Digital-to-Physical Visual Consistency Optimization for Adversarial Patch Generation in Remote Sensing Scenes

Jianqi Chen, Yilan Zhang, Chenyang Liu, Keyan Chen, Zhengxia Zou, and Zhenwei Shi*, *Senior Member, IEEE*

***Abstract*—In contrast to digital image adversarial attacks, adversarial patch attacks involve physical operations that project crafted perturbations into real-world scenarios. During the digital-to-physical transition, adversarial patches inevitably undergo information distortion. Existing approaches focus on data augmentation and printer color gamut regularization to improve the generalization of adversarial patches to the physical world. However, these efforts overlook a critical issue within the adversarial patch crafting pipeline—namely, the significant disparity between the appearance of adversarial patches during the digital optimization phase and their manifestation in the physical world. This unexplored concern, termed "Digital-to-Physical Visual Inconsistency", introduces inconsistent objectives between the digital and physical realms, potentially skewing optimization directions for adversarial patches. To tackle this challenge, we propose a novel harmonization-based adversarial patch attack. Our approach involves the design of a self-supervised harmonization method, seamlessly integrated into the adversarial patch generation pipeline. This integration aligns the appearance of adversarial patches overlaid on digital images with the imaging environment of the background, ensuring a consistent optimization direction with the primary physical attack goal. We validate our method through extensive testing on the aerial object detection task. To enhance the controllability of environmental factors for method evaluation, we construct a dataset of 3D simulated scenarios using a graphics rendering engine. Extensive experiments on these scenarios demonstrate the efficacy of our approach. Our code and dataset are publicly accessible at https://github.com/WindVChen/VCO-AP.**

***Index Terms*—Physical adversarial attack, adversarial patch, object detection, remote sensing.**

I. Introduction

DEEP learning models have exhibited remarkable performance in various domains, such as medical image analysis [1], [2], remote sensing image recognition [3], [4], denoising [5], captioning [6]–[8], *etc*. However, Szegedy *et al*. [9] discovered that these high-performing models can be vulnerable to carefully crafted perturbations, leading to erroneous predictions. This vulnerability has sparked concerns about the robustness of artificial intelligence systems.

To comprehensively understand and enhance model robustness, numerous studies [10]–[15] have delved into adversarial attack research, probing intelligent models for potential vulnerabilities. The majority of these investigations focus on digital perturbations [10], [11], involving the generation of perturbation maps applied to digital images, manipulating pixels to deceive recognition models. However, these digital attacks may not accurately simulate real-world scenarios. In practical situations, especially in domains like remote sensing, attackers often face constraints preventing direct manipulation of digital images within target systems due to sophisticated security mechanisms. Furthermore, applying a perturbation map across an entire physical scenario, particularly in remote sensing where images cover vast areas, is impractical and cost-prohibitive. To address these challenges in real-world scenarios, attackers [16]–[21] commonly resort to adversarial patches [22]. Unlike full-image digital perturbations, adversarial patches are small patches in regular shapes like squares or circles. Attackers can easily introduce these *physical perturbations* into the real environment using printing and inkjet technologies, indirectly affecting the digital input to visual models through the imaging of visual sensors. Due to pragmatic considerations and feasibility in real-world scenarios, our work focuses on physical adversarial patch attacks.

Typically, common adversarial patch attack crafting pipeline involve two stages: crafting the patch in the digital space and physically printing and applying it in the real world. In contrast to digital perturbations, information loss consistently poses a primary challenge for physical perturbations. Factors such as color gamut differences (*e.g*., RGB *vs.* CMYK) between the printer and the digital machine [23], camera imaging resolution [24], and environmental conditions [25] can lead to information loss within physical perturbations, inevitably impacting attack performance.

To mitigate information loss and ensure the generalization of digitally optimized adversarial patches to the physical world, various effective strategies have been employed within the adversarial patch domain. Sharif *et al*. [23] introduced the non-printability score (NPS) to align perturbation colors with printer color gamuts and minimized total variation (TV) to emulate natural patterns. Athalye *et al*. [25] proposed an expectation over transformation (EOT) strategy to enhance perturbation generalization through data augmentation.

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under the Grants 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (email: shizhenwei@buaa.edu.cn))*

Jianqi Chen, Yilan Zhang, Chenyang Liu, Keyan Chen, and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China.

