2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) | 978-1-6654-0898-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTAI52525.2021.00105

Blur the Eyes of UAV: Effective Attacks on UAV-based Infrastructure Inspection

Ashok Raja
Department of CIS
University of Massachusetts Dartmouth
Email: araja1@umassd.edu

Laurent Njilla
Air Force Research Laboratory
laurent.njilla@us.af.mil

Jiawei Yuan
Department of CIS
University of Massachusetts Dartmouth
Email: jyuan@umassd.edu

***Abstract*—Unmanned aerial vehicles (UAVs) are increasingly leveraged to perform infrastructure inspection tasks, especially with the support of rapidly evolving AI algorithms and hardware in recent years. While the integration of UAV and AI techniques enhances the efficiency and effectiveness of infrastructure inspection, it also raises security concerns due to the potential vulnerabilities existing in the underlying AI models. In this paper, we propose to investigate and discover these vulnerabilities with the case study on bridge inspection. In particular, we designed a two-stage approach that can construct effective adversarial perturbations that make the UAV miss the detection of risk-prone regions during the inspection. Spatial constraints, physical limits, as well as dynamic environmental changes are taken into consideration in our approach to make it practical in the physical world. We evaluate our approach using the COCO-Bridge dataset. Our experimental results demonstrate the effectiveness of our approach in both white-box attack and black-box attack settings.**

I. Introduction

Recent years have witnessed a significant growth in the adoption of UAVs, or drones, in various commercial services. According to a report from Philly By Air [1], there are over $522,000$ commercial drone registered in the United States with the Federal Aviation Administration (FAA) and this number is expected to double by 2024. Thanks to the high mobility and rich sensing capabilities of UAVs, they are now being increasingly leveraged for infrastructure monitoring and inspection tasks (e.g., bridge, pavement, and power utility) [2], especially for these hard-to-access areas. The UAV-based infrastructure inspection is made more effective with the rapid development of AI algorithms and hardware [3]–[6]. For example, Nvidia's Jetson AI modules [7] have been adopted by multiple UAV platforms to support AI operations. In a UAV-based infrastructure inspection task, the UAV leverages the deep neural network (DNN), such as YOLO [8] and Faster-RCNN [9], to detect risk-prone regions and then collect additional data from these detected regions for further analysis.

Although the integration of DNN and UAV can enhance the efficiency and effectiveness of inspection, it also raises security and safety concerns due to the fact that the DNN models used for inspection can be vulnerable to adversarial perturbations. In particular, multiple recent studies have demonstrated that crafted perturbations on visual input can confuse DNN models and make them misunderstand the input data to output wrong results [10]. If this kind of vulnerabilities are successfully exploited by adversaries, risk-prone regions of the infrastructure will be ignored during the inspection, which can cause severe safety consequences, such as bridge failures. In this paper, we choose bridge inspection as the case study to explore potential vulnerabilities in DNN-assisted UAV infrastructure inspection. According to a recent report from the American Society of Civil Engineers (ASCE) [11], 42% of all bridges are at least 50 years old, and 7.5% of the nation's bridges are considered structurally deficient. All these facts indicate the importance of accurate and timely inspection of bridge structures, and hence it is critical to understand and address the potential vulnerability in the inspection process.

The generation of effective adversarial perturbations to identify the potential vulnerabilities in UAV-based bridge inspection faces challenges from the following aspects. First, bridge's risk-prone regions exist in an unconstrained environment that has changing conditions, such as ambient light, weather, distance and angle of UAV's camera. Second, the adversarial perturbation applied on the bridge shall be sensitive to the UAV's camera, but inconspicuous to the human eyes. Third, instead of misclassifying the risk-prone regions to a wrong category (e.g., bearing to cover plate termination), successful adversarial attacks need to make the UAV completely skip risk-prone regions during the inspection. This is because a wrong classification will still trigger the UAV to collect additional data of that region for in-depth analysis.

