NumbOD: A Spatial-Frequency Fusion Attack Against Object Detectors

(/pdf?id=lkyxpj9FSn)

Ziqi Zhou (/profile?id=~Ziqi_Zhou2), Bowen Li (/profile?id=~Bowen_Li16), Yufei Song (/profile?id=~Yufei_Song3), Shengshan Hu (/profile?id=~Shengshan_Hu1), Wei Wan (/profile?id=~Wei_Wan2), Leo Yu Zhang (/profile?id=~Leo_Yu_Zhang1), Dezhong Yao (/profile?id=~Dezhong_Yao1), Hai Jin (/profile?id=~Hai_Jin1)

03 Apr 2024 (modified: 21 Jul 2024) Submitted to MM2024 Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=lkyxpj9FSn) BibTeX

Abstract:
With the advancement of deep learning, object detectors (ODs) with various architectures have achieved significant success in complex scenarios like autonomous driving. Previous adversarial attacks against ODs have been focused on desinging customized attacks targeting their specific structures (e.g., NMS and RPN), yielding some results but simultaneously constraining their scalability. Moreover, most efforts against ODs stem from image-level attacks originally designed for classification tasks, resulting in redundant computations and disturbances in object-irrelevant areas (e.g., backgrounds). Consequently, how to design a universal and efficient attack to comprehensively evaluate the vulnerabilities of ODs remains challenging and unresolved.

In this paper, we propose NumbOD, a brand-new spatial-frequency fusion attack against object detectors, aimed at disrupting object detection within images. We directly leverage the features output by the OD without relying on its any internal structures to craft adversarial examples. Specifically, we first design a dual-track attack target selection strategy to select high-quality bounding boxes from OD outputs for targeting. Subsequently, we employ directional perturbations to shift and compress predicted boxes and disrupt classification results to deceive ODs. Additionally, we focus on manipulating the high-frequency components of images to confuse ODs' attention on critical objects, thereby enhancing the attack efficiency. Our extensive experiments on eight ODs and two datasets show that our NumbOD achieves powerful attack performance and high stealthiness.

Primary Subject Area: [Experience] Multimedia Applications
Secondary Subject Area: [Content] Vision and Language
Relevance To Conference: This work introduces NumbOD, a pioneering approach for launching adversarial attacks against object detectors, which are a crucial component in multimedia and multimodal processing. They are extensively employed for identifying and localizing specific objects or targets within images or videos, forming the backbone of many multimedia analysis and search systems. However, this paper underscores a significant vulnerability in object detection systems—their susceptibility to adversarial examples. This research not only reveals a critical security flaw in object detection systems but also lays the groundwork for developing more robust, attack-resistant object detection algorithms. Therefore, the contributions of this paper are twofold: first, it advances the understanding of inherent vulnerabilities in object detection-based multimedia analysis and search systems; second, it drives the development of more secure and reliable systems, crucial for the integrity and trustworthiness of multimedia and multimodal processing applications.
Submission Guide: Yes
Open Discussion: Yes
Author Registration Confirmation: Yes
Reviewer Participation: Yes

Supplementary Material: zip (/attachment?id=lkyxpj9FSn&name=supplementary_material)
Submission Number: 1904

Search keywords...

Sort: Newest First

- = ≡

Everyone | Program Chairs | Submission1904 Authors | Submission1904...
Submission1904 Area... | Submission1904... | Submission1904... | Submission1904...
Submission1904... | Submission1904... | ✖

7 / 7 replies shown

Add: **Withdrawal**

Paper Decision

Decision Program Chairs 18 Jul 2024, 13:31 (modified: 21 Jul 2024, 16:04) Program Chairs, Authors
Revisions (/revisions?id=Y8wCg2unPW)

Decision: Reject

Meta Review of Submission1904 by Area Chair mUmC

Meta Review Area Chair mUmC 01 Jul 2024, 22:51 (modified: 21 Jul 2024, 13:51)
Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs
Revisions (/revisions?id=PkEhHJCBnM)

Metareview:
The paper proposes NumbOD, a spatial-frequency fusion-based adversarial attack against object detection models. The paper is written clearly and is easy to understand, and the technique sounds reasonable. However, there are many issues, such as the experimental results not supporting the article well and a lack of comparison with state-of-the-art attacks. After reading the authors' responses and the reviewers' comments, I believe this paper is not ready to be accepted at this time. I hope the authors can thoroughly revise the paper, as it has a good chance of being accepted in the future.

