Reproducibility Study on Adversarial Attacks against Robust Transformer Trackers

Anonymous authors
Paper under double-blind review

Abstract

New transformer networks have been integrated into object tracking pipelines and have demonstrated strong performances on the latest benchmarks. Our research is focused on understanding how transformer trackers behave under adversarial attacks and how different attacks perform on tracking datasets as their parameters change. We conducted a series of experiments to evaluate the effectiveness of existing adversarial attacks on object trackers with transformer backbones. Our research characterized the susceptibility of transformer trackers to adversarial perturbations. We assessed the accuracy of the object bounding box and binary mask outputs after applying different attack methods and evaluated the impact of perturbation levels on both white-box and black-box attacks. We used three types of transformer trackers and four attack approaches to conduct the assess performances and robustness. Interestingly, our study found that changing the perturbation level may not significantly affect the overall object tracking results after the attack. Additionally, we observed that, depending on the approach, the sparsity and imperceptibility of adversarial attacks on object trackers may remain stable against perturbation levels, albeit with the cost of generating several highly perturbed frames per sequence. Furthermore, we found that new transformer trackers with stronger cross-attention modelling exhibit greater adversarial robustness on tracking benchmarks. Our results also indicate that new attack methods are required to tackle the latest types of transformer trackers effectively.

1 Introduction

Adversarial perturbations deceive neural networks, leading to inaccurate outputs. Such adversarial attacks have been studied for vision tasks from image classification (Mahmood et al., 2021; Shao et al., 2022) to object segmentation (Gu et al., 2022) and tracker networks (Guo et al., 2020; Jia et al., 2020; Yan et al., 2020; Jia et al., 2021). With transformer networks, the object trackers surpassed the other deep learning-based trackers, showing a very robust performance on the state-of-the-art benchmarks (Kristan et al., 2023). However, the adversarial robustness of these trackers needs to be better studied in the literature.

First transformer trackers relied on relatively light relation modeling (Chen et al., 2021), using a relatively shallow feature extraction and fusion modelling. With the mixed attention module, the MixFormer (Cui et al., 2022) expanded the road for deeper relation modelling. Consequently, the Robust Object Modeling Tracker (ROMTrack) (Cai et al., 2023) proposed variation tokens to capture and preserve the object deformation across frames. Using transformers, especially those with deep relation modeling (Cui et al., 2022; Cai et al., 2023), the object tracker backbones made these models unshakable with many existing attack approaches (Guo et al., 2020; Jia et al., 2020). Indeed, the underlying concept of adversarial attacks against object trackers is to manipulate the tracker's output. By omitting the multi-head pipelines and substituting them with the transformer backbones (Cui et al., 2022; Cai et al., 2023), the tracker's output does not contain the object candidates, classification labels and/or regression labels. As a result, it is not straightforward to transfer adversarial attacks dealing with classification or regression labels (Guo et al., 2020; Jia et al., 2020) on these new transformer trackers. However, a question remains on whether other transferable attacks (Jia et al., 2021; Yan et al., 2020) can represent a sufficient challenge to transformer trackers.

This paper presents a study on the reproducibility of existing attack approaches for transformer trackers. We aim to recreate the attack outcomes on transformer trackers using the VOT2022ST (Kristan et al., 2023) and UAV123 (Mueller et al., 2016) datasets following two different evaluation protocols. We focused on transformer trackers susceptible to adversarial attacks on their prediction outputs, including the object bounding box and binary mask. Then, we analyzed two white-box attacks on a transformer tracker by varying the perturbation levels and checked its vulnerability to different levels of noise. For the black-box setting, we conducted a similar experiment to assess the attack performance on various noise levels by changing the upper bound of the added noise. Based on our study, we found the following:

1. The generated perturbations for those attacks applicable to transformer trackers have more impact on the object mask rather than on the object bounding box.
2. For a specific output, object bounding box or object mask, the transformer tracker with stronger cross-attention modelling presents a more robust performance against the same attack.
3. Although it was demonstrated that adding previous perturbations to the current frame for perturbed search regions generation have an impact on the attacker performance (Guo et al., 2020), these previous perturbations result in more stable performance against perturbation level shifts. For instance, such an increased stability has been observed on SPARK (Guo et al., 2020) attack performance against TransT tracker on UAV123 dataset (Mueller et al., 2016).
4. The SPARK algorithm generates temporally sparse perturbations, meaning that the added perturbation to the search region is small for majority of the frames. It results in imperceptible noise for many frames per video sequence, even though the perturbation level changes to a higher level.
5. Increasing the perturbation level on SPARK results in more super-perturbed regions, i.e. regions with perceptible noise, that grows by the perturbation level.
6. In other methods such as IoU (Jia et al., 2020) and RTAA (Jia et al., 2021) attacks, adding a higher perturbation level generates more perceptible noise for all frames, which damage more the overall tracking performance.
7. The ranking of attack performance is sensitive to the experiment settings, with possibly significant changes in the ranking with different datasets and evaluation protocol.
8. The outcome of the IoU attack is sensitive to its random noise-based initialization. Also, changes in the initial conditions have a smaller effect on transformer trackers with stronger cross-attention and object modelling compared to those with lighter and independent layers of cross-attention.

2 Related Work

Vision transformers have recently been employed in object trackers which enhances the tracking performance. By using transformers in different architectures (Chen et al., 2021; Cui et al., 2022; Chen et al., 2023; Cai et al., 2023), object trackers are capable of inferring object bounding box, binary mask and a prediction score with high robustness values over tracking benchmarks (Kristan et al., 2023; Mueller et al., 2016). These promising results though need to be revisited by assessing transformer trackers in handling the adversarial perturbations. The first transformer tracker, TransT (Chen et al., 2021), used the cross-attention and self-attention blocks to mix features of the moving target and the search region of the tracker. TransT presents a multi-head pipeline with classification and regression heads, unlike other transformer trackers. In TransT-SEG (Chen et al., 2023), the segmentation head is included in the pipeline. The multi-head pipelines follow the siamese-based trackers (Li et al., 2019) in dividing each task from target classification to discriminative tracking processing into individual blocks and fusing the results at the end of the tracker structure. Some light relation modelling layers called the Ego Context Augment (ECA) and Cross Feature Augment (CFA) are introduced by TransT to infer the output from combining the multi-head outputs. Next, the MixFormer (Cui et al., 2022) introduced Mixed Attention Module (MAM) to jointly extract and relate the information from video frames for the object tracking task. By attaching the MixFormer (Cui et al., 2022) tracker to the