With these factors in mind, this paper proposes to discover potential vulnerabilities of UAV-based bridge inspection by designing effective adversarial perturbations. In particular, this paper aims to contribute to the understanding of adversarial perturbations against DNN models used for bridge inspection, which can be leveraged to support the design of corresponding defenses in the future to secure the UAV-based infrastructure inspection. Specifically, we propose a two-stage approach towards the generation of practical and effective adversarial perturbations. The stage-one of our approach creates adversarial perturbations with bounded modifications to minimize the probability of detecting risk-prone regions using DNN. In stage-two, an adjustment is applied to the created adversarial perturbations to make them fit into the spatial constraints

DISTRIBUTION A. Approved for public release. Distribution unlimited. Case Number AFRL-2021-2600. Dated 09 Aug 2021

978-1-6654-0898-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICTAI52525.2021.00105

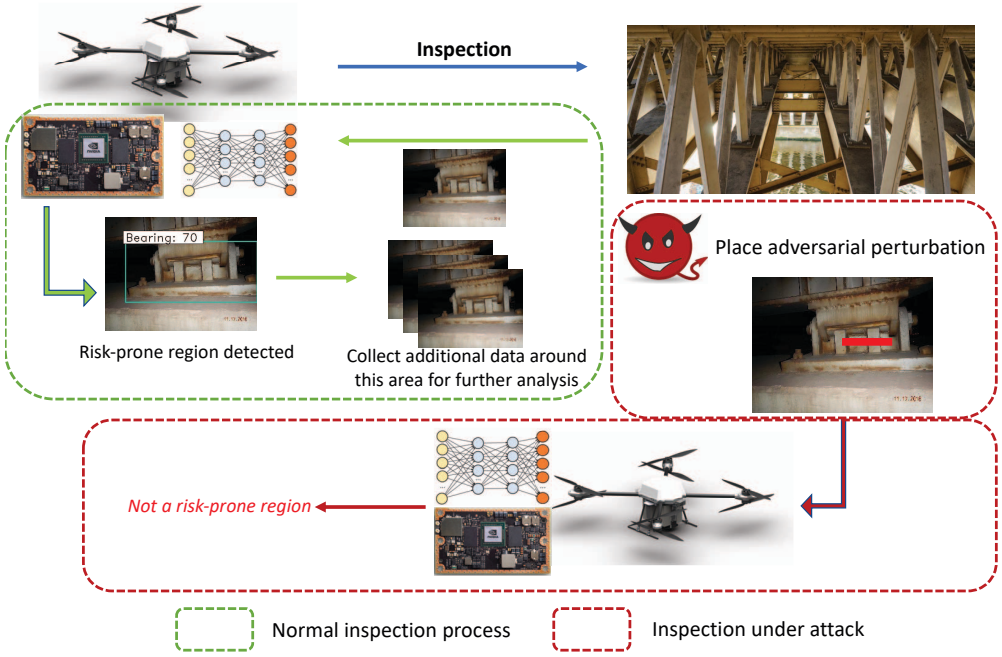

Fig. 1. Problem Overview

and physical limits. We evaluate our proposed approach using Faster R-CNN [9], which supports the deployment on UAVs with Nvidia Jetson for bridge inspection. Our experimental results show that white-box attacks and black-box attacks constructed using the proposed approach can cause the UAV to miss the detection of $48.28\% \sim 65.38\%$ and $41.38\% \sim 62.89\%$ risk-prone regions respectively under different environmental conditions.

The rest of this paper is organized as follows: In Section II, we review and discuss related works. Section III formulates the problem of this paper with the discussion of the system model and the adversary model. Section IV presents the detailed construction of our approach, which is followed by the evaluation in Section V. We conclude this paper in Section VI.

II. Related Work

The vulnerability of neural networks to adversarial examples was first pointed out by Szegedy et al. [12], which demonstrated that a small perturbation on an image can lead to the misclassification of the learning model with high confidence. Since then, a significant amount of research efforts have been put in towards adversarial perturbations on DNNs [10]. The core idea of these works is using carefully crafted modifications to the visual inputs of DNNs to cause the systems they control to misbehave in unexpected and even dangerous manner. Schemes proposed to generate adversarial examples can be classified into two major categories: 1) maximizing the loss function of the target model by adjusting the inputs of the model; and 2) using a surrogate objective function to cause the target model to misclassify the modified input. For both categories of schemes, per-instance generation is typically needed to optimize the performance of the adversarial examples, i.e., separate optimization for each image to generate an adversarially perturbed image.