Recommendation: Reject
Confidence: 5

Rebuttal by Authors

Rebuttal
Authors (Ziqi Zhou (/profile?id=~Ziqi_Zhou2), Bowen Li (/profile?id=~Bowen_Li16), Leo Yu Zhang (/profile?id=~Leo_Yu_Zhang1), Yufei Song (/profile?id=~Yufei_Song3), +4 more (/group/info?id=acmmm.org/ACMMM/2024/Conference/Submission1904/Authors))
18 Jun 2024, 11:52 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Rebuttal Pdf: pdf (/attachment?id=q0utUx4p6u&name=rebuttal_pdf)

Official Review of Submission1904 by Reviewer imtC

Official Review Reviewer imtC 26 May 2024, 20:24 (modified: 22 Jul 2024, 21:47)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer imtC, Authors

Revisions (/revisions?id=6kVUWJ1BI5)

Summary:
In this paper, the authors present NumbOD, a novel attack framework designed to degrade the performance of object detectors (ODs). Unlike previous adversarial attacks that typically target specific components of ODs or are adapted from image-level classification attacks, NumbOD offers a universal and efficient approach. This method leverages the features directly output by ODs and employs a dual-track attack strategy to select high-quality bounding boxes, designing a spatial-frequency fusion attack for targeted perturbation. Extensive experiments demonstrate the effectiveness and stealthiness of NumbOD in compromising various OD architectures.

Strengths:
1. This paper introduces an innovative method, the spatial-frequency fusion attack, which effectively disrupts object detectors' ability to identify objects in input images. The approach showcases substantial technical innovation.
2. NumbOD does not rely on internal components of object detectors, such as NMS and RPN, making it widely applicable to different architectures. It provides valuable insights into adversarial examples for object detection and significantly advances the field.
3. The proposed method exhibits robust attack capabilities, demonstrated through both qualitative and quantitative experiments. Comparative analysis indicates that this method surpasses existing attack techniques in performance.
4. The experiments in this paper are well-designed and comprehensive, sufficiently proving the method's effectiveness. Notably, defense experiment results show that NumbOD can overcome mainstream defense methods, further demonstrating its superiority.

Limitations:
1. The authors should adjust the font size of the text in the figures within the paper to enhance readability for the audience.
2. It is suggested that the authors further elaborate on the reasons for selecting the three stealthiness metrics and investigate their interrelationships.
3. There is a missing comma on line 198.

Suitability: Definitely suitable, e.g., contributes significantly to multimedia/multimodal processing
Rating: Weak Accept - I would need strong arguments to reject this submission.
Confidence: Confident
Final Rating: Weak Accept - I would need strong arguments to reject this submission.
Final Rating Justification:
The reviewer would like to appreciate the responses from the authors and will keep my original score.

Official Review of Submission1904 by Reviewer wVFp

Official Review Reviewer wVFp 23 May 2024, 15:38 (modified: 22 Jul 2024, 21:47)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer wVFp, Authors
Revisions (/revisions?id=vICM8WqBjL)

Summary:
The paper proposes NumbOD, a spatial-frequency fusion-based adversarial attack against object detection models.NumbOD specializes in white-box scenarios and can be applied to different detector architectures. Specifically, NumbOD first utilizes a two-track attack target selection strategy to select high-quality bounding boxes as attack targets, and then utilizes directional perturbation to corrupt the classification results. Meanwhile, NumbOD improves the attack efficiency by processing the high-frequency components of the image.

