}{HW} L_{\text{neg}}\,, \tag{1}$$

where λ is a coefficient that dynamically changes over iterations and H and W are the frame height and width, respectively. Here, L_{pos} and L_{neg} denote the cross-entropy loss restricted to correctly classified and wrongly classified pixels, respectively. The motivation for this reweighting stems from the observation that gradients of cross-entropy for wrongly classified examples tend to have larger amplitude, while the attack aims primarily at correctly classified pixels (Gu et al., 2022).

To adapt SegPGD for use against object trackers and to target the predicted binary masks, we must overcome two key challenges. Firstly, while SegPGD operates on multi-class segmentation networks, binary masks only have two classes, necessitating the switching of classification labels between pixels. Secondly, the labels in the majority of samples are imbalanced, with the number of object pixels being significantly fewer than background pixels. To address these hurdles, we introduce TrackPGD as a white-box attack for object trackers which predict binary masks.

4 Proposed method: TrackPGD

Let us formalize object tracking as using a tracker $\mathcal{F}(\cdot)$ that processes a video sequence $\mathcal{V} = \{I_0, \ldots, I_L\}$ with $L+1$ frames, having the first frame object mask M_0 provided to initialize the tracker. The transformer tracker is expected to estimate the object mask M_τ at each time step $\tau \in \{1, \ldots, L\}$ from the given video frame I_τ. The goal of the attack method is thus to create an adversarial video frame I_τ^{adv} at each time step to mislead the object tracker in its mask prediction. The transformer tracker predicts at each time τ a mask M_τ which assigns to each pixel a probability of being an object. The TrackPGD attack aims to mislead the tracker in making these assignments by creating the adversarial frame I_τ^{adv}. For object tracking, the mask annotation G_τ consists of two labels (object (1) and background (0)). It follows that finding adversarial examples can be achieved at optimization step t both by maximizing $L_{\text{SegPGD}}(M^t, G_\tau)$ or by minimizing $L_{\text{SegPGD}}(M^t, 1 - G_\tau)$, where we consider the mask $M_{\tau-1}$ predicted on the previous frame as the ground truth G_τ at step τ. In the TrackPGD framework, we therefore decide to combine these two complementary objectives into a difference loss L_Δ as follows:

$$L_\Delta = L_{\text{SegPGD}}(M^t, G_\tau) - L_{\text{SegPGD}}(M^t, 1 - G_\tau)\,. \tag{2}$$

Since the object pixels and background pixels are imbalanced in the video frame, we employed the focal loss (Lin et al., 2020). It is a useful loss for training highly imbalanced data, which is the case, here, for the object and background pixels. For TrackPGD, instead of using cross-entropy as in the original equation of focal loss (Lin et al., 2020), we use the L_Δ loss as follows:

$$\begin{aligned} p_t &= G_\tau \mathcal{P} + (1 - G_\tau)(1 - \mathcal{P}) \\ L_{\text{focal}} &= \alpha_t (1 - p_t)^\gamma L_\Delta\,, \end{aligned} \tag{3}$$

where $\mathcal{P}$ is a probability map derived from the application of the sigmoid function on the predicted binary mask M_t (before thresholding) and γ is the focusing parameter. This α-balanced variant of the focal loss used is similar to the TransT-SEG (Chen et al., 2023) except that L_Δ is used instead of cross-entropy. Then, we compute the dice loss (Milletari et al., 2016) between M^t and G_τ as follows:

$$L_{\text{dice}} = 1 - 2\text{IoU}(M^t, G_\tau), \tag{4}$$

where IoU is the Intersection over Union loss between the ground truth G_τ and the predicted binary mask M^t. The combination of focal loss (Equation 3), and dice loss (Equation 4), is used to compute the TrackPGD loss as follows:

$$L = \lambda_1 L_{\text{focal}} + \lambda_2 L_{\text{dice}}\,, \tag{5}$$

where the λ_1 and λ_2 balance the focal and dice terms.

For a detailed explanation of the TrackPGD algorithm and additional studies, please refer to the Supplementary materials.

5 Experiments

Datasets. Several video segmentation and tracking datasets. including DAVIS 2016 (Perazzi et al., 2016), VOT2022STS (Kristan et al., 2023), VOT2018 (Kristan et al., 2018), GOT10k (Huang et al., 2019), and UAV123 (Mueller et al., 2016), are adopted in our experiments.

Victim Trackers. We selected three popular transformer-based trackers, MixFormerM (Cui et al., 2022), OSTrackSTS (Ye et al., 2022), TransT-SEG (Chen et al., 2023) and one non-transformer tracker, RTS (Paul et al., 2022), to assess their adversarial robustness against adversarial attacks.