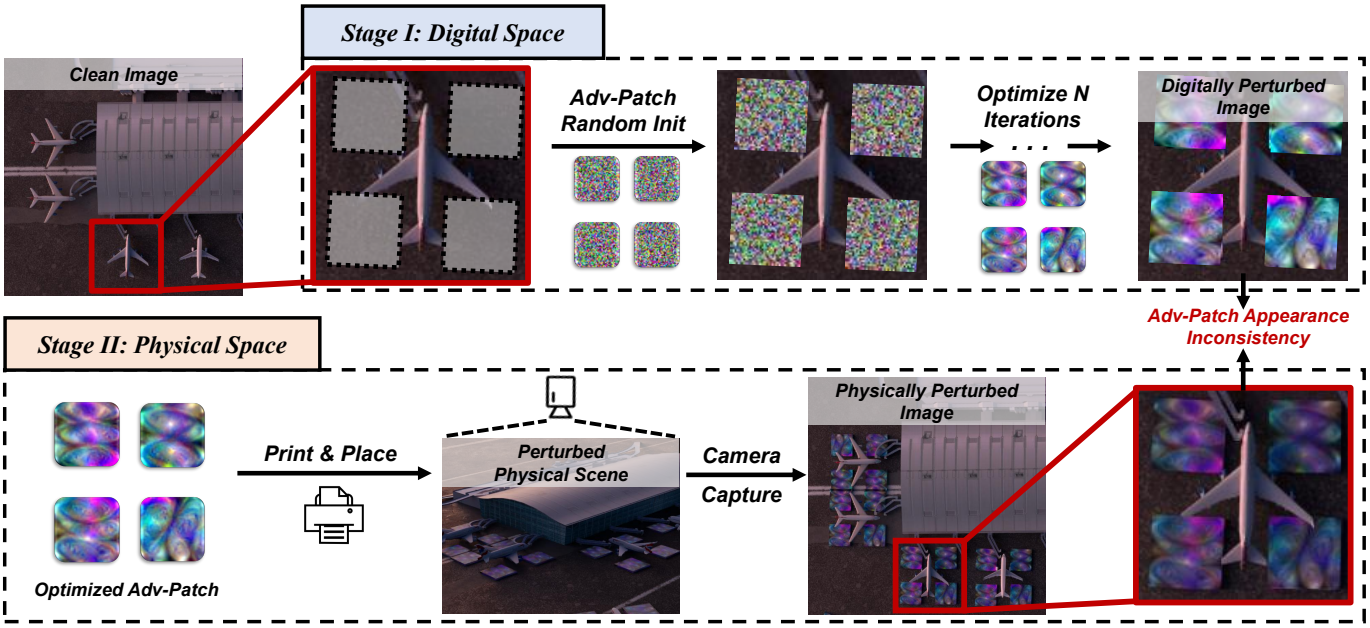

Fig. 1. **The visual illustration of the *digital-to-physical visual inconsistency* problem.** In the common pipeline of crafting the physical adversarial patch, there are two stages. The first stage takes place in the digital space, where the adversarial patch is directly overlaid on the captured images and iteratively optimized. Then, in the second stage, which takes place in the physical space, the optimized adversarial patch is printed out and placed into the real-world scenario. The final physical attack performance is then evaluated by capturing images from the physically perturbed scenario. However, due to the ignorance of the background environment, there is a significant appearance gap between the optimization direction in the digital space and the ultimate attack goal in the physical space. Please zoom in for a better view. ("Adv-Patch" denotes adversarial patches.)

Eykholt *et al*. [16] expanded on EOT by incorporating more diverse physical backgrounds. While these approaches enhance physical attack performance, they neglect a crucial aspect in the adversarial patch optimization phase—significant discrepancies in the appearance of adversarial patches between the digital and target physical domains.

We term this unexplored issue as *digital-to-physical visual inconsistency* and present a visual illustration in Fig. 1. The problem arises from the two-stage nature of the common physical adversarial patch crafting pipeline. Since digital-to-physical transitions are non-differentiable, gradients from final physical attack results cannot directly backpropagate for patch optimization. Consequently, the conventional approach optimizes the adversarial patch by directly overlaying it onto digital images. This straightforward operation disregards environmental factors (*e.g*., tone, illumination) in digital images, resulting in perceptible inconsistencies between the appearance of the adversarial patch in the digital space and in the physical space. This inconsistency reflects a misalignment between the ultimate physical attack goal and the digital optimization goal. Such optimization direction bias can compromise the generalization of adversarial patches from the digital space to the real physical world.