Strengths:
1. The paper is organized and easy to follow.
2. Overall, this paper is technically sound, as previous work has realized adversarial attacks in both the spatial and frequency domains.

Limitations:
1. Only 4/44 references were published in and after 2022, several latest object detection adversarial attack methods are not considered in the paper, e.g., [1],[2],[3]. [1] Huang H, Chen Z, Chen H, et al. T-sea: Transfer-based self-ensemble attack on object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. 2023: 20514-20523. [2] Tang G, Jiang T, Zhou W, et al. Adversarial patch attacks against aerial imagery object detectors[J]. Neurocomputing, 2023, 537: 128-140. [3] Yin M, Li S, Song C, et al. Adc: Adversarial attacks against object detection that evade context consistency checks[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 3278-3287.

2. In Section 3.2, the first limitation claimed by the paper is that previous methods rely on specific modules, but the paper discusses only one previous work (RAP). Does this limitation exist for most of the previous work? Also, the second limitation noted in Section 3.2 is not explicitly stated.
3. The experimental results do not support the article well. In Table 1, the proposed method is not compared with the baseline method. In Table 2, only 2 datasets and 2 models were used to evaluate the performance of NumbOD and the previous method. In most of the metrics, NumbOD shows only a weak performance improvement. As can be seen in Fig. 6 (a), Element B contributes the most to the proposed methodology, and the importance of the other elements needs more analysis to be verified.

Suitability: Moderately suitable, e.g., Unimedia/unimodal in nature but of sufficient interest to the MM community
Rating: Weak Reject - I would need strong arguments to accept this submission.
Confidence: Somewhat Confident
Final Rating: Weak Reject - I would need strong arguments to accept this submission.
Final Rating Justification:

The response from the authors did not fully address my concerns, some of them were also raised by other reviewers, so I would like to keep my decision.

Official Review of Submission1904 by Reviewer 1AcK

Official Review Reviewer 1AcK 19 May 2024, 10:26 (modified: 22 Jul 2024, 21:47)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer 1AcK, Authors

Revisions (/revisions?id=fBpt1VhUwY)

Summary:

The paper presents a novel adversarial attack strategy designed to undermine the performance of object detection systems. By fusing both spatial and frequency domain manipulations, the authors introduce an attack methodology that directly operates on the output features of object detectors, bypassing reliance on internal model structures. This approach includes a dual-track mechanism for selecting high-quality bounding boxes to target, applies directional perturbations to mislead box predictions and classification outcomes, and manipulates high-frequency components to confuse object-focused attention. The effectiveness and stealth of NumbOD are validated through comprehensive experimentation across eight different object detectors and two datasets.

Strengths:

Technical Correctness & Evaluation: The paper demonstrates a rigorous experimental setup, including a diverse set of object detectors and datasets, which underscores the broad applicability and reliability of the attack method. The results showing high attack performance and stealthiness provide strong evidence of the method's technical validity. The manuscript is well-structured and clearly articulates the motivation, methodology, and contributions. The description of the dual-track strategy and the utilization of high-frequency manipulation are particularly clear, facilitating understanding for readers.

Limitations:

1. Environmental Robustness Unclear: The study does not account for the impact of real-world environmental factors such as varying lighting, weather conditions, and camera quality on the attack's effectiveness. These variables can significantly affect image features and thus the performance of adversarial examples in practical applications.
2. Lack of Comparison with State-of-the-Art Attacks: While the paper discusses previous work, it could benefit from more direct quantitative comparisons with other recent adversarial attacks specifically designed for object detection. Such comparisons would further validate NumbOD's novelty and effectiveness in the context of the current research landscape.
3. Human Perception Study Absent: Though stealthiness is claimed through metrics like IW-SSIM, NMSE, and TV, a controlled study involving human participants to quantify the actual perceptibility of adversarial perturbations is missing. This is crucial for understanding the real-world implications where attacks need to evade not just machine but also human detection.

