frontiers | Frontiers in Computer Science

TYPE Review
PUBLISHED 21 November 2024
DOI 10.3389/fcomp.2024.1349206

OPEN ACCESS

EDITED BY
Xue-Cheng Tai,
Norwegian Research Institute (NORCE), Norway

REVIEWED BY
Akshay Agarwal,
Indian Institute of Science Education and Research, Bhopal, India
Sadadi Ojoatre,
BeZero carbon, United Kingdom

*CORRESPONDENCE
Shiyong Chu
csy319ldl@163.com

RECEIVED 06 December 2023
ACCEPTED 01 November 2024
PUBLISHED 21 November 2024

CITATION
Chen Y and Chu S (2024) On the adversarial robustness of aerial detection.
Front. Comput. Sci. 6:1349206.
doi: 10.3389/fcomp.2024.1349206

On the adversarial robustness of aerial detection

Yuwei Chen and Shiyong Chu*

Aviation Industry Development Research Center of China, Beijing, China

Deep learning-based aerial detection is an essential component in modern aircraft, providing fundamental functions such as navigation and situational awareness. Though promising, aerial detection has been shown to be vulnerable to adversarial attacks, posing significant safety concerns. The sparsity of a comprehensive analysis on the robustness of aerial detection exacerbates these vulnerabilities, increasing the risks associated with the practical application of these systems. To bridge this gap, this paper comprehensively studies the potential threats caused by adversarial attacks on aerial detection and analyzes their impact on current defenses. Based on the most widely adopted sensing strategies in aerial detection, we categorize both digital and physical adversarial attacks across optical sensing, infrared sensing, and Synthetic Aperture Radar (SAR) imaging sensing. Owing to the different imaging principles, attacks in each sensing dimension show different attack vectors and reveal varying attack potentials. Additionally, according to the operational life cycles, we analyze adversarial defenses across three operational phases: pre-mission, in-mission, and post-mission. Our findings reveal critical insights into the weaknesses of current systems and offer recommendations for future research directions. This study underscores the importance of addressing the identified challenges in adversarial attack and defense, particularly in real-world scenarios. By focusing future research on enhancing the physical robustness of detection systems, developing comprehensive defense evaluation frameworks, and leveraging high-quality platforms, we can significantly improve the robustness and reliability of aerial detection systems against adversarial threats.

KEYWORDS

adversarial robustness, aerial detection, adversarial attack, adversarial defense, physical adversarial attack

1 Introduction

Aerial detection, facilitated by airborne sensors mounted on aircraft platforms, is essential for capturing high-resolution images of the earth's surface, widely applied in defense (Fingas and Brown, 2001), agriculture (Everitt et al., 1991), mining (Maathuis and Genderen, 2004), and mapping (Connor et al., 2016). This aerial detection system, based on deep learning, is indispensable for intelligence gathering, surveillance, and reconnaissance operations. Unlike ground object detection, where the objects are typically on a stable surface, aerial detection deals with dynamic and three-dimensional spaces, introducing complexities related to altitude, speed, and varying perspectives (Wilkening, 2004). The additional dimension in aerial detection provides a more comprehensive understanding of the spatial relationships between objects in the airspace increasing the task success probability of object detection.

Compared with ground detection, aerial detection usually covers a larger geographical area and can detect a large range of ground targets in a short time to increase efficiency (Dhillon and Verma, 2020; Galvez et al., 2018). However, as edge devices, aircraft platforms

executing aerial detection cannot carry sufficient computing facilities and resources. So aerial detection should be executed under the circumstance of limited resources in practice. This brings the problems that the model used in this scenario is more lightweight than the traditional one, and it is more vulnerable to adversarial attacks in practical world.

In fact, the emergence of adversarial examples has precipitated significant concerns regarding the adversarial robustness of AI models. These visually imperceptible perturbations can induce misclassifications within deep learning models and systems, leading to a cascade of consequential issues (Kurakin et al., 2016a,b). For example, attackers post adversarial patches above or close to the target, preventing the optical object detector onboard the drone from making stable and accurate identities (Du et al., 2022). In addition, attackers can also change the infrared characteristics of targets by installing special light bulbs near the targets to attack target detection under the infrared system (Zhu et al., 2021). Given the complexity, openness, dynamics, and adversarial nature unique to scenarios in aerial detection, AI-based sensors confront a broader spectrum of attacks. These attacks compromise not only target detection accuracy but also pose security risks by inducing subsequent judgment or decision-making errors. Considering the safety-critical nature of aerial detection, it is of paramount importance to rigorously investigate and comprehensively study the robustness of adversarial examples.

Despite a substantial body of research on adversarial attacks and defenses in the context of aerial detection, a comprehensive investigation is conspicuously absent. This sparsity of research presents a severe risk to the safety of aerial detection systems, as it increases their vulnerabilities to attacks. In this article, we conduct an exhaustive review of existing research endeavors related to adversarial attacks and defenses in the context of aerial detection, aiming to provide an overall framework for better understanding the adversarial challenges for aerial detection. Based on the most widely adopted sensing strategies in aerial detection (Wilkening, 2004), we categorize both digital and physical adversarial attacks in the context of aerial detection from perspectives including optical sensing, infrared sensing, and SAR imaging sensing. Optical sensors observe within the visible electromagnetic wave range (Crawford, 1998), infrared sensors detect infrared radiation for target identification (Norton, 1991), and SAR imaging sensors utilize radar signals for target identification based on reflectivity, shape, and size. Owing to the different imaging principles, attacks in each sensing dimension show different attack vectors and reveal different attack potentials. Besides adversarial attacks, according to the operational life cycles, we also comprehensively analyze adversarial defenses for aerial detection from three operational phases including pre-mission, in-mission, and post-mission. Finally, we pinpoint several directions for future studies such as real-world attacks, defense, high-quality platforms, among others. It is imperative to clarify that, within the context of our study, the domain of aerial detection encompasses all surveillance activities conducted from an elevated vantage point at a specified altitude above the target of detection. This scope includes diverse practices, including ground imaging through remote sensing and satellite-based surveillance of terrestrial features. In summary, our primary contributions can be summarized as follows:

- We present a comprehensive investigation of both digital and physical adversarial attacks in the context of aerial detection based on the most widely used sensing principles, delivering an in-depth analysis of the evolution and progress within this research domain.
- Besides attacks, we also systematically investigate the existing adversarial defenses from the perspectives of the main operational phases of aerial detection missions.
- We engage in an extensive discussion regarding the challenges posed by real-world intelligent aerial detection and their implications on subsequent operational stages.

2 Preliminaries

2.1 Adversarial example

In 2013, researchers, including Szegedy et al. (2013), made a groundbreaking discovery within the realm of intelligent algorithms in computer vision. They identified a minute form of imperceptible noise, which, despite its inconspicuous nature to the human eye, had the potential to mislead deep neural network models. This newly defined phenomenon was termed an AE (Goodfellow et al., 2014) and is characterized as follows:

$$f_\Theta(x_{adv}) \neq y, s.t. \parallel x - x_{adv} \parallel \leq \epsilon, \tag{1}$$

In this context, where x represents the original data, x_{adv} denotes an adversarial example augmented with adversarial noise, y stands for the category label assigned to the original data x, $\parallel \cdot \parallel$ symbolizes the measure of the distance between x and x_{adv}, and ϵ signifies any positive value. This expression conveys the concept that even when the disparity between x and x_{adv} is infinitesimal, the neural network yields an erroneous classification outcome.

As research progresses, scholars have unveiled the extensive impact of AEs across a spectrum of domains, including natural language processing (NLP) (Zhang W. E. et al., 2020; Morris et al., 2020; Qiu et al., 2022; Chang et al., 2021), speech recognition (Qin et al., 2019; Cisse et al., 2017; Samizade et al., 2020; Schönherr et al., 2018), as well as a diverse array of AI paradigms and algorithms such as deep learning, reinforcement learning, and statistical machine learning. The perturbing effects of AEs are profound and multifaceted. Additionally, owing to its inherent versatility, AEs can be harnessed to orchestrate black-box attacks (Papernot et al., 2017; Liu et al., 2016; Guo et al., 2019; Ilyas et al., 2017; Jia et al., 2019), circumventing the need for specific knowledge concerning the target model. Notably, these attacks transcend the boundaries between the digital (Jan et al., 2019; Kong et al., 2020; Liu et al., 2020b) and physical (Kurakin et al., 2016a; Kong et al., 2020; Song et al., 2018; Xu et al., 2020; Athalye et al., 2018; Kurakin et al., 2018) realms, demonstrating their efficacy and reach across both domains.

In the present research landscape, scholars have undertaken exhaustive investigations into AEs, encompassing inquiries into the underlying mechanisms of AEs' successful incursions into intelligent algorithms (Liu et al., 2023a), methods for enhancing the likelihood of AE attack success (Liu et al., 2021; Yu et al., 2021),

and strategies for fortifying defenses against AE attacks (Guo et al., 2023; Liu et al., 2023b; Zhang et al., 2021). These research efforts span an array of application domains for intelligent algorithms, ranging from autonomous driving and navigation (Kong et al., 2020; Boloor et al., 2019; Cao et al., 2019b,a; Tu et al., 2020; Zhou et al., 2020; Liu et al., 2020a), facial recognition (Zhu et al., 2019; Zhang B. et al., 2020; Sharif et al., 2019; Massoli et al., 2021; Vakhshiteh et al., 2021), to target detection (Song et al., 2018; Xie et al., 2017; Liu et al., 2019; Yang et al., 2018; Smith and Gal, 2018), among others.