Table 2: The MixFormerM (Cui et al., 2022), OSTrackSTS (Ye et al., 2022) results in generating object binary masks, when attacked by different methods on the VOT-STS2022 (Kristan et al., 2023) dataset. TrackPGD is the only applicable white-box attack on MixFormerM and OSTrackSTS trackers. Bold/italic are the first/second ranks, respectively. The other white-box attacks, RTAA and SPARK are not applicable to these trackers due to difference in backbones. For further information, please check the text.

		Baseline				Unsupervised		Drop %	
Tracker	Attack	EAO	Accuracy	Robust.	AUC	EAO	Accuracy	Robust.	AUC
MixFormerM	No Attack	0.59	0.80	0.88	0.72	–	–	–	–
	CSA	0.56	0.80	0.86	0.65	4.58	-0.63	2.61	9.40
	IoU	*0.36*	*0.66*	*0.68*	*0.51*	*39.05*	*17.29*	*23.07*	*15.40*
	TrackPGD	**0.02**	**0.20**	**0.08**	**0.06**	**96.94**	**75.44**	**91.14**	**91.70**
OSTrackSTS	No Attack	0.59	0.77	0.87	0.66	-	-	-	-
	CSA	0.53	0.76	*0.82*	0.62	10.54	2.32	*5.29*	5.79
	IoU	*0.31*	*0.71*	**0.60**	*0.48*	*47.45*	*8.13*	**30.92**	*27.94*
	TrackPGD	**0.23**	**0.31**	0.86	**0.27**	**60.34**	**59.74**	0.80	**59.06**

Adversarial Attacks. We applied four attack methods including CSA (Yan et al., 2020), SPARK (Guo et al., 2020), RTAA (Jia et al., 2020), and IoU (Jia et al., 2021), against these trackers. The TrackPGD perturbs the search region of the segmentation network in MixFormerM (Cui et al., 2022) and OSTrackSTS (Ye et al., 2022) while the search region of the tracking network remains untouched. The hyperparameters (λ_1, λ_2) of the TrackPGD loss are determined through the validation experiment on the DAVIS2016 (Perazzi et al., 2016), as explained in the section Fine tuning the parameters.

Metrics. The expected average overlap (denoted as EAO and AO), robustness, accuracy and AUC from VOT2022 metrics, Mean Jaccared Index $J(M)\%$ from DAVIS2016 evaluation toolkit, AUC and precision rates for GOT01k and UAV123 datasets are used to rank the attack performance.

5.1 Binary mask assessment

Quantitative evaluation. We evaluated the TrackPGD attack compared to other methods on two transformer trackers – MixFormerM (Cui et al., 2022) and OSTrackSTS (Ye et al., 2022) – using the VOT-STS2022 dataset, see Table 2. Two streams of experiments from VOT-STS2022 (baseline and unsupervised) are conducted to indicate our attack performance under various protocols and metrics. Our findings demonstrate that TrackPGD was the most effective attack, reducing the robustness of MixFormerM (Cui et al., 2022) by around 91% in the baseline experiment of VOT-STS2022 evaluation. TrackPGD also caused higher degradations to other evaluation metrics – EAO, accuracy, and AUC – than the two attack methods (i.e., IoU (Jia et al., 2021) and CSA (Yan et al., 2020) methods) against MixFormerM (Cui et al., 2022), both on baseline and unsupervised streams of experiments. Among all measured metrics, TrackPGD ranked first by a significant margin on the VOT-STS2022 dataset. In attacking the OSTrackSTS tracker, TrackPGD also ranked first as an applicable attack per EAO, accuracy and AUC metrics. However, the IoU method (Jia et al., 2021) performs better than our method on the robustness metric. To check further examples of mask prediction after attacks, we refer the reader to the Supplementary materials.

5.2 Bounding box assessment

Quantitative evaluation. Table 5 presents the attack results on the TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers using the GOT-10k dataset (Huang et al., 2019). Although our method focuses on manipulating object binary masks, the reported numbers for bounding box evaluation are very close to those of competitive attacks aimed at misleading trackers into providing inaccurate bounding boxes. We surpassed the RTAA attack on the GOT-10k dataset in both success rate scores. However, RTAA performs better in terms of average overlap (AO). TrackPGD scores are also very close to SPARK scores, the most competitive attack using object bounding boxes on GOT-10k (Huang et al., 2019), in misleading trackers to predict accurate object bounding boxes. Despite our method's aim to misclassify object and background pixels, it proves to be a highly competitive approach compared to bounding box-centered attacks. Furthermore, our TrackPGD has an extra

Table 3: TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers evaluation after white-box applicable approaches, SPARK (Guo et al., 2020), RTAA (Jia et al., 2020) and TrackPGD (ours), on the GOT10k (Huang et al., 2019) dataset.