Suitability: Not Suitable, e.g., unimodal in nature and does not contribute to the MM community

Rating: Borderline Reject - I lean towards rejection, but I could be convinced otherwise.
Confidence: Confident
Final Rating: Weak Reject - I would need strong arguments to accept this submission.
Final Rating Justification:
After reading the other reviewers' comments and the author's rebuttal, I decide to lower my rating slightly.

Official Review of Submission1904 by Reviewer zBk6

Official Review Reviewer zBk6 17 May 2024, 17:53 (modified: 22 Jul 2024, 21:47)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer zBk6, Authors
Revisions (/revisions?id=Q2WXwJVHyp)

Summary:
This paper focuses on the attack on object detectors(OD). To this end, they propose a spatial-frequency fusion attack. Specifically, they generate adversarial examples by leveraging the feature output without relying on the internal structures of OD. Experimental results demonstrate the effectiveness of the proposed method.

Strengths:

1. This paper is well organized and well written.
2. Extensive experiments on eight ODs demonstrate their universal attack performance across different OD methods.

Limitations:

1. The authors claim that designing a universal attack for object detection remains challenging and unresolved. However, the definition of a universal attack is confusing. The authors did not explain what a universal attack is. Throughout the paper, it appears that the authors define a universal attack as an attack against different internal structures. As far as I know, a universal attack usually means that the adversarial perturbation can be added to any image-agnostic input[1]. In the context of object detection, universal perturbations have also been studied, such as in [2]. Hence, the current definition is confusing and unclear. The authors should provide a detailed explanation of their definition of a universal attack and how it differs from the one in [1]. Second, DPATCH[3] and T-SEA[4] have demonstrated their effectiveness across different ODs. Therefore, I believe claiming 'designing a universal attack for object detection remains challenging and unresolved' is arbitrary and incorrect, and overstates the contribution.
2. The paper lacks comparisons with state-of-the-art attacks for object detection. T-SEA[4] is the state-of-the-art method for object detection attacks. The authors should compare their method with T-SEA[4] to make the comparison more convincing.
3. The motivation for applying Spatial-Frequency Fusion Attack is missing in the ABSTRACT and not well explained in the paper.
4. Although the authors claim NumbOD is the first spatial-frequency fusion attack against ODs, Sec 3.3 seems to just apply adversarial loss to classification loss and location loss. There is significant overlap with existing OD attacks such as [5]. The spatial-frequency attack is also not new [6]. That makes me think the technical novelty is not enough.
5. Some typos: Line 21 We directly leverage the features output....

Line 615 DumbOD

[1] Universal adversarial perturbations.CVPR. 2017.

[2] Universal adversarial perturbations against object detection. PR 2020

[3] DPATCH: An Adversarial Patch Attack on Object Detectors, arxiv 2018

[4] T-SEA: Transfer-based Self-Ensemble Attack on Object Detection, CVPR 2023

[5] Towards Adversarially Robust Object Detection, ICCV 2019

[6] Exploring Frequency Adversarial Attacks for Face Forgery Detection, CVPR 2022

Suitability: Moderately suitable, e.g., Unimedia/unimodal in nature but of sufficient interest to the MM community
Rating: Weak Reject - I would need strong arguments to accept this submission.
Confidence: Confident
Final Rating: Weak Reject - I would need strong arguments to accept this submission.

Final Rating Justification:
The authors argue that comparing their method with T-SEA is unfair because their proposed approach uses adversarial perturbations, while T-SEA uses adversarial patches. What are the advantages of adversarial perturbations over adversarial patches in object detection? Why the proposed method is better than T-SEA or adversarial patch methods? Although the authors have responded to some extent, my concerns remain unresolved. Therefore, I am inclined to reject it.