AlphaRefine (Yan et al., 2021), MixFormerM (Kristan et al., 2023) can provide an binary object mask per frame for mask oriented evaluations (Kristan et al., 2023). The One-Stream Tracking (OSTrack) (Ye et al., 2022) developed a tracking pipeline that jointly extracts features and models the relation of search region and the template by bidirectional information flows. In contrast, with the Attention in Attention (AiA) mechanism, the AiATrack (Gao et al., 2022) suggested a three stream framework with long-term and short-term cross-attention modules for relation modelling. Following the further relation modeling, the Robust Object Modeling Tracker (ROMTrack) (Cai et al., 2023) is proposed to enable the interactive template learning using both self-attention and cross-attention modules. The ROMTrack has two main streams to learn discriminative features from hybrid (template and search) and inherent template. The newly introduced variation tokens enables ROMTrack with heavier relation modeling rather TransT (Chen et al., 2021) and MixFormer (Cui et al., 2022). The variation token carries the contextual appearance change to tackle object deformation in visual tracking task.

Adversarial attacks against object trackers adopt tracker outputs such as object candidates or classification labels as an attack proxy to generate the adversarial perturbations. For instance, spatial-aware online incremental attack (SPARK) (Guo et al., 2020) creates perturbation by manipulating the classification labels and Intersection of the Union (IoU) (Jia et al., 2021). It is developed to mislead trackers in providing an altered object bounding box based on the predicted bounding box. In Robust Tracking against Adversarial Attack (RTAA) (Jia et al., 2020), both classification and regression labels are used to generate the adversarial samples, similar to SPARK (Guo et al., 2020). In the RTAA (Jia et al., 2020) algorithm, the positive gradient sign is used to generate the adversarial frames. However, in SPARK, the gradient direction is set to negative following the decoupling of norm and direction of gradients in white-box attacks (Rony et al., 2019). Based on the decoupling direction and norm for efficient gradient-based attacks, the direction of the gradient is set in such a way that the generated perturbation has a smaller norm value and greater impact on the results.

Some other attacks, such as the Cooling-Shrinking Attack (CSA) (Yan et al., 2020) is developed specifically to impact the output of siamese-based trackers. In CSA (Yan et al., 2020), two GANs are trained to cool the hottest regions in the final heat-map of siamese-based trackers and shrink the object bounding box. Due to dependency on the siamese-based architecture and loss function, the generalization of the CSA attack (Yan et al., 2020) for other scenarios is harder. The black-box attack, called IoU attack (Jia et al., 2021), adds two types of noise into the frame to make the tracker predict another bounding box rather than the target bounding box. By considering the object motion in historical frames, the direction of added noise is adjusted according to the IoU scores of the predicted bounding box.

3 Transformer Trackers and Adversarial Attacks

In this section, we briefly review the transformer trackers used in our experiments. Also, we explain the adversarial attack methods which are transferred to attack transformer trackers. The codes and network of all of the investigated models are publicly available. The implementation of every tracker and every attack approach are the official repository announced by designers. The networks are also fine turned and released by the authors of original works.

3.1 Transformer Trackers

For the object trackers, we considered three types of the robust single object trackers as follows.

TransT and TransT-SEG In our studies, we used both Transformer Tracker (TransT) (Chen et al., 2021) and TransT with mask prediction ability (TransT-SEG) (Chen et al., 2023). By two discriminative streams and a lightweight cross-attention modeling in the end, the TransT introduced the first transformer-based tracker.

MixFormer and MixFormerM The MixFormer (Cui et al., 2022) is based on a flexible attention operation (MAM) to interactively exploit features and integrate them in a deep layer of the tracker. The MixFormer coupled with the AlphaRefine network has been proposed for the VOT2022 challenge (Kristan et al., 2023) as MixFormerM. It enables the original tracker to provide the object mask as an extra output. In our experiments, we tested both MixFormer and MixFormerM trackers.

ROMTrack The ROMTrack (Cai et al., 2023) is developed to generalize the idea of MixFormer by providing the template learning procedure. The template feature is processed both in self-attention (inherent template) and cross-attention (hybrid template) between template and search regions. This mixed feature avoids distraction in challenging frames and provides a more robust performance compared to TransT and MixFormer.

3.2 Adversarial Attacks

In our study, we examined four attack approaches against object trackers, as follows.

CSA In attention-based siamese trackers (Li et al., 2019), the loss function aims to locate the hottest region in the image where the correlation of the target and that location is the highest among all other regions. Using two GANs, one for the template perturbation and the other for the search region perturbation, the CSA attack (Yan et al., 2020) is developed to firstly cool the hot regions in the end of the network and shrink the object bounding box predicted by the tracker.

IoU The IoU attack (Jia et al., 2021) proposes a black-box setting attack to generate the adversarial frames based on the object motion with the purpose of decreasing the IoU between the predicted bounding box and the target. Two types of noises are added to achieve the final goal of the attack where the noise is bounded to a specific value for L1 norm.

SPARK In SPARK (Guo et al., 2020), the classification and regression labels are manipulated to create the white-box attack. The generated perturbation up to the last 30 frames is accumulated to create the adversarial search regions for the trackers in each time step.

RTAA Using RTAA (Jia et al., 2020), the classification and regression of object candidates are manipulated to generate the adversarial search regions. Only the last frame perturbation is used from the past in the current step of the attack.

4 Investigation

We conducted an analysis to determine how sensitive transformer trackers are to perturbations generated by existing attack methods under various conditions. We compared the difference in performance between the tracker's ability to provide accurate bounding boxes and binary masks by measuring the percentage difference from their original performance on clean data. We evaluated the impact of adversarial attacks on transformer trackers in predicting the object bounding boxes by varying the perturbation levels. Finally, we assessed the performance of the IoU attack when the generated perturbations were bounded at different noise levels. We then discussed the observations we drew from these sets of experiments.

4.1 Adversarial Attacks per Tracker Output

In this section, we have applied adversarial attack techniques against the TransT-SEG (Chen et al., 2023) and MixFormerM (Cui et al., 2022) trackers and compared the results based on different tracking outputs. The objective of this experiment is to determine the difference in each tracking metric before and after the attack when one of the tracker's outputs (bounding box or binary mask) is measured.