Besides leveraging adversarial examples to attack the classification of DNN models, recent research also proposed to apply the attacks to DNN-based object detection. In object detection tasks, region of interest (RoI) is typically recognized first and then classified to a specific object. Therefore, the core idea of generating adversarial attacks against object detection is to create adversarial examples against the ROIs. Bao et. al. [13] proposed a sparse adversarial attack on object detectors with bounded l_0 norm pertubation. Xie et al. [14] proposed a scheme using dense adversary generation that considers all objects at the same time and then extends the optimization problem on classification to object detection for the generation of adversarial examples. Also, attacks using project gradient descent and adversarial training on one-stage object detectors are studied in [15], [16]. Recently, Liu et al. proposed a momentum iterative fast gradient sign method [17], which improves the accuracy of adversarial attacks on Faster R-CNN for object detection.

III. Problem Formulation

As depicted in Fig.1, a UAV with a trained DNN model is performing inspections of a bridge. The UAV captures imagery data and analyzes in real-time to determine if the inspected region is risk-prone. For detected risk-prone regions, the UAV will collect additional data for further analysis. In this paper, we consider four categories of risk-prone regions in bridges that may contain defects, including bearing, out

of plane stiffener, gusset plate connection, and cover plate termination.

The purpose of the adversary is to confuse the UAV and make them miss the detection of risk-prone regions of the bridge during the inspection. The adversary launches the attack by placing adversarial perturbations to risk-prone regions of the bridge before it is inspected by the task UAV. In practice, the adversary can place adversarial perturbations with spray painting after the position to attack is determined. Alternatively, the adversary can have UAVs perform an inspection on the target bridge and collect data for analysis. We consider both white-box and black-box attacks from the adversary. In the white-box attack, the adversary has the DNN model and parameters to be used by the task UAV, and then analyzes the model and bridge data to figure out optimized adversarial perturbations used for attacking. In the black-box attack, the adversary does not have access to the DNN model used by the task UAV, and will generate the adversarial perturbations based on his/her own DNN models and the bridge data he/she collected.

IV. Adversarial Attacks towards Bridge Inspection

During the detection of risk-prone regions of a bridge, the DNN model first recognizes RoIs from the imagery data collected by the task UAV and place bounding boxes on them. Then, these RoIs will be classified into the corresponding category of risk-prone regions. Therefore, the goal of our design is to examine whether it is possible to create effective perturbations on the bridge that cause the DNN model to skip RoIs during the inspection.

We denote a single image of a bridge area captured by the UAV as x, which is used as the input of the DNN model. The perturbation to be added to x is denoted as β and hence $x' = x + \beta$ is the corresponding adversarial input. We also use C to denote the set of all categories of risk-prone regions, and B to denote the set of all potential RoIs. Given an RoI $b \in B$, the target of the added adversarial perturbation is to make the probability of classifying b as any category of risk-prone regions ($c \in C$) become smaller than the predefined threshold ($\mathcal{T}$). Therefore, our problem can be transformed as a constrained optimization problem to search an effective adversarial perturbation β for a risk-prone region, which achieves

$$\beta < D \quad s.t. \quad \mathcal{P}(b \to c) < \mathcal{T}, \forall\ c \in C \tag{1}$$

where D is the maximum size of the perturbation that can be applied to a risk-prone region of the bridge. To improve the effectiveness and robustness of the perturbations, we reformulate it as

$$\beta < D \quad s.t. \quad min\ \mathcal{P}(\gamma, x', C),\ c \in C \tag{2}$$

where γ is the set of model parameters and $\mathcal{P}()$ is the probability that x' is classified as a risk-prone region category $c, c \in C$. It is clear that adversarial perturbations can achieve the best performance by minimizing the output of $\mathcal{P}(\gamma, x', C)$.