Tracker	Attacker	AO	$SR_{0.5}$	$SR_{0.75}$
TransTSEG	No Attack	0.719	0.816	0.680
	RTAA	0.127	0.115	0.060
	SPARK	0.120	0.068	0.015
	TrackPGD (ours)	0.139	0.092	0.025
RTS	No Attack	0.736	0.830	0.681
	TrackPGD (ours)	0.292	0.291	0.162

Table 4: Evaluation of our proposed TrackPGD vs. other white-box attacks applied on TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers on the UAV123 dataset.

Tracker	Attacker	AUC	Precision
TranstSEG	No attack	0.670	0.859
	SPARK	0.050	0.074
	RTAA	0.058	0.077
	TrackPGD	0.135	0.205
RTS	No attack	0.666	0.873
	TrackPGD	0.427	0.674

advantage over other white-box attacks which is that other attack opponents, RTAA and SPARK, are not applicable as a white-box attack on new transformer and non-transformer trackers containing MixFormerM, OSTrackSTS and RTS. For the RTS tracker (Bhat et al., 2020), though, TrackPGD is the only white-box attack which is applicable and significantly affect the tracker performance.

The TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers robustness after a white-box attack is also assessed on the UAV123 dataset in Table 8. Once again, our method achieves scores that are closely competitive with SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020) attacks on bounding box prediction, even though TrackPGD is developed to mislead trackers in predicting binary masks. Also, our method is the first white-box attack applicable on RTS tracker and presents a very effective performance in decreasing the RTS scores on the UAV123 (Mueller et al., 2016) dataset. To check further examples of bounding box prediction after attacks, we refer the reader to the Supplementary materials.

6 Conclusion

Existing adversarial attacks for object trackers are ineffective as white-box attacks against transformer trackers. To evaluate the adversarial robustness of transformer trackers, new white-box attacks that utilize the tracker's own gradients are necessary. This paper introduces TrackPGD, a novel white-box attack that leverages object binary masks to assess the adversarial robustness of transformer trackers. We highlight the effectiveness of our proposed difference loss in impacting tracker performance compared to standard segmentation losses such as SegPGD. The efficacy of TrackPGD is validated through comprehensive experiments on various transformer and non-transformer networks on popular datasets. Despite the strong performance of our method, TrackPGD is limited to object trackers that generate a binary mask. Future work could extend these attacks to a broader range of trackers. Additionally, exploring defense strategies, such as adversarial purification with generative models (e.g., GANs or diffusion models) or adversarial training, could improve adversarial robustness in object tracking networks.

Acknowledgments

This work is supported by the DEEL Project CRDPJ 537462-18 funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Consortium for Research and Innovation in Aerospace in Québec (CRIAQ), together with its industrial partners Thales Canada inc, Bell Textron Canada Limited, CAE inc and Bombardier inc. [1]

References

Guo, Q.; Xie, X.; Juefei-Xu, F.; Ma, L.; Li, Z.; Xue, W.; Feng, W.; Liu, Y. SPARK: Spatial-Aware Online Incremental Attack Against Visual Tracking. IEEE Conf. Euro. Conf. Comput. Vis. 2020.

Jia, S.; Song, Y.; Ma, C.; Yang, X. IoU Attack: Towards Temporally Coherent Black-Box Adversarial Attack for Visual Object Tracking. IEEE Conf. on Comput. Vis. and Pattern Recog. 2021.

Jia, S.; Ma, C.; Song, Y.; Yang, X. Robust Tracking Against Adversarial Attacks. IEEE Conf. Euro. Conf. Comput. Vis. 2020.

Yan, B.; Wang, D.; Lu, H.; Yang, X. Cooling-Shrinking Attack: Blinding the Tracker With Imperceptible Noises. IEEE Conf. on Comput. Vis. and Pattern Recog. 2020.

Shao, R.; Shi, Z.; Yi, J.; Chen, P.-Y.; Hsieh, C.-J. On the Adversarial Robustness of Vision Transformers. *Trans. on Mach. Learn. Research* **2022**,

Mahmood, K.; Mahmood, R.; van Dijk, M. On the Robustness of Vision Transformers to Adversarial Examples. IEEE Conf. on Comput. Vis. 2021.

Gu, J.; Zhao, H.; Tresp, V.; Torr, P. H. S. SegPGD: An Effective and Efficient Adversarial Attack for Evaluating and Boosting Segmentation Robustness. IEEE Conf. Euro. Conf. Comput. Vis. 2022.