Evaluation Protocol We have conducted a baseline experiment for the VOT2022 (Kristan et al., 2023) short-term sub-challenge in two cases: object bounding box (STB) vs. object masks (STS) for target annotation and tracking. The Expected Average Overlap (EAO), accuracy and the anchor-based robustness metrics are calculated in this experiment. The EAO computes the expected value of the prediction overlaps with the ground truth. The accuracy measures the average of overlaps between tracker prediction and the ground truth over a successful tracking period. The robustness is computed as the length of a successful tracking period over the length of the video sequence. The successful period is a period of tracking in which the overlap between the prediction and ground truth is always greater than the pre-defined threshold. The baseline metrics are computed based on the anchor-based protocol introduced in VOT2020 (Kristan et al., 2020). In every video sequence evaluation under this protocol, the evaluation toolkit will reinitialize the

Table 1: Evaluation results of the TransT-SEG (Chen et al., 2021) tracker attacked by different methods on the VOT2022 (Kristan et al., 2023) Short-Term (ST) dataset and protocol for two stacks: bounding box prediction via bounding box annotations (STB) and binary mask prediction via binary mask annotations (STS). The "Clean" values are the original tracker performance without applying any attack.

		EAO			Accuracy			Robustness		
Stack	Method	Clean	Attack	Drop	Clean	Attack	Drop	Clean	Attack	Drop
STB	CSA	0.299	0.285	4.68%	0.472	0.477	-1.06%	0.772	0.744	3.63%
	IoU	0.299	0.231	22.74%	0.472	0.495	-4.87%	0.772	0.569	26.29%
	RTAA	0.299	0.058	83.28%	0.472	0.431	8.69%	0.772	0.157	79.66%
	SPARK	0.299	0.012	95.99%	0.472	0.244	48.30%	0.772	0.051	93.39%
STS	CSA	0.500	0.458	8.40%	0.749	0.736	1.73%	0.815	0.779	4.42%
	IoU	0.500	0.334	33.20%	0.749	0.710	5.21%	0.815	0.588	27.85%
	RTAA	0.500	0.067	86.60%	0.749	0.533	28.84%	0.815	0.146	82.08%
	SPARK	0.500	0.011	97.80%	0.749	0.266	64.48%	0.815	0.042	94.84%

tracker from the next anchor of the data to compute the anchor-based metrics wherever the tracking failure happens. For visualization usage, we employed some video sequences from DAVIS2016 (Perazzi et al., 2016) dataset.

Attacks Setting Four adversarial attacks are employed in this experiment namely: CSA (Yan et al., 2020), IoU (Jia et al., 2021), SPARK (Guo et al., 2020), and RTAA (Jia et al., 2020). However, not all of the attacks are applicable to both trackers. The SPARK and RTAA attacks manipulate the object candidates list, which includes classification labels and/or regression labels. If the tracker does not infer the candidates in the output, these attacks cannot be applied on, such as MixFormerM (Cui et al., 2022) that only outputs the predicted bounding box. In this experiment, we generated the perturbations using SPARK and RTAA attacks, i.e. white-box attacks, against trackers. However, the perturbation of CSA (Yan et al., 2020) is created by two GANs and passing the image into the SiamseRPN++ (Li et al., 2019) tracker to generate the adversarial loss depending on the SiamseRPN++ loss. Therefore, the CSA is a transferred black-box attack for both TransT (Chen et al., 2021) and MixFormer (Cui et al., 2022) trackers. The IoU method (Jia et al., 2021) is also a black-box approach that can perturb the whole frame using the tracker prediction for several times.

Results The following is a summary of the results obtained from an experiment conducted on the VOT2022 (Kristan et al., 2023) dataset using the TransT-SEG tracker (Chen et al., 2023) after adversarial attacks. The results are shown in Table 1 for both STB and STS cases. The most powerful attack against TransT-SEG (Chen et al., 2023) in all three metrics was found to be SPARK (Guo et al., 2020). The evaluation metrics revealed that the object binary mask was more affected by the adversarial attacks than the object bounding box.

It was observed that the CSA (Yan et al., 2020) attacks poorly degraded the outputs in evaluation metrics, except for the accuracy of the STB case. However, it was surprising to note that there was a negative difference in the accuracy metric after the CSA and IoU attack (Jia et al., 2021). Specifically, in the STB case, after the adversarial attacks, the accuracy of TransT-SEG (Chen et al., 2023) decreased by -1.06% for the CSA attack (Yan et al., 2020) and -4.87% for the IoU attack (Jia et al., 2021). This improvement was also observed for MixFormer (Cui et al., 2022) in Table 2 for EAO and robustness metrics after the CSA attack in the STB case and accuracy after the CSA in the STS case of the experiments.

According to Table 2, the most powerful attack against MixFormerM (Cui et al., 2022) is the IoU attack (Jia et al., 2021). Even after the IoU attack (Jia et al., 2021), the adversarial perturbation slightly improves accuracy. The EAO metric in the evaluation of the bounding box and binary mask is the most affected. When compared to the corresponding metric for TransT-SEG (Chen et al., 2023) as shown in Table 1, the IoU attack (Jia et al., 2021) had a more significant impact on binary mask creation for MixFormerM than

Table 2: Evaluation results of the MixFormerM (Cui et al., 2022) tracker attacked by different methods on the VOT2022 (Kristan et al., 2023) Short-Term (ST) dataset and protocol for two stacks: bounding box prediction via bounding box annotations (STB), binary mask prediction via binary mask annotations (STS). The "Clean" values are the original tracker performance without applying any attack.

		EAO			Accuracy			Robustness		
Stack	Method	Clean	Attack	Drop	Clean	Attack	Drop	Clean	Attack	Drop
STB	CSA	0.303	0.308	-1.65%	0.479	0.478	0.21%	0.780	0.791	-1.41%
	IoU	0.303	0.246	18.81%	0.479	0.458	4.38%	0.780	0.665	14.74%
STS	CSA	0.589	0.562	4.58%	0.798	0.803	-0.63%	0.880	0.857	2.61%
	IoU	0.589	0.359	39.05%	0.798	0.660	17.30%	0.880	0.677	23.07%