2.2 Aerial detection

In Aerial detection, the detection sensor on the aircraft carries out object detection on the ground target. The formula for the process is as follows:

$$\min \mathbb{E}_{(\mathbf{I},\{\mathbf{y},\mathbf{b}\})\sim\mathbb{D}}\mathcal{L}(f(\mathbf{I}),\{\mathbf{y},\mathbf{b}\}), \tag{2}$$

where $\mathcal{L}(\cdot)$ is the loss function that measures the difference between the output of the detector f and the ground truth, f is the detection processing and analysis system, $\mathbf{I}$ is the picture of the input detector, $\mathbf{y}$ denotes the true category of the target, and $\mathbf{b}$ denotes the real bounding box. When attacking the sensor in an aerial detection scenario, given an object detector f and an input image $\mathbf{I}$ with the ground truth label $\{\mathbf{y},\mathbf{b}\}$, an adversarial example $\mathbf{I}_{adv}$ satisfies the following:

$$f(\mathbf{I}_{adv}) \neq \{\mathbf{y},\mathbf{b}\} \quad s.t. \quad \|\mathbf{I}-\mathbf{I}_{adv}\| \leq \epsilon, \tag{3}$$

where $\|\cdot\|$ is a distance metric and commonly measured via ℓ_p-norm ($p \in \{1,2,\infty\}$).

In physical aerial detection scenarios, the items $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ on the ground are scanned via different sensors to produce an image, where $\mathcal{R}$ denotes the process of generating an image depicted as $\mathbf{I} = \mathcal{R}(\mathbf{X})$.

In conclusion, to discuss physical attacks, the ground target is scanned by the sensor into image $\mathbf{I}_{adv}$, which could deceive the object detector $f(\cdot)$ with attack mode $\mathcal{A}$, minimizing $\mathcal{M}$ that measures the performance of the detector:

$$\min \mathcal{M}\left[f(\mathcal{R}(\mathbf{x}_1, ..., \mathbf{x}_m, \mathbf{x}^{\mathbf{A}}_{adv}), \{\mathbf{y},\mathbf{b}\})\right]. \tag{4}$$

Further, sensor $\mathcal{R} \in \{\mathcal{R}_o, \mathcal{R}_i, \mathcal{R}_s\}$ which includes optical sensors $\mathcal{R}_o$ which rely on the wavelength in the visible range of electromagnetic waves for observation and imaging (Crawford, 1998), infrared sensors $\mathcal{R}_i$ which detect the infrared radiation emitted or reflected by the target to carry out identification (Norton, 1991), and SAR imaging sensor $\mathcal{R}_s$ which use the antenna to transmit and receive radar signals to the target, and identifies the target type according to the reflectivity, shape and size of the target.

For different sensors, the attacker designs the attack mode $\mathbf{A} \in \{\mathbf{P}, \mathbf{T}, \mathbf{C}, \mathbf{N}, \mathbf{E}\}$ according to the sensor detection mechanism: changing optical imaging results $\mathbf{P}$ to attack optical sensors, changing thermal signature $\mathbf{T}$ and infrared radiation $\mathbf{C}$ to attack infrared sensors, changing scattering properties $\mathbf{N}$ and texture properties $\mathbf{E}$ to attack SAR imaging sensors.

For aerial exploration, since aerial imagery primarily acquired from elevated platforms such as drones or satellites, provides an unconventional top-down or oblique perspective, which results in a counterattack on aerial imagery, Researchers must consider that the adversarial example is similar in size or pixel to the detection target, and consider the realizability of the adversarial example in the physical environment. At the same time, due to the potential air obstacles encountered by sensor imaging in aerial detection, such as wires, clouds, rain, and other environmental elements, researchers can use or eliminate this part when fighting attacks and defenses to ensure attack efficiency and defense success probability.

3 Adversarial attack on aerial detection

This study categorizes assaults on aerial detection sensors into three distinct types: optical sensors, infrared sensors, and SAR imaging sensors. These categorizations form the foundation for a comprehensive synthesis of attack methodologies, implementation modalities, underlying rationales, and specific mission objectives across these sensor categories, as shown in Figure 1. The resulting as shown in Table 1 encapsulates these findings for reference and analysis.

Our emphasis is on distinguishing between physical attacks and digital attacks. We assert that, within the domain of aerial detection, certain digital attacks, such as mapping alterations on remote sensing images, primarily aim to enhance existing algorithms. In contrast, physical attacks, such as strategically placing objects emitting counter signals near the target being detected, exert a more substantial disruptive influence on the functionality of aerial detection sensors.

However, physical attacks are susceptible to the influence of various natural conditions, including light, cloud cover, and other uncontrollable factors. A comprehensive description of these variables in experimental settings can be challenging. In contrast, digital attacks offer the advantage of being executed under ideal conditions, providing researchers with a more convenient platform to investigate strategies for countering attacks and to probe the vulnerabilities in intelligent algorithms.

3.1 Adversarial attack on optical sensors

Optical sensors systematically capture visible light emanating from objects within their natural surroundings, discerning object categories based on distinct visible light performance parameters. Subsequently, assailants often exploit optical sensors by orchestrating nuanced yet meticulously crafted alterations to optical characteristics or introducing image perturbations, thereby undermining the efficacy of target detection. These manipulations encompass subtle adjustments to target attributes such as color, texture, or brightness within the visible spectrum of imaging outcomes. These tactics aim to subtly distort the information received by the detector.

To investigate adversarial attacks on optical sensors, researchers have conducted extensive studies in both digital and physical environments. These investigations typically consider the

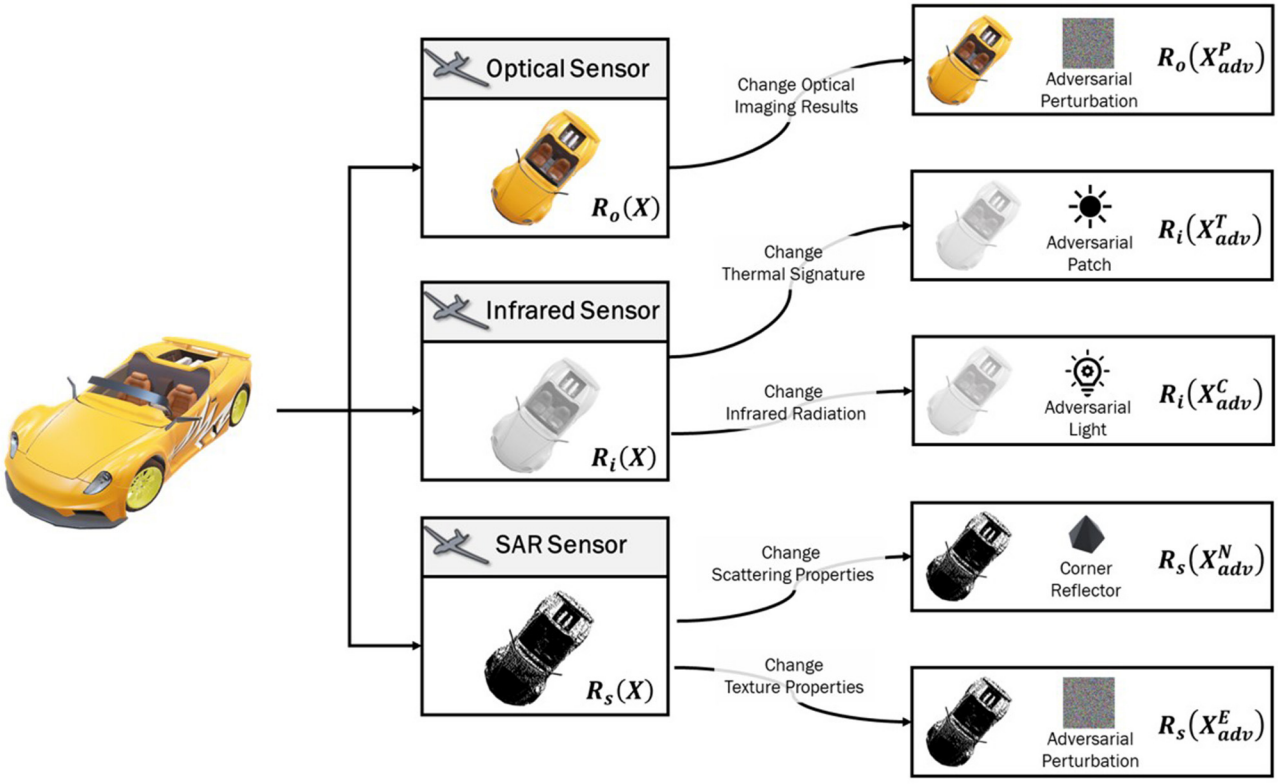

FIGURE 1
Adversarial attacks on aerial detection systems can be categorized based on the type of sensor targeted: optical sensors, infrared sensors, and SAR sensors. For *Optical sensors*, adversarial attacks typically involve altering optical imaging results. *Infrared sensors* are usually targeted by modifying the thermal signature or infrared radiation. In the case of *SAR sensors*, attacks are generally conducted by altering scattering properties or texture properties.

operational principles of optical sensor imaging in aerial detection, integrating the specific characteristics of aerial detection activities.

In the realm of fundamental model vulnerability research, Chen et al. (2020) explored the mechanisms by which models can be susceptible to attacks, emphasizing attack selectivity and model vulnerability. Attack selectivity underscores the variability in the impact of AE across diverse models, allowing attackers to achieve superior results by selecting specific models. Model susceptibility, on the other hand, pertains to a model's resilience against AE and its capability to correctly identify such examples. Higher susceptibility signifies a reduced ability of the model to recognize AEs.