Algorithm 1: Construction of Adversarial Perturbation

Input: x, target area θ to add perturbation
Output: Adversarial perturbation $(\beta_{optimal}, k_{optimal})$

Stage-1:
k: the position of β (denoted as k), $\mathcal{K}$: a queue;
set minProb = 100%;
for $\beta = 0.1D;\ \beta \leq D;\ \beta = \beta + 0.1D$ **do**
  Set k to the left-top corner of θ;
  while k *does not reach the right-bottom corner of* θ **do**
    p = $average(\mathcal{P}(\gamma, x + \beta, c)), c \in C$;
    push (β, k, p) to $\mathcal{K}$;
    if $p <$ *minProb* **then**
      minProb = p;
      $\beta_{optimal} = \beta, k_{optimal} = k$;
    end
    Move k rightward by w;
    if k *reaches the right bound of* θ **then**
      Move k downward by h;
      Move k leftward to the left bound of θ;
    end
  end
end

Stage-2:
if *(*$\beta_{optimal}, k_{optimal}$*) is a valid pertubation* **then**
  break;
end
sort $\mathcal{K}$ in an ascending order in terms of p;
for *each* k *in* $\mathcal{K}$ **do**
  if *(*β, k*) is a valid pertubation* **then**
    $\beta_{optimal} = \beta, k_{optimal} = k$;
    break;
  end
end

A. Construction of Adversarial Perturbations

We now present the two-stage approach for the construction of attacks using adversarial perturbations by solving Eq.2 with the consideration of spatial constraints and physical limits. The detailed design of our approach is presented in Algorithm.1. In our design, the adversary can identify the initial target area θ to add perturbations by flying his/her UAVs around the bridge and collect imagery data of its risk-prone regions. After target area θ are determined, our approach will completely scan it with different sizes of adversary perturbations from corner to corner in the *Stage-1*, in which the combination of $(\beta_{optimal}, k_{optimal})$ that minimize the average probability of $\mathcal{P}(\gamma, x + \beta, c), c \in C$ can be obtained. The *Stage-2* of our approach is designed to examine whether $(\beta_{optimal}, k_{optimal})$

obtained in *Stage-1* is a valid perturbation. If not, the perturbations sorted in queue $\mathcal{K}$ according to their probabilities will be checked until a valid perturbation is obtained.

To determine whether a perturbation is valid or not, we consider the following factors in terms of spatial constraints and physical limits introduced by the physical world. Unlike attacking digital images that can apply perturbations to any part of the image, the spatial constraints in our situation limit the application of perturbations on the physical object (i.e., bridge) only but not on the background areas (e.g., sky and river). In addition, attacks on digital images are able to modify any single pixel to the desired value for the generation of perturbation. However, it is impractical to generate physical perturbations (e.g., paint, sticker) to match the size and value of such optimal pixels for attacking. Therefore, the checking in *Stage-2* of our approach will exclude solutions that are restricted by these physical factors. For example, the values of the perturbations shall be in the set of printable colors.

B. Impact of Environmental Conditions

Besides the spatial constraints and physical limits, dynamic environmental conditions can also affect the effectiveness of the adversarial perturbations. For example, UAVs can perform the inspection task at different distances and angles with dynamic ambient light and weather conditions. Therefore, instead of obtaining a local optimal adversarial perturbation in a specific environmental condition, our approach shifts the emphasis to figuring out a global optimal adversarial perturbation that achieves the overall best performance in different conditions for the same risk-prone region. Given $p_1, p_2, \cdots, p_n$ as the probability of $average(\mathcal{P}(\gamma, x+\beta, c)), c \in C$ under n different conditions, we define the global optimal adversarial perturbation $\beta_{optimal}$ that achieves the minimized weighted average probability as

$$\beta_{optimal} \rightarrow min \frac{u_1 p_1 + u_2 p_2 + \cdots + u_n p_n}{n} \tag{3}$$

$u_1, u_2, \cdots, u_n$ are the weights for each condition, which are assigned based on the possibility to appear of each condition in practice. For example, a regular daylight condition is more common for inspection tasks compared with a low ambient light condition. To obtain data under different conditions, we first rotate the image during the analysis to simulate different angles. Then, data augmentation is applied to obtain images with different ambient light conditions (bright to low-light) and distances (zoom in/out) as examples shown in Fig.2.