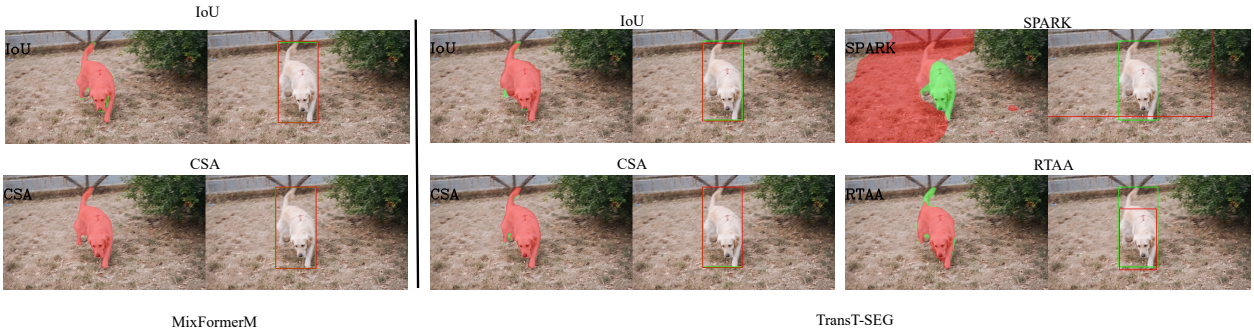


Figure 1: Mask vs. bounding box predictions as the output of transformer trackers, MixFormerM (Cui et al., 2022) and TransT-SEG Chen et al. (2023), while the adversarial attacks applied to perturb the input frame/search region. The TransT-SEG tracker's outputs harmed by the white-box methods, SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020), more than black-box attacks, IoU (Jia et al., 2021) and CSA (Yan et al., 2020).

for TransT-SEG (Chen et al., 2023). However, this result was reversed when the object bounding box was evaluated in both trackers. For this point, it is important to mention that tracking and segmentation is performed by two different networks in MixFormerM (Cui et al., 2022). Therefore, the object bounding box evaluation is the assessment of the tracker's network while the binary mask evaluation is the assessment the segmentation network (Yan et al., 2021). In contrast, the TransT-SEG (Chen et al., 2023) tracker performs both tracking and segmentation by a single transformer tracker.

Figure 1 demonstrates a sample frame perturbed by different attacks for MixFormerM and TransT-SEG trackers and the output results are depicted on the video frame. As the quantitative results indicated in Tables 1 and 2, the white-box attacks (SPARK and RTAA) have harmed the binary mask and bounding box more than the black-box attacks (IoU and CSA).

4.2 Adversarial Attacks per Perturbation Level

We test the effect of the perturbation levels on adversarial attack performance against transformer trackers. In white-box attacks such as SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020), the generated perturbation is used to update the frame patch in each attack step. The overview of pseudocode of these two attacks is presented in Algorithms 1 and 2 where the ϕ_ϵ indicates the operation of clipping the frame patch to the ϵ-ball. The α is the applied norm of gradients, and I is the search region. Although there are several differences in both settings, there is one similar step to generate the adversarial region from the input image gradient and previous perturbation(s). Line 4 of the RTAA pseudocode and line 6 of the SPARK pseudocode change the image pixels based on the computed gradients. By adjusting the ϵ-ball, the

Algorithm 1 RTAA (Jia et al., 2020) algorithm as the adversarial attack for object trackers

$\mathcal{P} \leftarrow \mathcal{P}(t-1)$ ▷ Initialize with perturbation map of previous frame
$I^{\text{adv}} \leftarrow I$ ▷ Initialize with clean current frame
for $i = 1, \ldots, i^{\max}$ **do**
  $I^{\text{adv}} \leftarrow I^{\text{adv}} + \phi^{\epsilon}\left(\mathcal{P} + \alpha \operatorname{sign}(\nabla_{I^{\text{adv}}} \mathcal{L})\right)$ ▷ Application of adversarial gradient descent
  $I^{\text{adv}} \leftarrow \max(0, \min(I^{\text{adv}}, 255))$ ▷ Clamp image values in [0, 255]
  $\mathcal{P} \leftarrow I^{\text{adv}} - I$ ▷ Update perturbation map
Return $I^{\text{adv}}, \mathcal{P}$ ▷ Return adversarial image and corresponding perturbation map

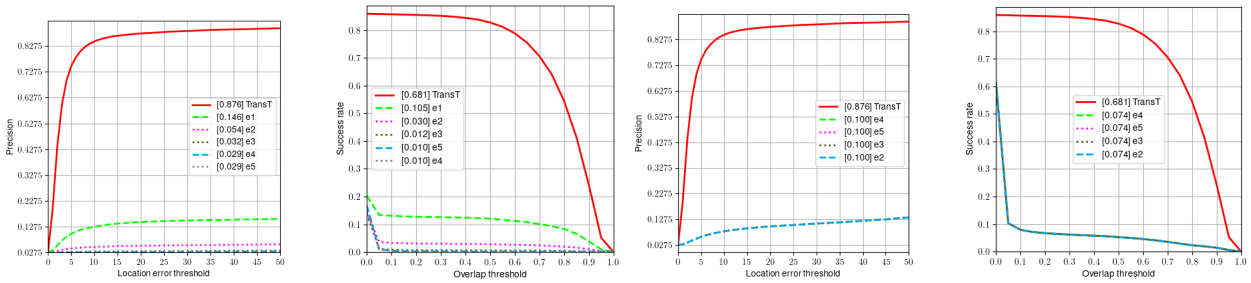
Algorithm 2 SPARK (Guo et al., 2020) algorithm as the adversarial attack for object trackers

$\mathcal{P} \leftarrow \mathcal{P}(t-1)$ ▷ Initialize with perturbation map of previous frame
$\mathcal{S} \leftarrow \sum_{i=1}^{K} \mathcal{P}(t-i)$ ▷ Sum of perturbation maps of last K frames
$I^{\text{adv}} \leftarrow I$ ▷ Initialize with clean current frame image
for $i = 1, \ldots, i^{\max}$ **do**
  $I' \leftarrow I^{\text{adv}}$ ▷ Get a copy of current adversarial image
  $I^{\text{adv}} \leftarrow I' + \phi^{\epsilon}\left(\mathcal{P} - \alpha \operatorname{sign}(\nabla_{I'} \mathcal{L})\right) + \mathcal{S}$ ▷ Application of adversarial gradient descent
  $I^{\text{adv}} \leftarrow \max(0, \min(I^{\text{adv}}, 255))$ ▷ Clamp image values in [0, 255]
  $\mathcal{P} \leftarrow I^{\text{adv}} - I' - \mathcal{S}$ ▷ Update perturbation map
$\mathcal{S} \leftarrow \sum_{i=1}^{K} \mathcal{P}(t-i)$ ▷ Update the sum of perturbation maps
$I^{adv} \leftarrow I + \mathcal{S}$ ▷ Generate the current adversarial frame
Return $I^{\text{adv}}, \mathcal{P}$ ▷ Return adversarial image and corresponding perturbation map

performance of attacks is evaluated to demonstrate the power of each adversarial idea for object trackers. The plus or minus sign of the gradient sign corresponds to the decoupling direction and norm research in the gradient-based adversarial attack (Rony et al., 2019). For instance, SPARK (Guo et al., 2020) uses minus, while RTAA (Jia et al., 2020) sums up the sign of gradients with image values. Furthermore, note that in the original papers of SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020), the attack parameters may have different names than those we used in this paper. Our goal was to unify their codes and approach into principle steps that make comparison more accessible for the audience.