3.1.1 Digital attacks on optical sensors

Numerous studies have investigated digital attacks on optical sensors, commonly involving the incorporation of adversarial patches or other elements capable of transmitting attack information into resulting remote sensing images. The majority of these studies emphasize enhancing attack strategies targeting the detector. While this approach offers the advantage of simulating attacks on optical sensors under ideal conditions, enabling exploration of the mechanisms behind successful attacks and algorithm vulnerabilities, it is essential to note its limitation in disregarding the intricate weather conditions typically encountered in aerial detection scenarios.

To facilitate multi-scale object detection in Unmanned Aerial Vehicle (UAV) remote sensing images, Zhang Y. et al. (2022) introduced a method for adapting the patch size to the scaling factor of the height label. Specifically, in digital attacks, the patch is resized in accordance with the height label of the image to accommodate targets of varying scales. Consequently, this approach ensures that the patch can yield effective attack outcomes on multi-scale targets across real-world scenarios, regardless of their size.

Den Hollander et al. (2020) employed adversarial patches on images of military targets to obscure them in automatic target detection. Their experiments demonstrated that adversarial patches are markedly more adept at concealing military targets compared to noisy patches, resulting in higher attack success rates. The authors also explored the impact of adversarial patch size on the attack's success rate, finding that larger patches are more effective in disguising the target. However, this comes at the expense of increased detectability of the attack. Consequently, selecting an appropriately sized adversarial patch can enhance the overall success rate of the attack.

To counter adversarial attacks on salient object detection models in remote sensing images, Sun et al. (2023) introduced an attack strategy termed "Adversarial Cloud". This method generates a cloud mask to simulate cloud cover in remote sensing images, transforming the original image into an adversarial cloud image. The objective is to deceive visual-based salient object detection

TABLE 1 Overview of recent adversarial attack research against airborne detection sensors.

References	Sensor type	Attack method	Attack mode	Adversarial's knowledge	Target task
Zhang Y. et al. (2022)	Optical	Adversarial patch	Digital	White box & black box	Detection
Den Hollander et al. (2020)	Optical	Adversarial patch	Digital	White box & black box	Detection
Sun et al. (2023)	Optical	Adversarial perturbation	Digital	White box	Detection
Lian et al. (2022)	Optical	Adversarial patch	Digital	White box	Detection
Wang et al. (2023)	Optical	Adversarial perturbation	Digital	White box & black box	Detection
Wang et al. (2024)	Optical	Adversarial perturbation	Digital	White box	Classification
Du et al. (2022)	Optical	Adversarial patch	Physical	Not mentioned	Detection
Shrestha et al. (2023)	Optical	Adversarial patch	Physical	White box	Detection
Lian et al. (2023)	Optical	Adversarial patch	Physical	White box	Detection
Tang et al. (2023)	Optical	Adversarial light	Physical	Black box	Detection
Xu and Ghamisi (2022)	Optical	Adversarial perturbation	Physical	Black box	Classification & detection
Wang et al. (2021)	Optical	Adversarial patch	Physical	Black box	Classification
Zhang et al. (2024)	Optical	Adversarial patch	Physical	White box	Detection
Zhou et al. (2024)	Optical	Adversarial Patch	Physical	White box & black box	Detection
Wei X. et al. (2023)	Infrared	Adversarial patch	Physical	White box & black box	Detection
Qi et al. (2022)	Infrared	Adversarial patch	Digital	White box	Detection
Wei H. et al. (2023)	Infrared	Adversarial patch	Physical	Black box	Detection
Zhu et al. (2021)	Infrared	Adversarial light	Physical	Black box	Detection
Zhang F. et al. (2022)	SAR	Adversarial perturbation	Digital	White box & black box	Classification & detection
Li et al. (2020)	SAR	Adversarial perturbation	Digital	White box & black box	Detection
Du et al. (2023)	SAR	Adversarial perturbation	Digital	Black box	Detection
Peng et al. (2021)	SAR	Adversarial perturbation	Digital	White box & black box	Detection
Peng et al. (2022)	SAR	Adversarial perturbation	Digital	Black box	Classification
Zhang L. A. et al. (2022)	SAR	Corner reflector	Physical	Black box	Detection

models designed for remote sensing images. Experimental results demonstrate the effectiveness of the Adversarial Cloud attack method, significantly compromising the performance of the salient target detection model and resulting in a substantial reduction in the F-measure value, decreasing from 0.8253 to 0.2572.

The work by Lian et al. (2022) outlined a physical attack framework based on adaptive patches. This framework executes covert attacks on designated targets by generating adaptive patches capable of concealing specific targets amidst varying physical dynamics and scales. Throughout the attack, the patch's placement, whether within or surrounding the target, uniformly impacts all targets belonging to the same category, while maintaining robustness within the physical domain.

Simultaneously, the study presented by Wang et al. (2023) provided a comprehensive assessment of success rates associated with diverse adversarial attacks targeting various semantic segmentation networks designed for aerial images. The evaluation metric employed is Pixel Accuracy (PA), reflecting the models' ability to correctly classify pixels. For example, in the FGSM attack-generated AE test set, the TCHNet model achieves a PA of 24.28%, while PA values for C&W, PGD, and UAP attacks are notably lower at 18.57, 17.82, and 15.16%, respectively. In stark contrast, the proposed GFANet exhibits exceptional resilience, consistently achieving PA values exceeding 80% across a spectrum of adversarial attack methods. This performance surpasses that of its counterparts in the realm of aerial image semantic segmentation networks, emphasizing its heightened robustness and efficacy in adversarial scenarios.

Wang et al. (2024) introduced an innovative adversarial attack framework called Background Attack via Dual-Adversarial-Induced Error Identification (BADEI), specifically targeting optical sensors used in aerial object detection. This framework leverages adversarial background manipulation to deceive detection systems, causing them to either misclassify targets as background or falsely identify background elements as targets. The methodology includes an Unoccluded Training Strategy (UTS), which strategically places adversarial backgrounds directly beneath targets, and a Dual Deceptive Loss Function (D2LF)

that facilitates both target concealment and misclassification. While the validation primarily occurred in a digital environment, the BADEI framework demonstrated exceptional attack efficacy across various optical detection models. The results underscore the vulnerability of optical sensors to such sophisticated attacks, significantly compromising their accuracy in detecting and identifying aerial targets.

3.1.2 Physical attacks on optical sensors

In the domain of physical attacks, researchers commonly manifest adversarial information by affixing stickers and patches to or in proximity of the target under surveillance. Aerial detectors are then deployed to assess detection and recognition capabilities in the area. This method offers the advantage of conducting experiments in real-world aerial detection scenarios. However, a drawback lies in the susceptibility of optical sensors to external factors like weather and illumination, rendering it challenging to precisely control natural conditions in physical attack experiments. This limitation may slightly diminish the experiment's reliability.

Du et al. (2022) conducted a comprehensive examination of the susceptibility of deep neural networks to adversarial perturbations in the context of aerial imagery. The research also investigates the influence of atmospheric variables and distances in the context of physical adversarial attacks. The methodology involves training adversarial patches placed on the ground, followed by capturing images from varying altitudes. These images are then processed using vehicle detectors to assess the efficacy of physical adversarial attacks. The evaluation relies on two key metrics: the Mean Objectness Score (MOS) and the Object Detection Success Rate (ODSR). Experimental findings elucidate that physical adversarial attacks can lead to a discernible reduction in target detection scores, spanning a range of 25–85%. The extent of this reduction is contingent upon specific vehicle and environmental conditions. Furthermore, the study reveals a notably high success rate associated with physical adversarial attacks, with values reaching between 60 and 100%.

Shrestha et al. (2023) introduced a novel approach for generating adversarial patches. This method consists of two key stages. Firstly, within the known white box setting of the DNN UAV target detector, an adversarial patch is created, taking into account potential variations in image brightness and perspective caused by the UAV's shooting angle and height. Subsequently, the generated patch is transferred to another DNN model and architecture. Experimental findings illustrate the efficacy of this proposed adversarial patch generation method in significantly undermining the reliability of the current UAV target detector, with attack success rates reaching as high as 75 and 78%.

In the context of addressing the contextual attack challenge in optical aerial inspection, Lian et al. (2023) presented a novel framework known as Contextual Background Attack (CBA). CBA introduces a fresh approach to background attacks in optical aerial inspection. By concealing the adversarial patch within the area of interest and optimizing the pixels outside this concealed region, the resulting adversarial patch effectively encompasses the critical contextual background area, thereby enhancing the attack's robustness and transferability. Specifically, within the CBA framework, target masking involves obscuring the target area with black pixels to ensure that target recognition remains unaffected during adversarial patch optimization. The CBA framework further enhances attack robustness and transferability by optimizing pixels outside the masked area, enabling the generated adversarial patch to encompass the crucial context background area.

The study conducted by Tang et al. (2023) introduced an innovative technique aimed at generating black-box adversarial attacks in the style of natural weather patterns for optical aerial detectors. The method proffers a departure from conventional adversarial attack strategies by harnessing natural weather-style perturbations to create adversarial instances that exhibit enhanced visual similarity to benign images. This approach proves notably efficacious when juxtaposed with extant methodologies that introduce adversarial perturbations directly onto unaltered images. Notably, the method exhibits a remarkable degree of effectiveness in black-box scenarios, thereby lending practical applicability to real-world contexts.

Xu and Ghamisi (2022) proposed two novel attack techniques: Mixup-Attack and Mixcut-Attack. Mixup-Attack involves the linear interpolation of two distinct samples to produce a new sample, while Mixcut-Attack is a derivative of Mixup-Attack that focuses on interpolation at various points within the input sample. Empirical findings demonstrate that the proposed attack methods can generate high-quality AE capable of deceiving a majority of the most advanced deep neural networks engaged in scene classification and semantic segmentation tasks. Furthermore, the authors have contributed to the UAE-RS dataset, representing a pioneering effort in the field of remote sensing by offering black-box AEs.