To address the digital-to-physical visual inconsistency problem, we propose a novel harmonization-based adversarial patch attack. Specifically, we break down the adversarial patch generation within the digital space into two sub-stages. The first sub-stage focuses on training an image harmonization method, while the second sub-stage incorporates the pre-trained harmonization module into the adversarial patch optimization to align the appearance of the patch with the background environment of the digital image. In this way, we correct the former optimization direction bias and can expect to simulate real-world attack scenarios in the digital space, ensuring robust attack performance in the physical world. To preserve the content structure of the adversarial patch, we refer to the harmonization network in [26] to predict parameters of a 3D lookup table (LUT) [27], making the model both lightweight and interpretable. Given that there is no large-scale image harmonization dataset in the remote sensing field, we propose a self-supervised strategy for training the harmonization network.

To ensure a reliable assessment of our method's effectiveness, we validate our attacks on aerial object detection, a prevalent task in remote sensing, specifically targeting rotated object detectors, which have demonstrated superior suitability and accuracy in remote sensing compared to traditional horizontal detection methods [28]. To validate our method effectively, comprehensive testing under various physical environmental factors is essential. Acknowledging the complexities of controlling real-world physical environments, we introduce a dataset of simulated 3D scenarios created using a graphics rendering engine. These scenarios faithfully replicate environmental factors such as illumination, tone, and shadow [29], offering robust evaluation platforms for physical adversarial attacks. Our evaluation primarily focuses on white-box attacks while also exploring the impact of our method on the transferability of attacks to black-box detectors. Notably, our visual consistency optimization acts as a versatile *plug-and-play* component, enabling seamless integration with techniques like EOT to further enhance the physical attack performance of adversarial patches.

In summary, our contributions encompass:

- We bring attention to the issue of *digital-to-physical visual inconsistency* in the optimization of adversarial patches, shedding light on a previously overlooked problem in this domain.
- We propose a novel harmonization-based adversarial patch attack to address the identified inconsistency. Additionally, we introduce a self-supervised training strategy to tackle the challenge of missing image harmonization datasets in the remote sensing field.
- We curate a distinctive dataset of 3D simulated scenarios using a graphics rendering engine. These scenarios serve as a robust platform for evaluating the performance of physical adversarial attacks, providing a solution to the difficulty of establishing a controllable physical environment.
- Our attacks are validated on rotated object detectors in remote sensing. Through experiments, we not only showcase the effectiveness of our method but also emphasize its positive impact on the transferability of attacks across black-box models. Furthermore, our approach is validated as flexible, enabling seamless integration with existing technologies to create robust adversarial patches.

II. Related Works

A. Adversarial Attacks.

The landscape of adversarial attack methods can be broadly categorized into digital attacks and physical attacks. Digital attack methods primarily manipulate the digital image input to compromise the victim model's performance. Szegedy *et al.* [9] were pioneers in identifying vulnerabilities in deep models and introduced the use of L-BFGS for optimizing adversarial perturbations. Goodfellow *et al.* [10] argued that even a single step of gradient backpropagation could deceive the victim model effectively and proposed the efficient fast gradient sign method (FGSM). Kurakin *et al.* [12] further extended FGSM into multiple iterations and proposed I-FGSM. Madry *et al.* [11] framed adversarial attack and defense as a zero-sum game, proposing a robust adversary using projected gradient descent (PGD), which essentially resembles I-FGSM. Recent works have explored frequency domain methods [13] and diffusion models [14] to enhance digital adversarial attack performance. Despite these achievements in the digital space, the evaluation of deep model robustness to physical world attacks remains limited, giving rise to the research branch of physical adversarial attacks.

To assess the generalization of digital perturbations to the physical world, Kurakin *et al.* [30] experimented by directly printing digital adversarial images. Jan *et al.* [31] developed an image-to-image translation network to bridge the appearance gap between printed images and their corresponding digital counterparts. However, these approaches fall short of simulating real-world attacks, as they rely solely on printed versions of digital images. Moreover, these methods often use full-scale perturbations applied to entire images, which is impractical for physical implementation. Subsequent research in physical adversarial attacks has thus shifted towards utilizing adversarial patches [22], a more practical method.