An important aspect of the SPARK algorithm, mentioned in (Guo et al., 2020), is its regularization term. This feature is convenient for maintaining sparse and imperceptible perturbations (Guo et al., 2020). The regularization term involves adding the $L_{2,1}$ Norm of previous perturbations to the adversarial loss, which helps generate sparse and imperceptible noises. We generated examples of SPARK perturbation (Guo et al., 2020) versus RTAA (Jia et al., 2020) perturbations to verify this claim.

Evaluation Protocol The test sets of the experiments on the perturbation level changes are the UAV123 dataset (Mueller et al., 2016) and VOTST2022 (Kristan et al., 2023). The UAV123 dataset comprises 123 video sequences with natural and synthetic frames in which an object appears and disappears from the frame captured by a moving camera. We calculate success and precision rates across various thresholds under the One Pass Evaluation (OPE) protocol. In this setup, the object tracker is initialized using the first frame and the corresponding bounding box. Subsequently, the tracker is evaluated for each frame's prediction for the rest of the video sequence. Precision is measured by calculating the distance between the center of the ground truth's bounding box and the predicted bounding box. The precision plot shows the percentage of bounding boxes that fall within a given threshold distance. The success rate is computed based on the Intersection over Union (IoU) between the ground truth and predicted bounding boxes. The success plot is generated by considering different thresholds over IoU and computing the percentage of bounding boxes that pass the given threshold. Furthermore, we computed the L1 norm and structural similarity (SSIM) (Wang et al., 2004) as the measurements of sparsity and imperceptibility of the generated perturbations per attack. We chose some video frames from the VOT2022ST (Kristan et al., 2023) dataset to visualize these metrics per frame.



(a) (b) (c) (d)

Figure 2: The precision and success plots related to the TransT (Chen et al., 2021) performance after RTAA (Jia et al., 2020) (a, b) and SPARK (Guo et al., 2020) (c,d) attack under different levels of noise on UAV123 (Mueller et al., 2016) dataset. The average score for each metric is shown in the legend of the plots. The 'red' plot is the original TransT performance without any attack applied on the tracker. The e's are corresponded to ϵ's in our experiment, changing from $e_1 = 2.55$ to $e_5 = 40.8$ to assess the TransT performance after the white-box attacks under various perturbation levels. The SPARK performances per perturbation level shifts did not change on UAV123dataset as one can observe the SPARK curves are overlieded.

Attacks Setting The SPARK (Guo et al., 2020) and RTAA (Jia et al., 2020) approaches applied on the TransT tracker (Chen et al., 2021) are assessed in this experiment using the OPE protocol. Both attacks generate the perturbed search region over a fixed number of iterations (10). While the step size α for the gradient's update is 1 for RTAA and 0.3 for SPARK. We used five levels of perturbation $\epsilon \in \{2.55, 5.1, 10.2, 20.4, 40.8\}$ to compare its effects on the TransT (Chen et al., 2021) performance on UAV123 (Mueller et al., 2016) and VOT2022ST (Kristan et al., 2023) datasets. The ϵ's are selected as a set of coefficients $\{0.01, 0.02, 0.04, 0.08, 0.16\}$ of the maximum pixel value 255 in an RGB image. It is worth mentioning that the ϵ for both attacks are set to 10 in their original settings. Therefore, the original performance of each attack is very close to the $\epsilon_3 = 10.2$ perturbation level.

Results Figure 2 shows the performance of TransT (Chen et al., 2021) under RTAA (Jia et al., 2020) and SPARK (Guo et al., 2020) attacks with different perturbation levels. The red curve indicates the clean performance of the tracker before applying any attacks. The other perturbation levels are demonstrated with different colors. Unlike classification networks with transformer backbones (Shao et al., 2022), the transformer tracker performances after the RTAA attack (Jia et al., 2020) using different ϵ's are minimally different but not after SPARK attacks (Guo et al., 2020). Adversarial perturbation methods against trackers use the previous perturbation to be added to the current frame. This setting may remove the sensitivity of the attack methods in the perturbation levels. In the RTAA attack (Jia et al., 2020), only one last perturbation is added to the current frame. In contrast, the SPARK (Guo et al., 2020) uses the previous perturbations in each time step for the last $K = 30$ frames, which reduces the sensitivity of the output to small changes in the inputs. For perturbation levels $\{\epsilon_3, \epsilon_4, \epsilon_5\}$, RTAA's performance (Jia et al., 2020) remains the same, whereas using smaller levels affects its performance. It is noteworthy that RTAA (Jia et al., 2020) outperforms SPARK (Guo et al., 2020) on UAV datasets (Mueller et al., 2016) in almost every perturbation level except for the most minor level $\epsilon_1 = 0.01$ in which SPARK is the stronger attack.

In the main paper of SPARK (Guo et al., 2020), it has been mentioned that the technique generates a temporally sparse and imperceptible noise. Figure 3 displays various examples of perturbed search regions and perturbation maps produced by the TransT tracker after applying the SPARK attack. Upon applying the attack, we noted that some frames, like frame number "7" of the 'bubble' sequence and the first two rows of Figure 3, generated search regions and perturbation maps with fixed values for imperceptibility (SSIM metric) and sparsity (L1 norm). Even by increasing the perturbation level, some frames retained the same level of imperceptibility and sparsity. However, there were also instances of super-perturbed search regions per video sequence, where the noise was noticeable, and the L1 norm had a high value, as shown in the last two rows of Figure 3. We consider a perturbed search region super-perturbed when the imperceptibility of the region is lower than 50%. The SPARK algorithm generates the most imperceptible noise with a constant