In the context of high-angle vehicle detection, Wang et al. (2021) introduced the Dual Attention Suppression (DAS) Attack method, which leverages environmental control and attention mechanisms to create adversarial texture tensors with robust transferability. When applied to a 3D solid object under specific environmental conditions, this method causes misclassification by neural network classifiers. Experimental evaluations involved multiple neural network classifiers and object detectors, including Inception-V3, VGG-19, ResNet-152, and DenseNet. The results demonstrate DAS's formidable attack capabilities and transferability in both digital and physical realms, producing high-quality adversarial texture tensors for diverse target categories and environmental conditions.

For optical sensors, the angle at which adversarial examples are imaged has traditionally been a significant factor affecting the success of attacks. However, recent studies have explored various methods to ensure that adversarial examples remain robust across different detection angles. Zhang et al. (2024) proposed an innovative approach to addressing the challenges posed by physical adversarial attacks in the context of aerial object detection. This study introduces feature-aligned expandable textures designed to deceive detection systems in real-world environments, emphasizing the decoupling of adversarial textures from specific shapes for flexible application. These textures are meticulously refined to align with their surrounding environment, ensuring seamless integration and increased concealment. Extensive experiments in both digital simulations and real-world conditions validate the robustness of

these textures, which maintain high efficacy across different viewing angles and environmental conditions.

Building on these advancements, Zhou et al. (2024) introduced the Direction-Guided Attack (DGA), a novel adversarial attack framework specifically designed to reduce the impact of varying camera angles on attack effectiveness against optical aerial detection systems. By employing affine transformations, the DGA framework aligns the orientation of adversarial patches with the target, ensuring consistent robustness across unpredictable shooting directions. Through comprehensive testing, including the development of the SJTU-4K dataset, the DGA framework demonstrated its ability to deceive a wide range of aerial detectors, highlighting its practical applicability in real-world scenarios. Together, these studies represent significant advancements in the field of adversarial attacks on aerial detection systems, particularly in their ability to maintain attack efficacy despite dynamic changes in viewing angles and environmental conditions.

3.2 Adversarial attack on infrared sensors

Infrared sensors systematically capture thermal signatures and infrared radiation data from targets, employing these distinctive infrared characteristics for object detection. This mechanism, however, renders infrared sensors susceptible to attacks wherein assailants manipulate the thermal signature or infrared radiation of the target. In practical terms, attackers might subtly alter the heat distribution surrounding a target or utilize materials with properties that either absorb or reflect infrared radiation. Such interventions aim to perturb the information, influencing the outcome encapsulated in the object detector.

In line with the principles of infrared imaging outlined in He et al. (2021), attacks on infrared sensors typically manipulate the recognition algorithm by altering the thermal attributes of identified targets. In the context of aerial detection, post-infrared imaging results in a notably reduced target size, posing challenges to the identification process. Consequently, the body of research on adversarial attacks targeting infrared sensors remains relatively limited.

3.2.1 Digital attacks on infrared sensors

In the realm of digital attacks, interference, such as AE, is primarily introduced into infrared remote-sensing images, leading to errors in intelligent recognition. However, this approach does not address the practical challenges that infrared detectors in aerial detection may confront in real-world scenarios.

In Qi et al. (2022), researchers investigated an antagonistic algorithm for infrared remote sensing target recognition based on a multi-channel self-attention mechanism GAN network. The proposed attack mode, evaluated using various examples from different infrared video sequences in diverse remote sensing scenarios, achieved a substantial impact on the targeted detector while introducing minimal adversarial disturbance. Compared to established target recognition countermeasure algorithms, this attack method exhibited clear advantages in terms of physical feasibility, portability, and generation speed.

3.2.2 Physical attacks on infrared sensors

In current research on physical attacks targeting infrared sensors, prevalent methods involve placing temperature-regulating materials on the target to manipulate the thermal radiation characteristics of the detected object. While this approach offers a higher degree of realism, its feasibility for aerial detection remains uncertain.

In the work outlined in Wei X. et al. (2023), researchers introduce a physically viable method for infrared attacks, termed "Adversarial Infrared Patches". This approach manipulates the heat distribution of a target object by affixing a thermally insulating material patch with a specific shape and position onto the object, thus misleading the detection outcomes of the infrared sensor. Experimental findings demonstrate an impressive Attack Success Rate (ASR) exceeding 90% when applied to pedestrian detectors.

Wei H. et al. (2023) introduced the HOTCOLD Block attack method, utilizing temperature control materials, specifically warming paste and cooling paste, to manipulate thermal infrared detectors. This method offers distinct advantages, characterized by its stealthiness, practicality, ease of acquisition and deployment, and resilience to detection by both human observers and detection models. The experimental outcomes demonstrate the effectiveness of HOTCOLD Block in duping thermal infrared detectors, achieving a success rate of over 90% in successful attacks.

In addition to the utilization of cooling materials or patches, Zhu et al. (2021) delineated an attack algorithm employing miniature light bulbs to deceive thermal infrared detectors. The authors devised a cardboard apparatus adorned with small light bulbs, combining both pixel-level patches and Gaussian function patches on the cardboard, along with the incorporation of additional miniature light bulbs. This approach effectively deceived the YOLOV3-based thermal infrared pedestrian detector. Empirical findings underscore the real-world effectiveness of this attack method, which costs less than $5 to implement.

3.3 Adversarial attack on SAR imaging sensors

The SAR imaging sensor employs an antenna for the transmission and reception of radar signals to and from a target. It discerns the target's type by analyzing its characteristics, including reflectivity, shape, and size. However, this mechanism becomes vulnerable to manipulation by attackers seeking to induce the SAR system into producing deceptive imaging results. This can be achieved by strategically introducing reflectors in the target area or adjusting the characteristics of the radar signal. These interventions are orchestrated with the intent to modify the information, thereby influencing the outcome encapsulated in the object detector.

Li et al. (2020) explored the susceptibility of SAR images to AEs and introduced a novel method known as AE Selective Analysis (AESD). AESD is employed to assess various non-target attack algorithms and represents an innovative approach for analyzing the selectivity of AEs. The method is based on the distance between a sample and the nearest decision boundary, offering a rational explanation for why certain adjacent samples exhibit greater vulnerability than others. Geometrically, AESD

signifies that a given raw sample, when subjected to adversarial perturbations, may traverse the nearest decision boundary, resulting in misclassification by the classifier.

3.3.1 Digital attacks on SAR imaging sensors

Digital attacks targeting SAR imaging sensors involve the introduction of adversarial noise into SAR imaging photos. While this approach serves as a foundational concept for physical attacks, it tends to be highly theoretical and often lacks consideration for the practical aspects of reconstructing the electromagnetic signal from the adversarial signal in the physical world.

Zhang F. et al. (2022) provided a comprehensive overview of current adversarial attacks against SAR target recognition networks, encompassing methods such as I-FGSM (BIM), ILCM, and DBA. Additionally, the paper introduces a novel adversarial attack algorithm against SAR target recognition networks, capable of generating AEs to induce incorrect classification results. Experimental findings demonstrate the algorithm's ability to produce high-quality AE, characterized by a higher success rate in deception, enhanced recognition confidence, and a reduced disturbance coverage area. This approach exhibits superior attack efficacy compared to existing methods.

Du et al. (2023) introduced a novel counterattack method named TAN, or Transferable Adversarial Network, specifically designed for efficient black-box attacks. The TAN method achieves its objective by introducing a generator G and an attenuator A to target DNN-based SAR-ATR models. The paper provides a comprehensive account of the TAN method's implementation and training process, detailing the specific procedures for both the generator G and attenuator A. Through experimental verification, the effectiveness of the TAN method is demonstrated, showcasing its ability to efficiently conduct black-box attacks with a remarkable success rate of 98.5%.

In Peng et al. (2021), an attack method targeting SAR image segmentation is introduced. Target Segmentation-based Adversarial Attack (TSAA) is specifically designed to undermine deep learning models employed in SAR image analysis. Distinguishing itself from existing counterattack methods, TSAA places particular emphasis on the practical feasibility of generating perturbations within the target region, rendering it notably effective for SAR images. Empirical findings substantiate the superior performance of TSAA, showcasing its enhanced attack capabilities compared to existing methods across eight common deep-learning model attack tasks.

Peng et al. (2022) introduced the Scattering Model Guided AE (SMGAA) method, aimed at enhancing the adversarial robustness of SAR Automatic Target Recognition (ATR) models based on DNN. In the context of SAR imaging, the scattering response of a target object can be represented as a collection of scattering centers (SC), each characterized by its unique position, amplitude, and phase properties. During an attack, SMGAA is employed to design and generate adversarial scatterers, strategically placing them on the target object. This optimizes their properties, including position, amplitude, and phase, to maximize the impact of the adversarial attack.

3.3.2 Physical attacks on SAR imaging sensors

In contrast to digital attacks, physical attacks inherently entail addressing the fundamental challenge of altering electromagnetic signals in the real world. One existing approach involves the design of a suitable angular reflector strategically positioned near the target, significantly impacting the recognition capabilities of SAR imaging sensors.

In accordance with Zhang L. A. et al. (2022), corner reflectors are trihedral structures composed of highly reflective materials, manifesting as conspicuous bright features in SAR imaging. Notably, a single 1-foot-wide angular reflector can yield a radar cross-section of 500 square meters, sufficiently large to obscure nearby objects. These reflectors exhibit operational viability, owing to their cost-effectiveness and portability, and contribute to the reduction of object visibility in SAR images.