Kristan, M.; Leonardis, A.; Matas, J. The Eighth Visual Object Tracking VOT2020 Challenge Results. IEEE Conf. Euro. Conf. Comput. Vis. Worksh. 2020.

Cui, Y.; Jiang, C.; Wang, L.; Wu, G. MixFormer: End-to-End Tracking With Iterative Mixed Attention. IEEE Conf. on Comput. Vis. and Pattern Recog. 2022.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. IEEE Conf. Euro. Conf. Comput. Vis. 2022.

Chen, X.; Yan, B.; Zhu, J.; Lu, H.; Ruan, X.; Wang, D. High-Performance Transformer Tracking. *IEEE Trans. on Pattern Analy. and Machine Intel.* **2023**, *45*, 8507–8523.

Bhat, G.; Lawin, F. J.; Danelljan, M.; Robinson, A.; Felsberg, M.; Van Gool, L.; Timofte, R. Learning What to Learn for Video Object Segmentation. IEEE Conf. Euro. Conf. Comput. Vis. 2020; pp 777–794.

Paul, M.; Danelljan, M.; Mayer, C.; Van Gool, L. Robust Visual Tracking by Segmentation. IEEE Conf. Euro. Conf. Comput. Vis. 2022; pp 571–588.

Yan, B.; Zhang, X.; Wang, D.; Lu, H.; Yang, X. Alpha-Refine: Boosting Tracking Performance by Precise Bounding Box Estimation. IEEE Conf. on Comput. Vis. and Pattern Recog. 2021.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M. The Tenth Visual Object Tracking VOT2022 Challenge Results. IEEE Conf. Euro. Conf. Comput. Vis. Worksh. 2023.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Pflugfelder, R. The Visual Object Tracking VOT2016 Challenge Results. IEEE Conf. Euro. Conf. Comput. Vis. 2016; pp 777–823.

Kristan, M.; Leonardis, A.; Matas, J.; Felsberg, M.; Feng, Z.-H. The Sixth Visual Object Tracking VOT2018 Challenge Results. IEEE Conf. Euro. Conf. Comput. Vis. Worksh. 2018.

Fan, H.; Ling, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. IEEE Conf. on Comput. Vis. and Pattern Recog. 2019.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks. IEEE Conf. on Comput. Vis. and Pattern Recog. 2019.

[1] https://deel.quebec

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. on Pattern Analy. and Machine Intel.* **2020**, *42*, 2980–2988.

Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Int. conf. 3D Vis. 2016.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; Sorkine-Hornung, A. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. IEEE Conf. on Comput. Vis. and Pattern Recog. 2016.

Huang, L.; Zhao, X.; Huang, K. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. on Pattern Analy. and Machine Intel.* **2019**, *43*, 1562–1577.

Mueller, M.; Smith, N.; Ghanem, B. A Benchmark and Simulator for UAV Tracking. IEEE Conf. Euro. Conf. Comput. Vis. 2016.

Rony, J.; Hafemann, L. G.; Oliveira, L. S.; Ayed, I. B.; Sabourin, R.; Granger, E. Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses. IEEE Conf. on Comput. Vis. and Pattern Recog. 2019.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer Tracking. IEEE Conf. on Comput. Vis. and Pattern Recog. 2021.

A Supplementary materials

A.1 Algorithm

The principle steps of TrackPGD algorithm is presented in Algorithm 1. The sign of step size α, line 9 in Algorithm 1, is determined according the best performance of each tracker, following the decoupling of the norm and direction of gradients in adversarial attacks (Rony et al., 2019).

Algorithm 1 TrackPGD to attack transformer trackers with segmentation capability

Require: Tracker $\mathcal{F}(\cdot)$, current frame I_τ, previous binary mask $M_{\tau-1}$, perturbation range ϵ, step size α, loss trade-offs λ_1 and λ_2, maximum iteration T, focusing parameter γ, variant of focal loss α_t, probability map p_t