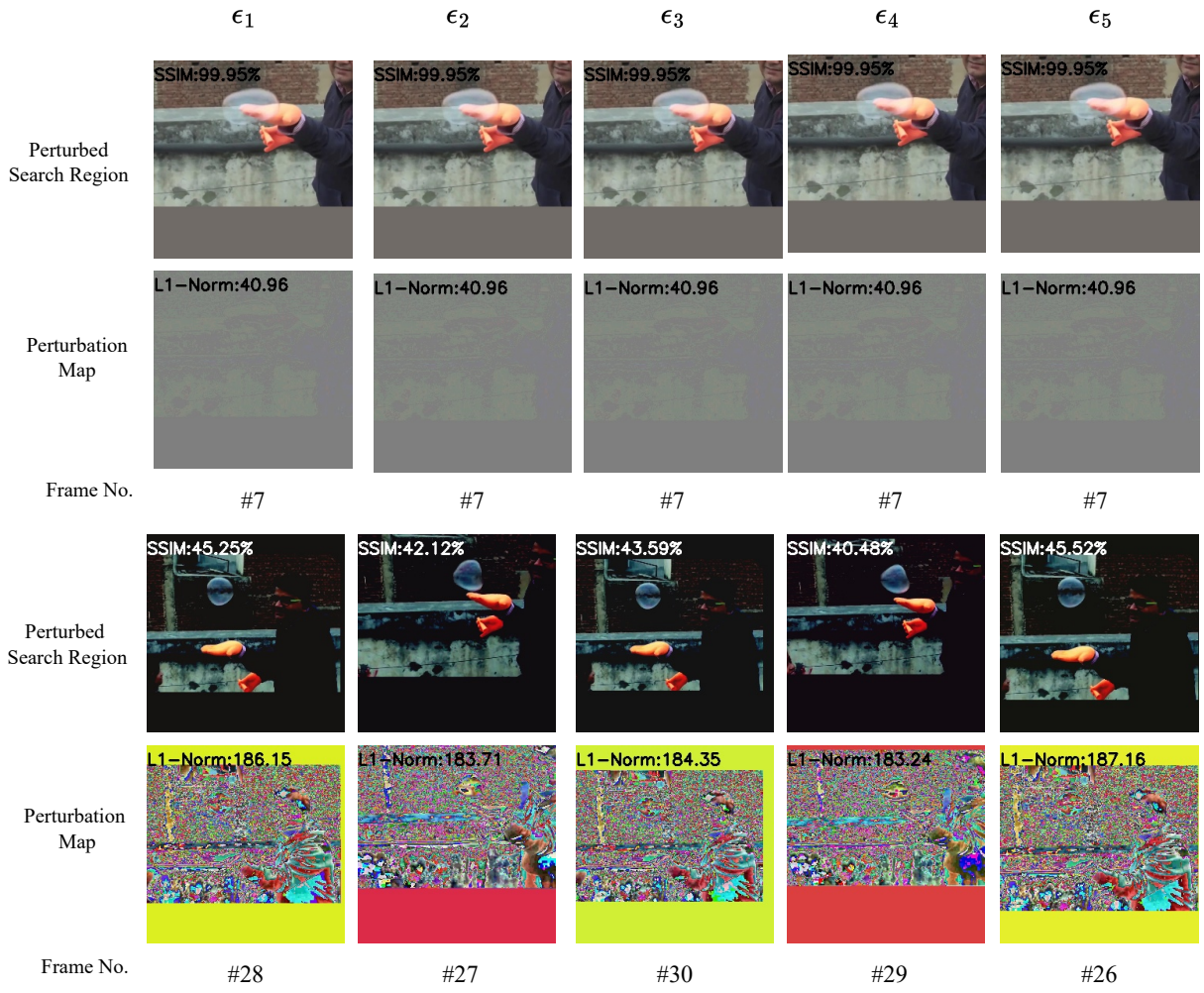


Figure 3: The search regions related to the "bubble" sequence in the VOT2022ST dataset (Kristan et al., 2023) after applying SPARK (Guo et al., 2020) attack on TransT (Chen et al., 2021) tracker. The perturbed search region is labeled with the SSIM (Wang et al., 2004) measured between search regions before and after the attack. The perturbation maps, following the work of (Yan et al., 2020), are created to demonstrate the added noise in colors. The L1 norm for perturbation maps are calculated to show the perturbation density/sparsity.

and high SSIM value of 99.95% and sparse noise with L1 norm of 40.96 in all of the perturbation levels given the same frame. This fixed number of L1 norm is also the result of the regularization term discussed in the SPARK paper (Guo et al., 2020). This stability of SSIM and L1 norm have been repeated for many frames of the "bubble" sequence for SPARK attack (Guo et al., 2020) while in some frames, the imperceptibility and sparsity are not stable per perturbation levels.

In Figure 3, we indicate some super-perturbed regions with their perturbation maps per perturbation level. Interestingly, as we increase the perturbation levels, the number of super-perturbed regions also increases. In the attack settings, the perturbation of previous frames considered in the loss function is erased every 30 frames for RTAA (Jia et al., 2020) and SPARK (Guo et al., 2020) algorithms. Table 3 provides information about the number of highly perturbed frames during a video sequence and the average imperceptibility (SSIM) and sparsity (L1 norm) scores. With higher levels of perturbations, more frames become highly perturbed, resulting in a greater L1 norm of the perturbations. Furthermore, in the lower perturbations

Table 3: The perturbation levels versus number of highly perturbed search regions generated by the SPARK algorithm (Guo et al., 2020) applied on TransT (Chen et al., 2021) tracker. The SSIM and L1 norm are computed as the average number of highly perturbed regions on the "bubble" sequence of VOT2022 (Kristan et al., 2023) dataset.

ϵ	No. of frames	SSIM%	L1 norm
2.55	#7	36.86	176.04
5.1	#7	40.96	181.86
10.2	#13	41.08	181.33
20.4	#13	41.97	182.53
40.8	#14	42.53	183.98