Furthermore, the prospect of training AI systems as "through-angle" reflectors is explored in this paper. Using the RetinaNet architecture, the impact of angle reflectors on SAR object detection is investigated. The model, denoted as the Blue model, is fine-tuned to accurately detect and identify vehicles in the presence of central-corner reflectors, achieving a 56% mean Average Precision (mAP). However, it's worth noting that the paper does not explore the potential for complete occlusion with a large angular reflector. Nevertheless, it is anticipated that a suitably substantial corner reflector, or a combination of such reflectors, could potentially fully obstruct the scene.

3.4 Discussion on adversarial attack in aerial detection

In this chapter, we review the existing research on adversarial attacks against different types of sensors used in aerial detection, such as optical, infrared, and SAR imaging sensors. We categorize the attacks into digital and physical modes and analyze their attack methods, vectors, potentials, and objectives. At the same time, we also provide a table that summarizes the main characteristics and references of the recent adversarial attack research against airborne detection sensors.

In general, adversarial attacks targeting aerial detection face several challenges and limitations. These include the difficulties in conducting realistic and effective physical attacks on aerial detection sensors due to factors such as environmental conditions, target size, and attack feasibility. Additionally, there are trade-offs between attack success rate, stealthiness, and cost, as well as challenges in measuring and evaluating these factors. Furthermore, the transferability and generalization of adversarial attacks across different sensor types, models, and scenarios remain significant obstacles.

4 Adversarial defense in aerial detection

Within the realm of aerial detection, we classify potential adversarial defense methods based on distinct stages of aerial detection missions: pre-takeoff, on-mission, and post-processing.

The resulting as shown in Table 2 encapsulates these findings for reference and analysis. The pre-takeoff phase represents the preparatory stage, where the aircraft or satellite responsible for aerial detection or remote sensing photography is in readiness, allowing for potential modifications and redeployment of onboard algorithms (Figure 2). Therefore, pre-takeoff defense primarily revolves around algorithm design and training.

As the mission progresses into the on-mission stage, the aircraft enters its designated airspace, and the algorithms become fixed. During this phase, the onboard algorithm must autonomously address defense challenges. Hence, the mission's defense strategy centers on algorithm fusion and the detection of AE edges under limited resources. The post-processing stage involves the analysis of images captured by the aircraft. Regardless of whether the onboard intelligent algorithm provides recognition or classification results, these image results are either uploaded to the cloud or returned to the ground. In this context, defense mechanisms primarily encompass ground-based AE detection, defense based on demonstrable robustness, or other approaches such as human-in-the-loop processing.

4.1 Pre-takeoff defense strategies

Pre-takeoff defense methods predominantly involve adversarial training and the refinement of model structures. Adversarial training, a pivotal aspect, encompasses two subtypes: AE or adversarial noise training and randomized training.

4.1.1 Adversarial training

In the context of optical sensors, Lu et al. (2023a) proposed a proactive defense framework that employs a deeply integrated model to enhance UAV vision system robustness. The framework introduces Feature Suppression and Recalibration modules, reactivating suppressed non-robust features, and suggests methods for enhancing loss field orthogonality. Additionally, Lu et al. (2023b) argued that the introduction of AE can enhance model adaptability to adversarial attacks, thereby improving overall model robustness.

For infrared sensors, Spasiano et al. (2022) presented an adversarial training strategy to bolster model robustness. This method incorporates AE into the training process, enhancing the model's resilience against adversarial attacks. Ortiz et al. (2018) proposed the Adaptive Multiband Selection Framework, dynamically selecting optimal band combinations based on evolving adversarial attack scenarios. In multispectral image classification, this approach aids in selecting pertinent input features, reducing data dimensionality, and improving model focus on relevant information.

In the domain of SAR imaging sensors, Peng et al. (2022) delved into AE generation and inversion using a scattering model. This process enables the restoration of original images through an AE reconstruction method, involving AE generation and optimization steps. Additionally, Li et al. (2022) introduced SAR-AD-BagNet, a model employing AE in training to enhance its ability to recognize AE. Leveraging BagNet characteristics and bagging techniques, SAR-AD-BagNet mitigates model variance, enhancing generalization capacity.

4.1.2 Design of model structure

Wang et al. (2023) introduced the Global Feature Attention Network (GFANet) to enhance model robustness in semantic segmentation for aerial images. GFANet employs a dual-branch network architecture and an attention mechanism to dynamically fuse local and global features. This facilitates the effective utilization of global feature information while preserving local details, strengthening the model's robustness and resistance to adversarial attacks.

Experimental evaluations across diverse attack methods highlight GFANet's superior robustness, consistently achieving performance exceeding 80% on an AE test set. This outperformance positions GFANet as a robust solution for aerial image semantic segmentation networks against adversarial challenges.

4.2 On-mission defense strategies

During in-flight detection missions, aircraft heavily rely on onboard intelligent algorithms for adversarial defense. Feasible approaches at this stage involve model integration and partial noise reduction, requiring model compression for reliable edge deployment.

4.2.1 Ensemble model

Utilizing ensemble methods for defending against adversarial perturbations involves constructing multiple classifiers that classify new data points based on weighted or unweighted averages of predictions. These classifiers can be of the same or different types, aiming to mitigate vulnerabilities specific to individual models (Strauss et al., 2017).

Lu et al. (2023a) proposed a deep integration model for reactive defense, aggregating output confidence from multiple DNNs to enhance robustness. They introduce Feature Suppression and Recalibration modules, empirically validating their defense effectiveness. Similarly, Lu et al. (2023b) employed an integrated approach as a reactive defense strategy, averaging output from multiple detectors to improve detection accuracy. Experimental results indicate its efficacy in various scenarios, showcasing flexibility and practicality without requiring model retraining.

However, He et al. (2017) raised concerns about the integrated defense approach, emphasizing the need for further research to enhance its resistance to AE compared to individual components.

4.2.2 Noise reduction

Addressing cloud-like noise or adversarial cloud attacks, Sun et al. (2023) introduced DefenseNet. This defense mechanism trains a model to eliminate clouds by learning features from adversarial cloud images, showcasing effectiveness in object detection. The follow-up work by Sun et al. (2024) further validated the robustness of DefenseNet, expanding its applicability across a broader range of adversarial scenarios and providing additional experimental evidence supporting its efficacy in real-world remote sensing environments. Ortiz et al. (2018) presented the detection network for defending against material responses within the short-wave infrared spectrum, offering robustness without the need for retraining. Detection networks provide the option to refrain from

TABLE 2 Overview of recent adversarial defense research against airborne detection sensors.

References	Mission phase	Sensor type	Defense type	Defense method
Lu et al. (2023a)	Pre-takeoff	Optical	Adversarial training	Feature suppression and recalibration
Lu et al. (2023b)	Pre-takeoff	Optical	Adversarial training	Adversarial example
Spasiano et al. (2022)	Pre-takeoff	Infrared	Adversarial training	Training on faster R-CNN
Ortiz et al. (2018)	Pre-takeoff	Infrared	Adversarial training	Adaptive multiband selection framework
Peng et al. (2022)	Pre-takeoff	SAR	Adversarial training	Scattering model guided adversarial attack
Li et al. (2022)	Pre-takeoff	SAR	Adversarial training	SAR-AD-BagNet
Wang et al. (2023)	Pre-takeoff	Optical	New model structure	Global feature attention network
Lu et al. (2023a)	On-mission	Optical	Ensemble model	Multiple DNNs
Lu et al. (2023b)	On-mission	Optical	Ensemble model	Multiple detectors
Sun et al. (2023)	On-mission	Optical	Noise reduction	DefenseNet
Sun et al. (2024)	On-mission	Optical	Noise reduction	DefenseNet
Ortiz et al. (2018)	On-mission	Infrared	Noise reduction	Detection network
Zhang Z. et al. (2022)	Post-processing	SAR	Noise reduction	Energy-based AE detection
Madry et al. (2017)	Post-processing	Optical	Robustness certification	Linear programming-based
Dong et al. (2020)	Post-processing	Optical	Robustness certification	Gradient-based
Chen and Chu (2023)	Post-processing	Optical	Miscellaneous	Adaptive defense pipeline