Adversarial Patch. Brown *et al.* [22] introduced the concept of adversarial patches, which differ from full-image perturbations in the digital space [10], [11]. Adversarial patches typically take the form of regular shapes (*e.g.*, squares or circles). Their printability, flexibility, practicality, and strong attack performance have made them a popular choice for physical attacks [16]–[19]. However, transitioning from digital signals to analog signals (patterns printed on physical media) introduces information distortion and loss. To address this, Sharif *et al.* [23] proposed the non-printability score (NPS) to ensure consistent color reproduction between the digital and printed worlds. They also utilized a smooth pattern achieved through total variation (TV) loss to mimic natural images. To enhance the generalization of adversarial patches in the physical world, Athalye *et al.* [25] and Chen *et al.* [24] introduced extensive data augmentations during the digital optimization phase, a technique known as Expectation over Transformation (EOT). These techniques have been widely adopted in subsequent research, including recent efforts aimed at remote sensing object detectors [20], [21], [32].

In this work, we identify an additional challenge in the optimization of adversarial patches, which we term *digital-to-physical visual inconsistency*. By addressing this challenge, we anticipate further improvements in the performance of adversarial patches. We validate our findings and designs in the context of aerial rotated object detection.

B. Image Harmonization.

Image harmonization is a technique within the field of image compositing used to address inconsistencies between a foreground object and its background environment. Traditional methods [33]–[35] relied on low-level statistics like histogram distribution, which were challenging to generalize to complex scenarios. Tsai *et al.* [36] pioneered the use of deep learning methods, offering superior performance. This data-driven approach gained popularity, particularly with the introduction of a large-scale image harmonization dataset by Cong *et al.* [37]. Ke *et al.* [38] manually designed hand-crafted filters and predicted their parameters for image harmonization. Chen *et al.* [26] adopted an implicit neural representation paradigm that processes images at arbitrary resolutions and included an optional 3D LUT prediction branch for efficient harmonization. More recently, Chen *et al.* [39] proposed a diffusion-based harmonization approach that is zero-shot and requires no paired ground truth.

In our work, we pioneer the introduction of image harmonization techniques into adversarial patch optimization to tackle the unexplored digital-to-physical visual inconsistency problem. Similar to [26], we directly predict 3D LUT parameters for efficient and interpretable prediction, and also for preserving the content of the adversarial patch. Furthermore, given the absence of a large-scale harmonization dataset in remote sensing, we propose a self-supervised training scheme.

III. Analysis of Visual Inconsistency

In this section, we delve into an empirical investigation to illuminate the issue of digital-to-physical visual inconsistency during the adversarial patch optimization phase. To

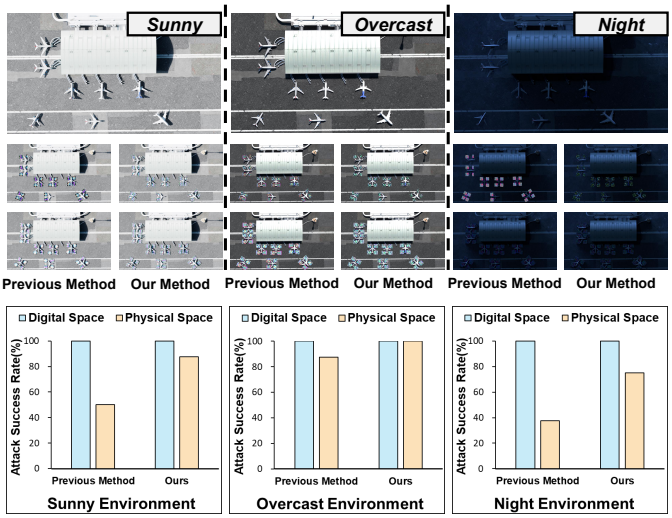

Fig. 2. **Analysis of the Visual Inconsistency Problem.** The first row represents the target scenarios for the adversarial attack, featuring different environmental conditions. The second row displays the digital appearance of the adversarial patches, while the third row showcases their physical appearance which is rendered by a graphics engine (Note: Incomplete patches in the digital world are a result of the splitting strategy explained in Section VI-A). The final row presents a comparison of the attack success rates between the conventional method and our approach in both digital and physical spaces. Please zoom in for a better view.