1: $I^0_{\text{adv}} \leftarrow I_\tau$ ▷ initialization
2: $G_\tau \leftarrow M_{\tau-1}$ ▷ use last predicted binary mask as ground truth
3: **for** $t = 1 \dots T$ **do**
4: $\quad M^t \leftarrow \mathcal{F}(I^{t-1}_{\text{adv}})$ ▷ predict binary mask
5: $\quad L_\Delta \leftarrow L_{\text{SegPGD}}(M^t, G_\tau) - L_{\text{SegPGD}}(M^t, 1 - G_\tau)$ ▷ compute difference of SegPGD losses
6: $\quad L_{\text{focal}} \leftarrow \alpha_t(1 - p_t)^\gamma L_\Delta$ ▷ compute focal loss
7: $\quad L_{\text{dice}} \leftarrow 1 - 2\,\text{IoU}(M^t, G_\tau)$ ▷ compute dice loss
8: $\quad L \leftarrow \lambda_1 L_{\text{focal}} + \lambda_2 L_{\text{dice}}$ ▷ compute TrackPGD loss
9: $\quad I^t_{\text{adv}} \leftarrow I^{t-1}_{\text{adv}} + \alpha\,\text{sign}(\nabla_{I^{t-1}_{\text{adv}}} L)$ ▷ update adversarial example
10: $\quad I^t_{\text{adv}} \leftarrow \phi^\epsilon\left(I^t_{\text{adv}}\right)$ ▷ clip to the ϵ-ball
11: **end for**

A.2 Ablation study

Role of L_Δ in TrackPGD Inspired by the SegPGD attack (Gu et al., 2022), TrackPGD relies on the predicted binary mask to generate the adversarial frames. The main challenge in following SegPGD for binary mask impairment to attack object trackers is that the SegPGD works well as a multi-class segmentation attack; while here, we have only two classes to switch the pixel labels. The object class and the background class pixels should switch their labels after TrackPGD. However, the object pixels are limited in comparison to the background pixels in almost all of the video frames. For this specific reason, it is necessary to define our method based on both object and background masks to involve all of the frame pixels in generating the adversarial perturbation. Plus, given the availability of two classes in the tracking scenario, we customized the segmentation loss using the tracking loss of the TransT (Chen et al., 2021) to effectively attack binary masks. We also demonstrate the role of L_Δ in TrackPGD versus using the vanilla SegPGD losses for the object and background masks, in Figure 2. The after attack masks show the essential role of our proposed combination loss L_Δ in

Figure 2: The before (green) and after attack masks (red) generated by MixFormerM (Cui et al., 2022), while the vanilla SegPGDs and difference loss are used instead of binary cross entropy in the focal loss for the TrackPGD attack.

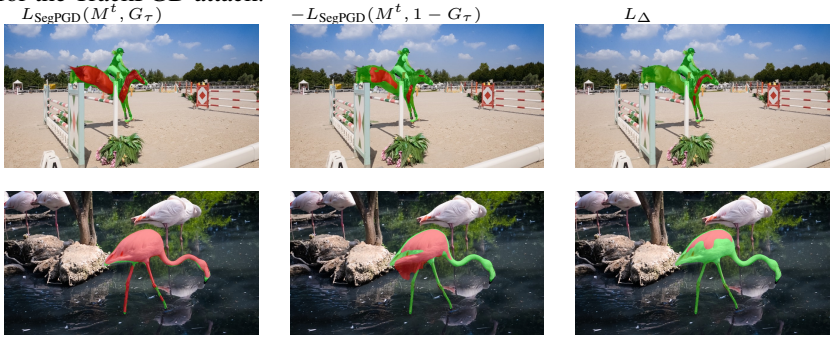

Table 5: The impact of cross-entropy substitutions in focal loss on $J(M)\%$. The vanilla SegPGD losses versus the difference loss in the TrackPGD against MixFormerM (Cui et al., 2022) on DAVIS2016 (Mueller et al., 2016).

Original	$L_{\text{SegPGD}}(M^t, G_\tau)$	$-L_{\text{SegPGD}}(M^t, 1 - G_\tau)$	L_Δ
85.82	52.86	37.28	**30.30**

crashing the object mask inferred by MixFormerM (Cui et al., 2022). Although, the vanilla SegPGD losses also generate inaccurate binary masks, the mask impairment caused by L_Δ is significantly greater.

In addition, we conducted an ablation study to demonstrate the difference loss role in the focal loss. We predicted the binary masks by the MixFormerM tracker attacked by TrackPGD where the binary cross entropy is substituted with the vanilla SegPGD loss for the object mask, vanilla SegPGD for the background mask, and the difference loss L_Δ. The DAVIS2016 (Perazzi et al., 2016) dataset was used to compute the percentage of the mean Jaccard index $J(M)$ of the binary masks indicating the average overlap between predicted and annotated binary masks. TrackPGD attack is stronger with the proposed difference loss, see Table 5, in comparison with the vanilla SegPGD losses for the object mask G_τ, and the background mask $1 - G_\tau$. The difference loss presented the most effective attack based on the generated binary masks by MixFormerM (Cui et al., 2022), which has the smaller $J(M)\%$ on average.