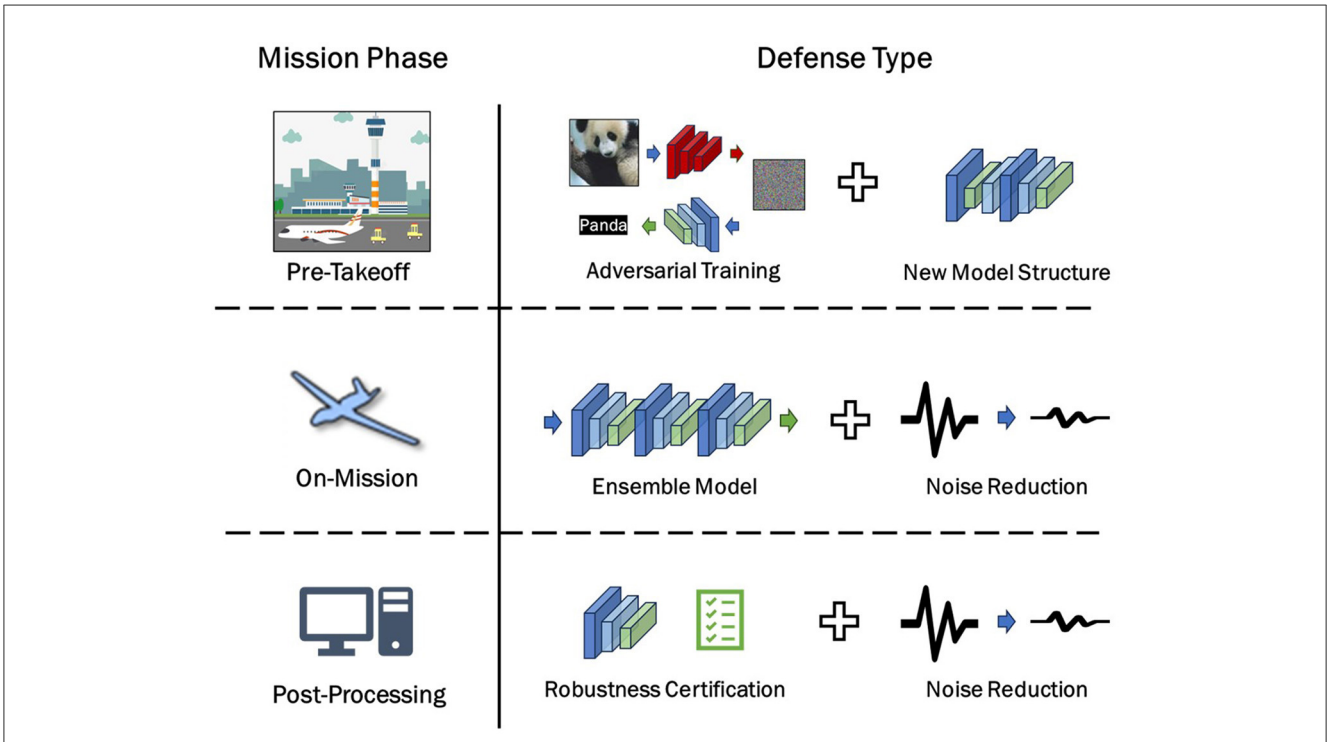

FIGURE 2
Adversarial defenses in aerial detection systems can be categorized based on the operational phase during which they are implemented: Pre-Takeoff, On-Mission, and Post-Processing. In the *Pre-Takeoff phase*, defenses typically involve adversarial training or the development of new model structures. During the *On-Mission phase*, defenses may include the use of ensemble models or noise reduction techniques. In the *Post-processing phase*, robustness certification or further noise reduction methods can be employed to enhance security.

classification when an input image is identified as an adversarial attack, enhancing overall model robustness.

4.3 Post-processing defense strategies

In the post-processing stage, resources initially allocated to onboard intelligent algorithms become available for various defensive measures, including detection, robustness certification, and other methods.

4.3.1 Detection for defense

A novel approach for detecting AE, called Energy-Based AE Detection, is introduced by Zhang Z. et al. (2022). This method considers SAR AE as low-probability instances that deviate from clean datasets and leverages an energy model to capture the intrinsic energy distinctions between SAR AE and clean samples. Importantly, this energy-based approach remains robust even in the presence of perturbations. Building upon this foundation, an energy-based AE detector is proposed without requiring any modifications to the pre-trained model. To enhance the discrimination between clean and AE, energy regularization is applied to fine-tune the pre-trained model. Experimental results demonstrate the method's high accuracy and robustness in detecting AE, outperforming other existing approaches.

4.3.2 Robustness certification

Robustness Certification aims to establish neural network robustness across a range of inputs. This objective can be achieved by determining the maximum disturbance range within the input space. Specifically, techniques for validating robustness fall into two subcategories: linear programming-based methods and gradient-based methods.

Linear programming-based approaches calculate the maximum disturbance range by solving a linear programming problem. This problem aims to identify the largest perturbation range within which the neural network remains robust for all inputs. Linear programming solvers are employed to address this problem. For instance, Madry et al. (2017) employed linear programming to calculate the maximum disturbance range, which is subsequently used for detecting and filtering AE.

Gradient-based methods determine the maximum disturbance range by computing the gradient of the neural network. The underlying concept is that if the neural network exhibits a small gradient for a given input, the input is considered robust. Therefore, the maximum disturbance range can be estimated by evaluating the network's gradient concerning the input. For instance, Dong et al. (2020) employed gradient descent to compute the maximum disturbance range, which is then used for detecting and filtering AE.

Proving robustness typically incurs high computational costs, primarily because it involves determining the maximum perturbation range in the input space, which can be extensive. Moreover, robustness certification methods often require the use of intricate mathematical tools such as linear programming and convex optimization, further elevating computational costs. In summary, this method is not well-suited for real-time applications like aircraft operations due to its resource-intensive nature. Instead, it is more suitable as a post-processing defense mechanism.

4.3.3 Miscellaneous

Other post-processing methods include techniques like re-identification or incorporating human-in-the-loop during classification, as well as data fusion from multiple sensors for improved recognition.

In Chen and Chu (2023)'s research, an Adaptive Defense Pipeline (ADP) is introduced to enhance the robustness of AI algorithms in target detection. ADP adjusts the weight coefficients of detection results from multiple sensors to synthesize these results based on different weather conditions and further incorporates a secondary confirmation step. This approach effectively fortifies the defense against adversarial attacks and elevates the accuracy of target detection. To validate the effectiveness of the proposed method, a comparison is made between the outcomes of traditional single-sensor aerial detection and ADP-weighted detection. The results affirm that this method significantly enhances the efficiency of aerial detection using artificial intelligence algorithms in adversarial environments.

The document emphasizes the pivotal role of human-in-the-loop approaches in counter-defense strategies. By amalgamating the knowledge and expertise of human experts with machine learning algorithms, detection and defense against attacks can be considerably ameliorated. Moreover, human involvement contributes to enhanced detection accuracy through the secondary validation of detection results, allowing for manual intervention when algorithms face challenges.

However, it's crucial to acknowledge that this method constitutes a defense strategy that comes at the cost of artificial intelligence efficiency. While it does provide a certain degree of protection against the misleading effects of AE attacks on identification results, it significantly diminishes overall processing time and efficiency.

4.4 Discussion on adversarial defense in aerial detection

In this chapter, we classified the potential adversarial defense methods based on the distinct stages of aerial detection missions: pre-takeoff, on-mission, and post-processing. We discussed the advantages and limitations of different defense strategies, such as adversarial training, model structure design, ensemble model, noise reduction, robustness certification, and adaptive defense pipeline. Meanwhile, We have also provided a table that summarizes the main characteristics and references of the recent adversarial defense research against airborne detection sensors.

In general, for adversarial defense, we believe the following challenges remain: First, the robustness and adaptability of adversarial defense methods to various attack scenarios, and how to test and verify them. Second is the compatibility and integration of adversarial defense methods with existing aerial detection systems, and how to optimize them. In addition to that, the ethical and social

implications of adversarial defense methods for aerial detection, and how to ensure their safety and accountability.

5 Outlook and future directions

There's a lot of work that's been done to show adversarial attack and defense methods for aerial detection systems. In the field of adversarial attack, researchers have explored various methods to deceive aerial detection sensors, such as optical, infrared, and SAR imaging sensors, by introducing subtle perturbations or patches in the digital or physical domain. These attacks aim to cause misclassification, false detection, or occlusion of the target objects, affecting the accuracy and reliability of the aerial detection systems. In the adversarial defense field, researchers have proposed various strategies to enhance the robustness and resilience of aerial detection models, such as adversarial training, model ensemble, noise reduction, robustness certification, and adaptive defense pipeline. These defenses aim to improve the performance and generalization of the aerial detection models under adversarial scenarios.

However, there are still many open problems and directions that need further exploration and investigation in this field. In the following, we highlight some of the most important and promising ones.

1. **Lack of physical attack testing in real scenarios**
   Most of the existing studies focus on simulated or idealized attacks, which may not reflect the practical challenges and constraints in real-world scenarios, such as weather conditions, sensor noise, target dynamics, among others. More studies are needed to investigate the feasibility and effectiveness of real-world attacks and their impact on aerial detection applications.
2. **Lack of defense evaluation**
   Most of the existing studies evaluate the defense methods based on specific attack methods or datasets, which may not capture the diversity and transferability of adversarial examples. More studies are needed to develop comprehensive and standardized evaluation metrics and benchmarks for aerial detection defense methods and to compare their strengths and limitations.
3. **Lack of high-quality platforms**
   Most of the existing studies rely on low-quality or limited platforms, such as low-resolution images, small-scale datasets, or simple models, which may not reflect the state-of-the-art or the potential of aerial detection systems. More studies are needed to leverage high-quality or large-scale platforms, such as high-resolution images, large-scale datasets, or advanced models, to explore the challenges and opportunities of aerial detection adversarial attack, and defense.

To solve these problems, we will carry out corresponding research in two research directions in the future. On the one hand, we will use aircraft platforms equipped with different sensors to conduct real experiments on ground targets and collect real multi-angle, multi-scale, and multi-modal attack and defense data as a way to enhance the efficiency of adversarial attack and defense. On the other hand, due to the importance of physical experiments and the difficulty of the actual operation of aerial detection experiments, we are wondering whether there is a digital twin solution. That is, the natural environmental conditions, detection mechanism, effect of adversarial attack, and other elements of aerial detection are restored in the virtual environment, and adversarial experiments are conducted in a more convenient, low-cost, and efficient form to accelerate the research on attack and defense methods.