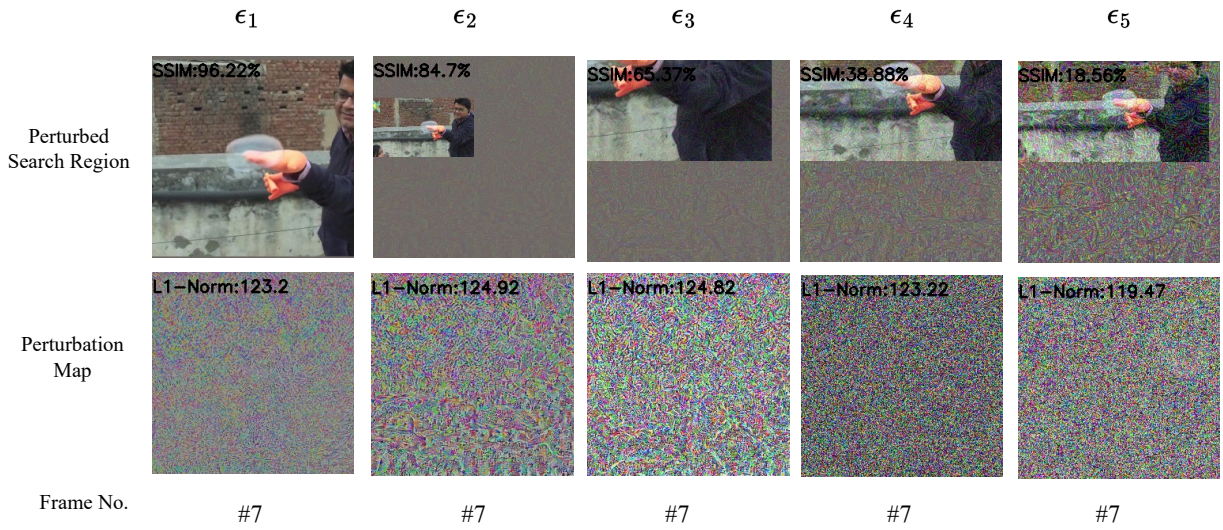


Figure 4: The search regions related to the "bubble" sequence in the VOT2022ST dataset (Kristan et al., 2023) after applying RTAA (Jia et al., 2020) attack on TransT (Chen et al., 2021) tracker. The perturbed search region is labeled with the SSIM (Wang et al., 2004) measured between search regions before and after the attack. The perturbation maps, following the work of (Yan et al., 2020), are created to demonstrate the added noise in colors. The L1 norm for perturbation maps are calculated to show the perturbation density (i.e. sparsity).

levels, the highly perturbed search regions generate more perceptible noise, i.e. the imperceptibility of generated perturbations have grown by boosting the perturbation level.

For the RTAA (Jia et al., 2020) attack applied on the TransT (Chen et al., 2021) tracker, whenever the perturbation level boosts the imperceptibility and sparsity declines. Figure 4 demonstrates the result of applying RTAA against TransT tracker for the same frame #7 of the 'bubble' sequence. The RTAA attack perturbs search regions with higher SSIM values at the lowest ϵ level, i.e. the first level $\epsilon = 2.55$. By increasing the perturbation levels, the perceptibly of the RTAA perturbation has been increased while the sparsity changes are small.

4.3 Adversarial Attack per Upper Bound

We conducted an experiment to test the vulnerability of the IoU attack (Jia et al., 2021) to noise bounding, using different upper bounds. The IoU method (Jia et al., 2021) is a black-box attack that misleads trackers by adding various noises to the frame using object bounding boxes. The essential steps of the IoU attack (Jia et al., 2021) involve creating two levels of noise perturbations: orthogonal and normal direction noises. Our

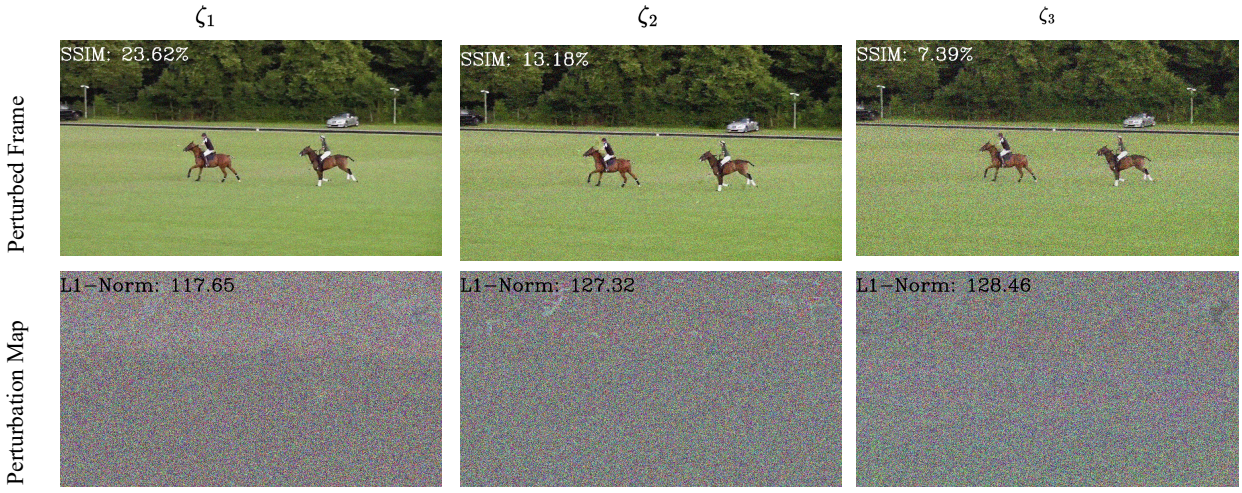


Figure 5: The perturbed frames and perturbation maps generated by the IoU method (Jia et al., 2021) against ROMTrack (Cai et al., 2023) using three upper bounds of $\zeta \in \{8k, 10k, 12k\}$. The imperceptibility and L1 norm of the generated perturbations are shown in the frames representing the noise imperceptibility and sparsity of perturbation maps.

study aims to manipulate the attack settings in the second part of perturbation generation, which is in the normal direction.

Dataset and Protocol The performance of the IoU attack is assessed against ROMTrack (Cai et al., 2023) on UAV123 (Mueller et al., 2016) dataset using the OPE protocol. The success rate, precision rate, and normalized precision rate are computed to compare the results. For this experiment, we report the average of the success rate called Area Under Curve (AUC), as well as the average precision and norm precision on the thresholds.

Attack Setting The IoU method (Jia et al., 2021) is a black-box attack on object trackers. It adds two types of noise to the frames: one in the tangential direction and the other in the normal direction. In the original setting, there was no limit on the number of times that noise could be added in the normal direction. This noise was limited only by the upper bound ζ and the S_{IoU} value in the original setting. The S_{IoU} value is the weighted average of two computed IoU values: 1) the IoU of the noisy frame prediction and the first initialized frame I_1^{adv} in the attack algorithm, and 2) the IoU of the noisy frame prediction and the last frame bounding box in the tracking loop. In our experiment, we set a limit of 10 steps in the algorithm's last loop to reduce the processing time, especially for the larger upper bound ζ values. We tested the IoU attack under three upper bounds: $\zeta \in \{8000, 10000, 12000\}$. The middle value of $\zeta = 10000$ corresponds to the original setting of the IoU attack (Jia et al., 2021).