Coefficients of the TrackPGD Loss In this study, we show the effect of each loss term, dice and focal losses, in the TrackPGD attack. We conducted experiments on DAVIS2016 (Perazzi et al., 2016) using two evaluation metrics: the mean Jaccard index $J(M)$ and contour accuracy $F(M)$ of the predicted masks. The results were presented in Table 6 where the average $J\&F(M)$ percentage was used to demonstrate the impact of each losses term, focal vs. dice loss. Table 6 shows that the focal loss is more effective than the dice loss, resulting in smaller $J\&F(M)$ values. The $J\&F(M)$ value is used to measure the effectiveness of the attack, where a smaller value represents a more significant influence. We selected powers of 10 as coefficients of focal loss and powers of 2 as coefficients of dice loss. As in the literature, the loss related to IoU is normally 1 or 2 (Chen et al., 2023; Cui et al., 2022), while the loss terms related to binary cross entropy (or focal loss) are selected as 1000 (Yan et al., 2021) or 1 (Chen et al., 2023). An interesting observation in Table 6 is that the MixFormerM is significantly impacted by TrackPGD, more than the TransT-SEG tracker using the same set of experiments. In both tables, the MixFormerM tracker achieves the smallest values of $J\&F(M)$ meaning more significant attack results.

A.3 Fine tuning the parameters

We used the DAVIS2016 dataset as the validation set to adjust the λ_1 and λ_2 parameters. The DAVIS protocol calculates the Jaccard Index $\mathcal{J}$ and contour accuracy $\mathcal{F}$ via three error statistic measures: mean, recall, and decay. The Jaccard index measures the Intersection over Union (IoU) metric, while the contour accuracy is computed by calculating the F-measure of the predicted mask and the ground truth. To obtain the F-measure, the predicted mask is considered as a set of closed contours, and precision P_c and recall R_c are calculated between contour points by comparing these contours with the ground-truth contours. Then, the F-measure $\frac{2P_cR_c}{P_c+R_c}$ is computed to represent the accuracy of the predicted mask. To fine-tune the TrackPGD parameters, we computed

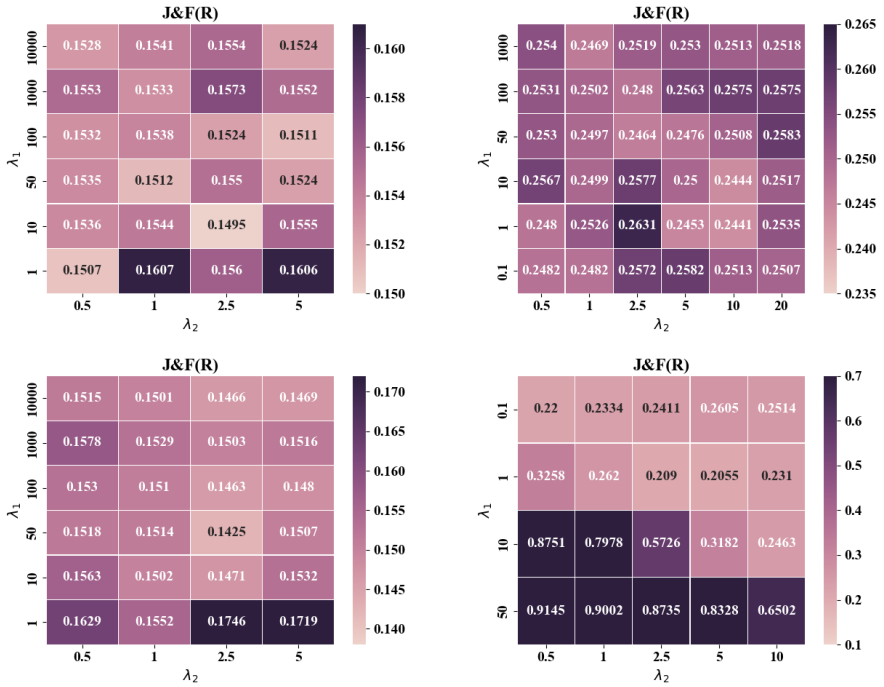

Figure 3: The TrackPGD parameter heatmaps from left to right, first row: MixFormerM(Cui et al., 2022; Yan et al., 2021), OSTrackSTS (Ye et al., 2022; Yan et al., 2021) and second row: TransT-SEG (Chen et al., 2023), RTS (Paul et al., 2022) trackers.

the average value of $\mathcal{J}(R)$ and $\mathcal{F}(R)$, represented as $\mathcal{J}\&\mathcal{F}(R)$. This approach considers both IoU and contour accuracy over time across video frames. Since this measurement is close to the robustness definition in VOT challenge (Kristan et al., 2023), we chose this metric to identify the best hyperparameters to determine the optimal coefficient of TrackPGD loss.