6 Conclusion

This article aimed to review the adversarial attacks and defenses in the context of aerial detection, which is an important problem in practical world for the deployment of deep learning. We presented a systematic and comprehensive investigation of both digital and physical adversarial attacks based on the most widely used sensing principles, delivering an in-depth analysis of the evolution and progress within this research domain. We also systematically investigated the existing adversarial defenses from the perspectives of the main operational phases of aerial detection missions, such as pre-takeoff, on-mission, and post-processing. Moreover, we engaged in an extensive discussion regarding the challenges posed by real-world intelligent aerial detection and their implications on subsequent operational stages and pinpointed several directions for future studies such as real-world attacks, defense, high-quality platforms, among others. This article provides a valuable resource and a reference guide for researchers and practitioners in the field of aerial detection, as it highlights the importance and urgency of addressing the adversarial robustness of aerial detection, which involves safety-critical applications such as defense, agriculture, mining, and mapping. We also suggest some potential benefits of adversarial attacks and defenses, such as enhancing existing algorithms, improving model adaptability, and exploring new attack and defense scenarios. However, we also acknowledge the limitations and challenges of the current research, such as the lack of standardized datasets and benchmarks, the difficulty of evaluating the effectiveness and efficiency of adversarial attacks and defenses, and the ethical and legal issues of adversarial manipulation. Therefore, we call for more collaboration and communication among researchers, practitioners, and policymakers to advance more robust and secure aerial detection systems in the future.

Author contributions

YC: Writing – original draft, Writing – review & editing. SC: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2018). "Synthesizing robust adversarial examples," in *International Conference on Machine Learning* (Stockholm: Stockholmsmässan).

Boloor, A., He, X., Gill, C., Vorobeychik, Y., and Zhang, X. (2019). "Simple physical adversarial examples against end-to-end autonomous driving models," in *2019 IEEE International Conference on Embedded Software and Systems (ICESS)* (Las Vegas, NV).

Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., et al. (2019a). "Adversarial sensor attack on lidar-based perception in autonomous driving," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London).

Cao, Y., Xiao, C., Yang, D., Fang, J., Yang, R., Liu, M., et al. (2019b). Adversarial objects against lidar-based autonomous driving systems. *arXiv* [preprint]. doi: 10.48550/arXiv.1907.05418

Chang, K.-W., He, H., Jia, R., and Singh, S. (2021). "Robustness and adversarial examples in natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts* (Punta Cana).

Chen, L., Li, H., Zhu, G., Li, Q., Zhu, J., Huang, H., et al. (2020). Attack selectivity of adversarial examples in remote sensing image scene classification. *IEEE Access*. 8, 137477–137489. doi: 10.1109/ACCESS.2020.3011639

Chen, Y., and Chu, S. (2023). "Adversarial defense in aerial detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Vancouver, BC).

Cisse, M. M., Adi, Y., Neverova, N., and Keshet, J. (2017). Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. *Adv. Neur. Inf. Process. Syst*.

Connor, D., Martin, P. G., and Scott, T. B. (2016). Airborne radiation mapping: overview and application of current and future aerial systems. *Int. J. Remote Sens*. 37, 5953–5987. doi: 10.1080/01431161.2016.1252474

Crawford, F. (1998). Electro-optical sensors overview. *IEEE Aerospace Electron. Syst. Mag*. 13, 17–24. doi: 10.1109/62.722416

Den Hollander, R., Adhikari, A., Tolios, I., van Bekkum, M., Bal, A., Hendriks, S., et al. (2020). "Adversarial patch camouflage against aerial detection," in *Artificial Intelligence and Machine Learning in Defense Applications II*.

Dhillon, A., and Verma, G. K. (2020). Convolutional neural network: a review of models, methodologies and applications to object detection. *Progr. Artif. Intell*. 9, 85–112. doi: 10.1007/s13748-019-00203-0

Dong, X., Zhu, Y., Zhang, Y., Fu, Z., Xu, D., Yang, S., et al. (2020). "Leveraging adversarial training in self-learning for cross-lingual text classification," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Xi'an).

Du, A., Chen, B., Chin, T.-J., Law, Y. W., Sasdelli, M., Rajasegaran, R., et al. (2022). "Physical adversarial attacks on an aerial imagery object detector," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI).

Du, M., Sun, Y., Sun, B., Wu, Z., Luo, L., Bi, D., et al. (2023). TAN: a transferable adversarial network for DNN-based UAV SAR automatic target recognition models. *Drones*. 7:205. doi: 10.3390/drones7030205

Everitt, J., Escobar, D., Villarreal, R., Noriega, J., and Davis, M. (1991). Airborne video systems for agricultural assessment. *Remote Sens. Environ*. 32, 155–167. doi: 10.1016/0034-4257(91)90015-X

Fingas, M., and Brown, C. (2001). Review of ship detection from airborne platforms. *Can. J. Remote Sens*. 27, 379–385. doi: 10.1080/07038992.2001.10854880

Galvez, R. L., Bandala, A. A., Dadios, E. P., Vicerra, R. R. P., and Maningo, J. M. Z. (2018). "Object detection using convolutional neural networks," in *TENCON 2018-2018 IEEE Region 10 Conference* (Jeju).

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint* arXiv:1412.6572.

Guo, C., Gardner, J., You, Y., Wilson, A. G., and Weinberger, K. (2019). "*Simple black-box adversarial* attacks," in International Conference on Machine Learning.

Guo, J., Bao, W., Wang, J., Ma, Y., Gao, X., Xiao, G., et al. (2023). A comprehensive evaluation framework for deep model robustness. *Pattern Recognit*. 137:109308. doi: 10.1016/j.patcog.2023.109308

He, W., Wei, J., Chen, X., Carlini, N., and Song, D. (2017). "Adversarial example defense: ensembles of weak defenses are not strong," in *11th USENIX Workshop on Offensive Technologies (WOOT 17)* (Vancouver, BC).

He, Y., Deng, B., Wang, H., Cheng, L., Zhou, K., Cai, S., et al. (2021). Infrared machine vision and infrared thermography with deep learning: a review. *Infrared Phys. Technol*. 116:103754. doi: 10.1016/j.infrared.2021.103754

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. (2017). Query-efficient black-box adversarial examples (superceded). *arXiv* [preprint]. doi: 10.48550/arXiv.1712.07113

Jan, S. T., Messou, J., Lin, Y.-C., Huang, J.-B., and Wang, G. (2019). "Connecting the digital and physical world: improving the robustness of adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI).

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. (2019). "Memguard: defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London).

Kong, Z., Guo, J., Li, A., and Liu, C. (2020). "Physgan: generating physical-world-resilient adversarial examples for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA).

Kurakin, A., Goodfellow, I., and Bengio, S. (2016a). *Adversarial Examples in the Physical World*. Toulon: Curran Associates, Inc.

Kurakin, A., Goodfellow, I., and Bengio, S. (2016b). Adversarial machine learning at scale. *arXiv* [preprint]. doi: 10.48550/arXiv.1611.01236

Kurakin, A., Goodfellow, I. J., and Bengio, S. (2018). "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security* (London: Chapman & Hall/CRC), 99–112.

Li, H., Huang, H., Chen, L., Peng, J., Huang, H., Cui, Z., et al. (2020). Adversarial examples for cnn-based sar image classification: an experience study. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. 14:3038683. doi: 10.1109/JSTARS.2020.3038683

Li, P., Hu, X., Feng, C., Shi, X., Guo, Y., and Feng, W. (2022). SAR-AD-BagNet: an interpretable model for sar image recognition based on adversarial defense. *IEEE Geosci. Remote Sens. Lett*. 20:3230243. doi: 10.1109/LGRS.2022.3230243

Lian, J., Mei, S., Zhang, S., and Ma, M. (2022). Benchmarking adversarial patch against aerial detection. *IEEE Transact. Geosci. Remote Sens*. 60:3225306. doi: 10.1109/TGRS.2022.3225306

Lian, J., Wang, X., Su, Y., Ma, M., and Mei, S. (2023). CBA: contextual background attack against optical aerial detection in the physical world. *IEEE Transact. Geosci. Remote Sens*. 61, 1–16. doi: 10.1109/TGRS.2023.3264839

Liu, A., Huang, T., Liu, X., Xu, Y., Ma, Y., Chen, X., et al. (2020a). "Spatiotemporal attacks for embodied agents," in *European Conference on Computer Vision* (Glasgow).

Liu, A., Liu, X., Yu, H., Zhang, C., Liu, Q., and Tao, D. (2021). Training robust deep neural networks via adversarial noise propagation. *IEEE Transact. Image Process*. 30:82317. doi: 10.1109/TIP.2021.3082317

Liu, A., Tang, S., Chen, X., Huang, L., Qin, H., Liu, X., et al. (2023a). Towards defending multiple lp-norm bounded adversarial perturbations via gated batch normalization. *Int. J. Comput. Vis*. 132, 1881–1898. doi: 10.1007/s11263-023-01884-w

Liu, A., Tang, S., Liang, S., Gong, R., Wu, B., Liu, X., et al. (2023b). "Exploring the relationship between architecture and adversarially robust generalization," in *CVPR* (Vancouver, BC).

Liu, A., Wang, J., Liu, X., Cao, b., Zhang, C., and Yu, H. (2020b). "Bias-based universal adversarial patch attack for automatic check-out," in *European Conference on Computer Vision* (Glasgow).

Liu, J., Zhang, W., Zhang, Y., Hou, D., Liu, Y., Zha, H., et al. (2019). "Detection based defense against adversarial examples from the steganalysis point of view," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA).