V. Evaluation

A. Experimental Settings

To evaluate the performance of our proposed approach, we implemented it on Faster-RCNN based bridge inspection. ResNet and VGG are adopted as the backbone architectures respectively. We use COCO-Bridge dataset [6] for our evaluation, which contains 719 bridge images with 2337 annotated risk-prone regions for training and another 55 images for testing. Four categories of risk-prone regions of bridges are considered, including bearing, out of plane stiffener, gusset plate connection, and cover plate termination. The training of DNN models and generation of adversarial perturbations are performed on a desktop computer with i7 8-core CPU, 32GB memory, and one RTX 3070 GPU.

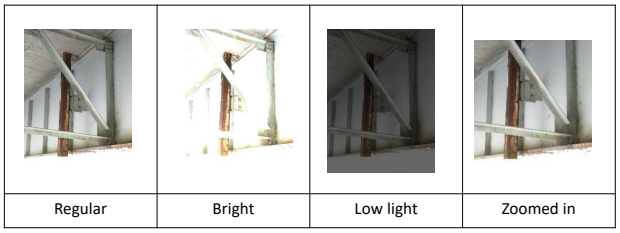

Fig. 2. Examples of the Same Bridge Area under Different Conditions

In our evaluation, the IoU threshold is set as 0.7. To measure the effectiveness of our attacks, we compare the numbers of RoIs detected for risk-prone regions with and without our attacks. Therefore, we define the attack success rate as

$$\frac{\#\ \ RoIs\ without\ attack - \#\ \ RoIs\ with\ attack}{\#\ \ RoIs\ without\ attack} \tag{4}$$

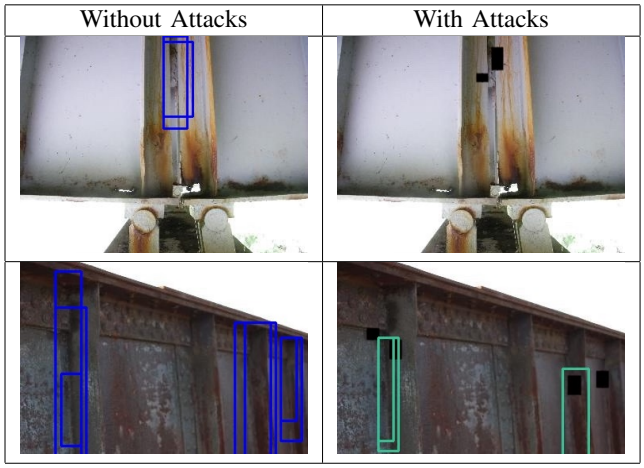

Fig. 3. Examples of Inspection with and without Attacks

B. Experimental Results

White-box Attack: We first evaluate the performance of our approach in terms of white-box attacks, in which the Faster-RCNN (ResNet) is used. Fig.3 presents examples of the inspection results with and without attacks. As summarized in Table I, our approach achieves attack success rates from 48.28% to 65.38% under different environmental conditions, i.e., the attacks will confuse the UAV to miss more than half of the risk-prone regions during the inspection of most cases. If defects exist in these missed risk-prone regions are not addressed in time, they can cause severe safety consequences. It is noteworthy that the DNN model detects a relatively lower number of RoIs in the low-light environment. This is because

the reduced light environment significantly affects the quality of imagery data collected by the UAV's camera.

TABLE I
WHITE-BOX ATTACK RESULTS

Scenario	# of RoIs without Attack	# of RoIs with Attack	Success Rate
Regular	98	44	55.1%
Bright	78	27	65.38%
Low-light	29	15	48.28%
Changed Distance	97	37	61.86%

Black-box Attack: With regard to the black-box attacks, we apply the adversarial perturbations generated using the adversary's Faster-RCNN model (VGG) to the Faster-RCNN model (ResNet) used by the UAV. As presented in Table II, the effectiveness of our approach for black-box attacks is comparable with that in white-box attacks.

TABLE II
BLACK-BOX ATTACK RESULTS

Scenario	# of RoIs without Attack	# of RoIs with Attack	Success Rate
Regular	98	43	56.12%
Bright	78	32	58.97%
Low-light	29	17	41.38%
Changed Distance	97	36	62.89%

C. Discussion

Attack Success Rate: In practical attacks towards UAV-based infrastructure inspection, a high attack success rate (e.g., $> 75\%$) can lead the inspector to doubt about the data collected by the UAV. When the attacks make the UAV detect only a small portion of risk-prone regions, the inspector has a high chance to realize the potential problems in the detection tools, because bridges typically contain multiple risk-prone regions. Therefore, practical adversarial perturbation may try to reduce the number of RoIs by $30\% \sim 40\%$ for a UAV's inspection. For infrastructures like bridge, the missed detection of a small amount of risk-prone regions with defects can still lead to severe consequences if they are not addressed in time.