The following heatmaps show the average values of the Jaccard index $\mathcal{J}$ and contour accuracy $\mathcal{F}$ for the tracker's recall metric. The better performance of the tracker is represented by the greater $\mathcal{J}\&\mathcal{F}(R)$. As our reported numbers are based on attack results, a smaller number indicates a more effective attack. Therefore, the best parameters for MixFormerM are $(\lambda_1, \lambda_2) = (10, 2.5)$, while the TrackPGD is strongest against OSTrackSTS with $(\lambda_1, \lambda_2) = (1, 10)$, against TransT-SEG with $(\lambda_1, \lambda_2) = (50, 2.5)$ and against RTS with $(\lambda_1, \lambda_2) = (1, 5)$. It is also worth mentioning that optimal performance is achieved when both terms are active at intermediate values, as shown in Figure 3. This suggests that neither term alone yields the best results.

A.4 Binary mask assessment

Qualitative evaluation. Figure 4 represents the performance of TrackPGD in generating inaccurate binary masks for MixFormerM (Cui et al., 2022) and OSTrackSTS (Ye et al., 2022) and RTS (Paul et al., 2022) trackers on the successive frames of several video sequences. The green color represents the original trackers' output,

Table 6: $J\&F(M)\%$ values with varying focal and dice losses. Present the role of focal loss (changing λ_1 when $\lambda_2 = 0$) and dice loss (changing λ_2 when $\lambda_1 = 0$) in the TrackPGD attack. $J\&F(M)\%$ values are computed on DAVIS2016 (Mueller et al., 2016) dataset.

λ_1 ($\lambda_2 = 0$)	10000	1000	100	10	1	0.1
MixFormerM	27.97	27.87	27.76	27.90	**27.73**	27.93
TransT-SEG	28.98	**28.52**	28.74	28.71	28.65	28.67
λ_2 ($\lambda_1 = 0$)	4	2	1	0.5	0.25	0.125
MixFormerM	30.00	30.27	**29.92**	30.04	30.14	29.92
TransT-SEG	33.62	33.62	33.52	34.05	**33.14**	33.55

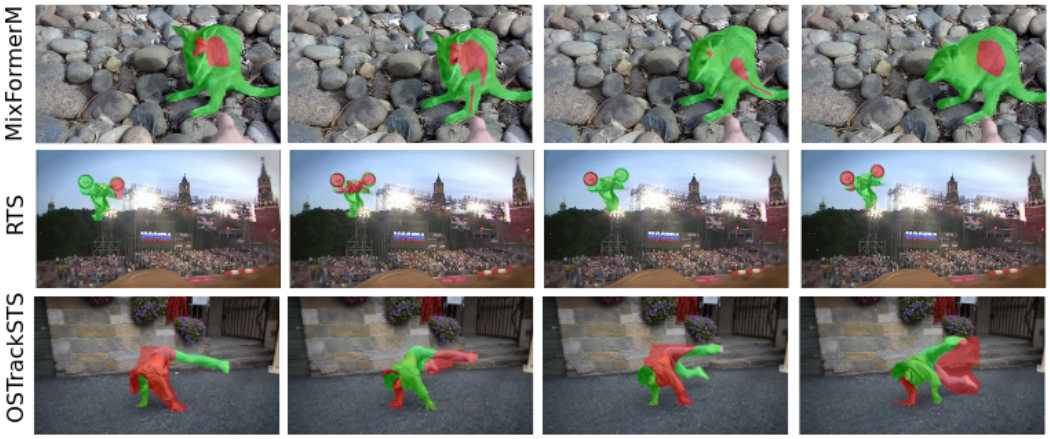

Figure 4: The results of MixFormerM (Cui et al., 2022), RTS (Paul et al., 2022), and OSTrackSTS (Ye et al., 2022) attacked by TrackPGD. The green mask is the original output, while the red mask is the after attack mask.

Table 7: TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers evaluation after white-box applicable approaches, SPARK (Guo et al., 2020), RTAA (Jia et al., 2020) and TrackPGD (ours), on the GOT10k (Huang et al., 2019) dataset.

Tracker	Attacker	AO	$SR_{0.5}$	$SR_{0.75}$
TransTSEG	No Attack	0.719	0.816	0.680
	RTAA	0.127	0.115	0.060
	SPARK	0.120	0.068	0.015
	TrackPGD (ours)	0.139	0.092	0.025
RTS	No Attack	0.736	0.830	0.681
	TrackPGD (ours)	0.292	0.291	0.162

while the red color represents the tracker output after the TrackPGD attack. As it is presented in Table 2, the effect of the TrackPGD attack on the MixFormerM performance is more significant.