Liu, Y., Chen, X., Liu, C., and Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. *arXiv* [preprint]. doi: 10.48550/arXiv.1611.02770

Lu, Z., Sun, H., Ji, K., and Kuang, G. (2023a). Adversarial robust aerial image recognition based on reactive-proactive defense framework with deep ensembles. *Remote Sens*. 15:4660. doi: 10.3390/rs15194660

Lu, Z., Sun, H., and Xu, Y. (2023b). Adversarial robustness enhancement of uav-oriented automatic image recognition based on deep ensemble models. *Remote Sens*. 15:3007. doi: 10.3390/rs15123007

Maathuis, B. H., and Genderen, J. V. (2004). A review of satellite and airborne sensors for remote sensing based detection of minefields and landmines. *Int. J. Remote Sens*. 25, 5201–5245. doi: 10.1080/01431160412331270803

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv* [preprint]. doi: 10.48550/arXiv.1706.06083

Massoli, F. V., Carrara, F., Amato, G., and Falchi, F. (2021). Detection of face recognition adversarial attacks. *Comp. Vis. Image Understand*. 202:103103. doi: 10.1016/j.cviu.2020.103103

Morris, J. X., Lifland, E., Yoo, J. Y., and Qi, Y. (2020). "Textattack: a framework for adversarial attacks in natural language processing," in *Proceedings of the 2020 EMNLP* (Barceló Bávaro Convention Centre).

Norton, P. R. (1991). Infrared image sensors. *Opt. Eng.* 30:56001. doi: 10.1117/12.56001

Ortiz, A., Fuentes, O., Rosario, D., and Kiekintveld, C. (2018). "On the defense against adversarial examples beyond the visible spectrum," in *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)* (Los Angeles, CA).

Papernot, N., Mcdaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519. doi: 10.1145/3052973.3053009

Peng, B., Peng, B., Zhou, J., Xie, J., and Liu, L. (2022). Scattering model guided adversarial examples for sar target recognition: attack and defense. *IEEE Transact. Geosci. Remote Sens.* 60, 1–17. doi: 10.1109/TGRS.2022.3213305

Peng, B., Xia, J., Peng, B., Zhou, J., Zhi, S., and Liu, Y. (2021). "Target segmentation based adversarial attack for sar images," in *2021 CIE International Conference on Radar (Radar)* (Haikou: IEEE), 2146–2150.

Qi, J., Zhang, Y., Wan, P., Li, Y., Liu, X., Yao, A., et al. (2022). Object detection adversarial attack for infrared imagery in remote sensing. *Aero Weaponry* 29, 47–53.

Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., and Raffel, C. (2019). "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *International Conference on Machine Learning* (Long Beach, CA).

Qiu, S., Liu, Q., Zhou, S., and Huang, W. (2022). Adversarial attack and defense technologies in natural language processing: a survey. *Neurocomputing* 492, 278–307. doi: 10.1016/j.neucom.2022.04.020

Samizade, S., Tan, Z.-H., Shen, C., and Guan, X. (2020). "Adversarial example detection by classification for deep speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona).

Schönherr, L., Kohls, K., Zeiler, S., Holz, T., and Kolossa, D. (2018). Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv* [preprint]. doi: 10.14722/ndss.2019.23288

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. (2019). A general framework for adversarial examples with objectives. *ACM Transact. Privacy Sec.* 22, 1–30. doi: 10.1145/3317611

Shrestha, S., Pathak, S., and Viegas, E. K. (2023). "Towards a robust adversarial patch attack against unmanned aerial vehicles object detection," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 3256–3263. doi: 10.1109/IROS55552.2023.10342460

Smith, L., and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. *arXiv* [preprint].

Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., et al. (2018). "Physical adversarial examples for object detectors," *in 12th USENIX Workshop on Offensive Technologies (WOOT 18)* (Baltimore, MD).

Spasiano, F., Gennaro, G., and Scardapane, S. (2022). "Evaluating adversarial attacks and defences in infrared deep learning monitoring systems," in *2022 International Joint Conference on Neural Networks (IJCNN)* (Rome).

Strauss, T., Hanselmann, M., Junginger, A., and Ulmer, H. (2017). Ensemble methods as a defense to adversarial perturbations against deep neural networks. *arXiv* [preprint]. doi: 10.48550/arXiv.1709.03423

Sun, H., Fu, L., Li, J., Guo, Q., Meng, Z., Zhang, T., et al. (2023). Defense against adversarial cloud attack on remote sensing salient object detection. *arXiv* [preprint]. doi: 10.1109/WACV57701.2024.00816

Sun, H., Fu, L., Li, J., Guo, Q., Meng, Z., Zhang, T., et al. (2024). "Defense against adversarial cloud attack on remote sensing salient object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI), 8345–8354.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., and Fergus, R. (2013). *Intriguing Properties of Neural Networks* (Banff, AB).

Tang, G., Yao, W., Jiang, T., Zhou, W., Yang, Y., and Wang, D. (2023). Natural weather-style black-box adversarial attacks against optical aerial detectors. *IEEE Transact. Geosci. Remote Sens.* doi: 10.1109/TGRS.2023.3315053

Tu, J., Ren, M., Manivasagam, S., Liang, M., Yang, B., Du, R., et al. (2020). "Physically realizable adversarial examples for lidar object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA).

Vakhshiteh, F., Nickabadi, A., and Ramachandra, R. (2021). Adversarial attacks against face recognition: a comprehensive study. *IEEE Access.* 9, 92735–92756. doi: 10.1109/ACCESS.2021.3092646

Wang, J., Liu, A., Yin, Z., Liu, S., Tang, S., and Liu, X. (2021). "Dual attention suppression attack: generate adversarial camouflage in physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN).

Wang, X., Mei, S., Lian, J., and Lu, Y. (2024). Fooling aerial detectors by background attack via dual-adversarial-induced error identification. *IEEE Transact. Geosci. Remote Sens.* doi: 10.1109/TGRS.2024.3386533

Wang, Z., Wang, B., Liu, Y., and Guo, J. (2023). Global feature attention network: Addressing the threat of adversarial attack for aerial image semantic segmentation. *Remote Sensing*. 15:1325. doi: 10.3390/rs15051325

Wei, H., Wang, Z., Jia, X., Zheng, Y., Tang, H., Satoh, S., et al. (2023). "Hotcold block: fooling thermal infrared detectors with a novel wearable design," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington, DC).

Wei, X., Yu, J., and Huang, Y. (2023). "Physically adversarial infrared patches with learnable shapes and locations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, BC).

Wilkening, D. A. (2004). Airborne boost-phase ballistic missile defense. *Sci. Glob. Sec.* 12, 1–67. doi: 10.1080/08929880490464649

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., and Yuille, A. (2017). "Adversarial examples for semantic segmentation and object detection," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice).

Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., et al. (2020). "Adversarial t-shirt! evading person detectors in a physical world," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V 16* (Glasgow).

Xu, Y., and Ghamisi, P. (2022). Universal adversarial examples in remote sensing: methodology and benchmark. *IEEE Transact. Geosci. Remote Sens.* 60, 1–15. doi: 10.1109/TGRS.2022.3156392

Yang, K., Liu, J., Zhang, C., and Fang, Y. (2018). "Adversarial examples against the deep learning based network intrusion detection systems," in *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)* (Los Angeles, CA).

Yu, H., Liu, A., Li, G., Yang, J., and Zhang, C. (2021). Progressive diversified augmentation for general robustness of dnns: a unified approach. *IEEE Transact. Image Process.* 30, 8955–8967. doi: 10.1109/TIP.2021.3121150

Zhang, B., Tondi, B., and Barni, M. (2020). Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability. *Comp. Vis. Image Understand*. 2020:102988. doi: 10.1016/j.cviu.2020.102988

Zhang, C., Liu, A., Liu, X., Xu, Y., Yu, H., Ma, Y., et al. (2021). Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transacti. Image Process.* 30:83. doi: 10.1109/TIP.2020.3042083

Zhang, F., Meng, T., Xiang, D., Ma, F., Sun, X., and Zhou, Y. (2022). Adversarial deception against sar target recognition network. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.* 15, 4507–4520. doi: 10.1109/JSTARS.2022.3179171

Zhang, L. A., Hartnett, G. S., Aguirre, J., Lohn, A. J., Khan, I., Herron, M., et al. (2022). *Operational Feasibility of Adversarial Attacks Against Artificial Intelligence*. Santa Monica, CA: RAND Corporation.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: a survey. *ACM Transact. Intell. Syst. Technol.* 11, 1–41. doi: 10.1145/3374217

Zhang, Y., Chen, J., Peng, Z., Dang, Y., Shi, Z., and Zou, Z. (2024). Physical adversarial attacks against aerial object detection with feature-aligned expandable textures. *IEEE Transact. Geosci. Remote Sens.* 62:3426272. doi: 10.1109/TGRS.2024.3426272

Zhang, Y., Zhang, Y., Qi, J., Bin, K., Wen, H., Tong, X., et al. (2022). Adversarial patch attack on multi-scale object detection for uav remote sensing images. *Remote Sensing.* 14:5298.doi: 10.20944/preprints202210.0131.v1

Zhang, Z., Gao, X., Liu, S., Peng, B., and Wang, Y. (2022). Energy-based adversarial example detection for SAR images. *Remote Sens.* 14:5168. doi: 10.3390/rs14205168

Zhou, H., Li, W., Kong, Z., Guo, J., Zhang, Y., Yu, B., et al. (2020). "Deepbillboard: Systematic physical-world testing of autonomous driving systems," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (Seoul).

Zhou, Y., Sun, S., Jiang, X., Xu, G., Hu, F., Zhang, Z., et al. (2024). DGA: direction-guided attack against optical aerial detection in camera shooting direction agnostic scenarios. *IEEE Transact. Geosci. Remote Sens.* doi: 10.1109/TGRS.2024.3387486

Zhu, X., Li, X., Li, J., Wang, Z., and Hu, X. (2021). "Fooling thermal infrared pedestrian detectors in real world using small bulbs," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Vancouver, BC).

Zhu, Z.-A., Lu, Y.-Z., and Chiang, C.-K. (2019). "Generating adversarial examples by makeup attacks on face recognition," in *2019 IEEE International Conference on Image Processing (ICIP)* (Taipei).