Results The images in Figure 5 display the results of the IoU attack (Jia et al., 2021) against ROMTrack (Cai et al., 2023) under various upper bounds for a single frame. The L1 norm of the perturbation has increased as the upper bounds were raised. Additionally, the imperceptibility, measured by the SSIM values, decreased as the perturbations became more severe. Since the IoU attack starts by generating some random noise, it is highly dependant on the initialization points. For some cases, the algorithm did not process a single video sequence even after 48 hours. One solution that worked for proceeding was to stop the processing without saving any results about the current sequence to restart the evaluation. After re-initialization, the attack began from another random point (noise) and it proceeded to the next sequence in less than 2 hour.

The results of the attack on ROMTrack (Cai et al., 2023) using the IoU method (Jia et al., 2021) with different upper bounds are presented in Table 4. It is clear that a higher upper bound leads to a more effective attack across all metrics. Despite the most substantial level of perturbation using the IoU method (Jia et al., 2021) resulting in an 8.79% decrease in the AUC metric, this outcome is insignificant. As shown in Figure 5,

Table 4: Evaluation results of the ROMTrack (Cai et al., 2023) attacked by the IoU approach (Jia et al., 2021) on the UAV123 (Mueller et al., 2016) dataset and protocol for three different upper bounds on the added noise in normal direction up to 10 processing steps.

	AUC			Precision			Norm Precision		
ζ	Original	Attack	Drop	Original	Attack	Drop	Original	Attack	Drop
8k	69.74	66.85	4.14%	90.83	89.31	1.67%	85.30	83.00	2.70%
10k	69.74	65.46	6.14%	90.83	87.81	3.32%	85.30	81.73	4.18%
12k	69.74	63.61	8.79%	90.83	86.31	4.98%	85.30	79.71	6.55%

increasing ζ generates a perceptible perturbation with a lower SSIM to the original frame, resulting in a more noisy frame that damages the tracking performance of ROMTrack (Cai et al., 2023) even more. However, the robust tracking performance is not affected more than 9% per metric, even in the highest perturbation level. In other words, ROMTrack (Cai et al., 2023) demonstrates good adversarial robustness against IoU attack on UAV123 dataset.

5 Discussion and Conclusion

We conducted a study on adversarial attack methods for object trackers with the aim of testing their impact on transformer trackers. Our paper includes experiments on tracking datasets, trackers, and attack settings. We evaluated three transformer trackers, ranging from light to deep relation modelling, on two tracking benchmarks. At the same time, we perturbed their inputs using four attack methods, including two white-box and two black-box approaches.

Our analysis revealed that binary mask prediction is more affected than object-bounding box predictions for the same tracker. White-box attacks were found to be stronger than black-box attacks against all trackers. For a single attack, the transformer tracker with deeper relation modelling showed greater adversarial robustness. The level of perturbation shifts may or may not change the overall tracking results depending on the attack approach. A higher level of perturbation always led to a greater number of super-perturbed search regions with greater L1 norm for perturbations in both white-box and black-box settings. In the case of black-box attacks dealing with random noises, the initialization point is a critical aspect. The IoU method becomes highly time-consuming due to inappropriate initialization.

The most effective attack idea may vary depending on the benchmark, as the strongest attack on the VOT2022 benchmark was SPARK, whereas RTAA outperformed the SPARK attack on the UVA123 benchmark. The only transferrable attacks for transformers with deep relation modelling are the black-box, IoU and CSA methods. Among black-box methods, the IoU attack outperformed CSA for TransT-SEG and MixFormerM trackers. However, the effect of the IoU method against ROMTrack is trivial. The ROMTrack and MixFormerM bounding box predictions were harmed by the IoU method up to 9 % and 18 % on UAV123 and VOT2022 datasets, respectively. It indicates that these trackers were not being challenged enough with existing applicable attack methods.

We tested two tracking protocols, one-pass evaluation and anchor-based short-term tracking, using the TransT, MixFormer, and ROMTrack trackers. Our study revealed that black-box attacks are the only applicable methods on transformer trackers with deep relation modelling. However, they are not challenging these trackers as much as white-box attacks challenge TransT. We also discovered that changes in the perturbation level do not necessarily affect the tracking performance over a tracking dataset. The sparsity and imperceptibility of the perturbations can be managed by advising a proper loss function. Finally and above all, further research on the adversarial robustness of transformer trackers is needed for a more in-depth exploration.

References

Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *IEEE Conf. on Comput. Vis.*, 2023.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2021.

Xin Chen, Bin Yan, Jiawen Zhu, Huchuan Lu, Xiang Ruan, and Dong Wang. High-performance transformer tracking. *IEEE Trans. on Pattern Analy. and Machine Intel.*, 45(7):8507–8523, 2023.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2022.

Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *IEEE Conf. Euro. Conf. Comput. Vis.*, 2022.

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip H. S. Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *IEEE Conf. Euro. Conf. Comput. Vis.*, 2022.

Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu. Spark: Spatial-aware online incremental attack against visual tracking. In *IEEE Conf. Euro. Conf. Comput. Vis.*, 2020.

Shuai Jia, Chao Ma, Yibing Song, and Xiaokang Yang. Robust tracking against adversarial attacks. In *IEEE Conf. Euro. Conf. Comput. Vis.*, 2020.

Shuai Jia, Yibing Song, Chao Ma, and Xiaokang Yang. Iou attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2021.

Matej Kristan, Aleš Leonardis, and Jiří Matas. The eighth visual object tracking vot2020 challenge results. In *IEEE Conf. Euro. Conf. Comput. Vis. Worksh.*, 2020.

Matej Kristan, Aleš Leonardis, Jiří Matas, and Michael Felsberg. The tenth visual object tracking vot2022 challenge results. In *IEEE Conf. Euro. Conf. Comput. Vis. Worksh.*, 2023.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2019.

Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *IEEE Conf. on Comput. Vis.*, 2021.

Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *IEEE Conf. Euro. Conf. Comput. Vis.*, 2016.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2016.

Jerome Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2019.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *Trans. on Machine Learning Research*, 2022. ISSN 2835-8856.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Process.*, 13(4):600–612, 2004.

Bin Yan, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Cooling-shrinking attack: Blinding the tracker with imperceptible noises. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2020.

Bin Yan, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In *IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2021.

Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *IEEE Conf. Euro. Conf. Comput. Vis.*, 2022.