A.5 Bounding box assessment

Quantitative evaluation. Table 5 presents the attack results on the TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers using the GOT-10k dataset (Huang et al., 2019). Although our method focuses on manipulating object binary masks, the reported numbers for bounding box evaluation are very close to those of competitive attacks aimed at misleading trackers into providing inaccurate bounding boxes. We surpassed the RTAA attack on the GOT-10k dataset in both success rate scores. However, RTAA performs better in terms of average overlap (AO). TrackPGD scores are also very close to SPARK scores, the most competitive attack using object bounding boxes on GOT-10k (Huang et al., 2019), in misleading trackers to predict accurate object bounding boxes. Despite our method's aim to misclassify object and background pixels, it proves to be a highly competitive approach compared to bounding box-centered attacks. Furthermore, our TrackPGD has an extra advantage over other white-box attacks which is that other attack opponents, RTAA and SPARK, are not applicable as a white-box attack on new transformer trackers containing MixFormerM, OSTrackSTS and RTS. For the RTS tracker (Bhat et al., 2020), though, TrackPGD is the only white-box attack which is applicable and significantly affect the tracker performance.

The TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers robustness after a white-box attack is also assessed on the UAV123 dataset in Table 8. Once again, our method achieves scores that are closely competitive with SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020) attacks on bounding box prediction, even though TrackPGD is developed to mislead trackers in predicting binary masks. Also, our method is the first white-box attack applicable on RTS tracker and presents a very effective performance in decreasing the RTS scores on the UAV123 (Mueller et al., 2016) dataset.

Table 8: Evaluation of our proposed TrackPGD vs. other white-box attacks applied on TransT-SEG (Chen et al., 2023) and RTS (Paul et al., 2022) trackers on the UAV123 dataset.

Tracker	Attacker	AUC	Precision
TranstSEG	No attack	0.670	0.859
	SPARK	0.050	0.074
	RTAA	0.058	0.077
	TrackPGD	0.135	0.205
RTS	No attack	0.666	0.873
	TrackPGD	0.427	0.674

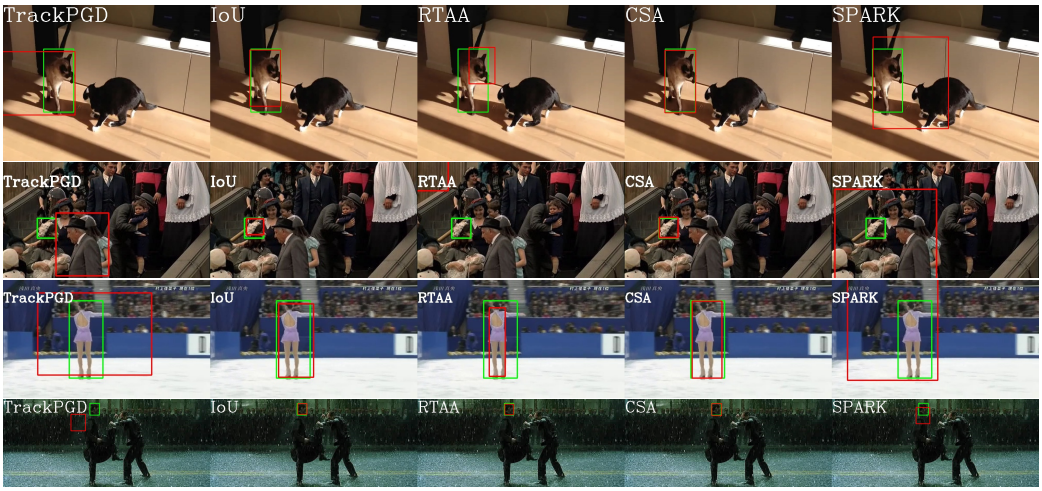

Figure 5: The TransT-SEG (Chen et al., 2023) tracker after SPARK (Guo et al., 2020), RTAA (Jia et al., 2020), CSA (Yan et al., 2020), IoU (Jia et al., 2021) and TrackPGD. The green/red colors correspond to before and after the attack.

Qualitative evaluation. We also illustrate the bounding box evaluation of TransT-SEG tracker (Chen et al., 2023) in several frames from the VOT2018 (Kristan et al., 2018) sequence in Figure 5. The greed/red colors indicates the before and after attack predictions. Our proposed method, TrackPGD, results in misleading the tracker while its perturbation is generated only based on the binary mask. Other attacks such as SPARK and RTAA, which fully built upon bounding box(es) also significantly degrade the target bounding boxes. As the quantitative results in Table 7 indicated, the IoU (Jia et al., 2021) and CSA (Yan et al., 2020) attack, black-box attacks, are behind the white-box methods' performance.