Size of Perturbation: During the practical UAV-based inspection, the size of the same perturbation can vary in images captured by UAVs at different distances, i.e., a smaller distance indicates a larger perturbation. In practical attacks, the size of perturbation can be determined by images collected at a relatively large distance using UAVs. This is because the perturbation will be enlarged in the image when the UAV flying towards the bridge, which guarantees that the perturbation still covers the attacking area.

VI. CONCLUSION

In this paper, we investigate and discover the potential security vulnerabilities in UAV-based infrastructure inspection with the focus on bridge inspection. A two-stage approach is proposed in this paper for the generation of effective and physically realizable adversarial perturbations. The evaluation results on the real-world dataset demonstrate the effectiveness of adversarial perturbations generated using our approach in both white-box and black-box attacks. Evaluation with the consideration of different environmental conditions is also conducted to validate the robustness of our approach. This research contributes to the understanding of adversarial attacks against DNN models used for infrastructure inspection. Based on the discovery of this paper, corresponding defenses shall be designed in the future to protect the UAV-based infrastructure inspection from adversarial attacks.

ACKNOWLEDGMENT

This work is supported by Air Force VFRP Award, US NSF Award (DGE-1956193), and the UMass Dartmouth Cybersecurity Center Fellowship.

REFERENCES

[1] Federal Aviation Administration. UAS by the Numbers. https://www.phillybyair.com/blog/drone-stats/, 2021.
[2] Hazim Shakhatreh, Ahmad H. Sawalmeh, Ala Al-Fuqaha, Zuochao Dou, Eyad Almaita, Issa Khalil, Noor Shamsiah Othman, Abdallah Khreishah, and Mohsen Guizani. Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges. *IEEE Access*, 7:48572–48634, 2019.
[3] Weidong Wu, Murad A. Qurishee, Joseph Owino, Ignatius Fomunung, Mbakisya Onyango, and Babatunde Atolagbe. Coupling deep learning and uav for infrastructure condition assessment automation. In *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–7, 2018.
[4] Van Nhan Nguyen, Robert Jenssen, and Davide Roverso. Intelligent monitoring and inspection of power line components powered by uavs and deep learning. *IEEE Power and Energy Technology Systems Journal*, 6(1):11–21, 2019.
[5] Neshat Bolourian and Amin Hammad. Path planning of lidar-equipped uav for bridge inspection considering potential locations of defects. In Ivan Mutis and Timo Hartmann, editors, *Advances in Informatics and Computing in Civil and Construction Engineering*, pages 545–552, Cham, 2019. Springer International Publishing.
[6] Eric Bianchi. Coco-bridge: Common objects in context dataset and benchmark for structural detail detection of bridges. 2019.
[7] Nvidia Jetson Solutions for Drones & UAVs. https://www.nvidia.com/pt-br/autonomous-machines/uavs-drones-technology/.
[8] YOLOv5. https://ultralytics.com/yolov5.
[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.
[10] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.
[11] The American Society of Civil Engineers. 2021 Infrastructure Report Card. https://infrastructurereportcard.org/wp-content/uploads/2020/12/Bridges-2021.pdf.
[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.
[13] Jiayu Bao. Sparse adversarial attack to object detection, 2020.
[14] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection, 2017.
[15] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 421–430, 2019.
[16] Yutong Wang, Kunfeng Wang, Zhanxing Zhu, and Fei-Yue Wang. Adversarial attacks on faster r-cnn object detector. *Neurocomputing*, 382:87–95, 2020.
[17] Zhenghao Liu, Wenyu Peng, Jun Zhou, Zifeng Wu, Jintao Zhang, and Yunchun Zhang. *MI-FGSM on Faster R-CNN Object Detector*, page 27–32. ACM, New York, NY, USA, 2020.