offer a concise yet comprehensive insight into the problem, we specifically focus on a singular airport scenario from our datasets, applying diverse environmental conditions to it (please refer to Section VI for larger-scale experiments). For each environment, we capture an image of the airport, optimize the adversarial patches on the image, and subsequently place the optimized adversarial patch back into the airport scenario. We then capture another image of this physically perturbed airport and evaluate the physical attack performance. This results in a complete white-box attack. We adhere to the prevalent adversarial patch optimization pipeline, directly backpropagating gradients of the detection loss. The victim detector chosen for this study is ROITransformer [40]. The experimental results are presented in Fig. 2, and we summarize our observations below:

1) Obvious Visual Difference: As illustrated in Fig. 2, a conspicuous disparity emerges between the appearance of the adversarial patch in the digital optimization phase and its manifestation in the real-world scenario. Despite the consistent content of the adversarial patch in both the digital and real spaces, their visual characteristics markedly differ due to the complex environmental factors present in the physical world. This stark contrast highlights a fundamental optimization bias within the digital phase of adversarial patch generation, emphasizing the critical need for a solution to bridge this gap.

2) Deteriorated Attack Performance: The results depicted in the bar chart of Fig. 2 demonstrate the inadequacy of conventional optimization strategies in generating robust physical adversarial patches. While these patches achieve a flawless 100% attack success rate in the digital realm, their performance significantly falters when transferred to the real world. These outcomes underscore the challenges faced by current methods and highlight the urgent necessity for mitigating the visual difference issue. Ensuring that images utilized for optimization closely resemble real-world conditions presents a promising avenue for enhancing attack performance significantly. This revelation substantiates the value of our proposed solutions and underscores the critical importance of addressing the visual inconsistency problem.

In light of these observations, we assert that aligning the digital optimization objective with real-world physical goals holds profound advantages. This alignment not only diminishes the disparity between digital and physical domains but also implicitly resolves the fundamental challenge of directly backpropagating gradients from physical detection results. By enabling the imitation of real-world scenarios within the digital space, our work harnesses these insights to pave the way for more effective and robust physical adversarial attacks.

IV. Methodology

A. Problem Formulation and Pipeline Overview

To formulate our approach, here we consider the digitally optimized adversarial patch as $AdvP$, the physical scenario we intend to attack as S, and the region for placing the adversarial patch as R. We denote the adversarial examples within the digital image space (I_{dig}) and the physical space (I_{phy}) as follows:

$$I_{dig} = \mathrm{Cap}(S) \overset{R}{\oplus} \mathcal{F}(AdvP) \tag{1}$$

$$I_{phy} = \mathrm{Cap}(S \overset{R}{\oplus} \mathrm{Phy}(AdvP)) \tag{2}$$

where $\mathrm{Cap}(\cdot)$ represents the camera image capture, $\mathrm{Phy}(\cdot)$ represents operations for physicalization (*e.g.*, printing out the patch), and $\oplus$ represents either the physical or the digital blending operation. $\mathcal{F}(\cdot)$ pertains to operations applied to $AdvP$, traditionally representing an identical mapping in prior attack methods (*i.e.*, $\mathcal{F}(AdvP) = AdvP$). However, in our approach, it embodies the harmonization method we introduce to ensure digital-to-physical visual consistency for the adversarial patch. Our primary objective, aligning the adversarial patch appearance during digital optimization with its real-world practical appearance, can thus be formulated as:

$$I_{dig} \approx I_{phy} \Rightarrow \mathcal{F}(AdvP) \approx \mathrm{Cap}(\mathrm{Phy}(AdvP)) \tag{3}$$

We present our entire pipeline in Fig. 3. The pipeline comprises two main stages. In the first stage, we train the harmonization network $\mathcal{F}$. This network takes both the composite image $\widetilde{I}$ and the corresponding mask M as input and aims to generate a harmonized result $\overline{I}$. In the second stage, we incorporate the pre-trained harmonization network $\mathcal{F}$ into the adversarial patch optimization phase. Given the clean image $I = \mathrm{Cap}(S)$ of the target physical scenario S, we place the adversarial patch $AdvP$ in a suitable location and obtain the resulting digital adversarial example I_{dig}. Before sending I_{dig} to detectors for gradient backpropagation, we harmonize I_{dig} using the network $\mathcal{F}$ to obtain $\overline{I}_{dig}$, which serves as the input for the detectors. After finishing the optimization, we can place the physicalized adversarial patch in the real-world scenario to achieve an attack.

In the subsequent subsections, we delve into the finer details of our pipeline's components and designs.

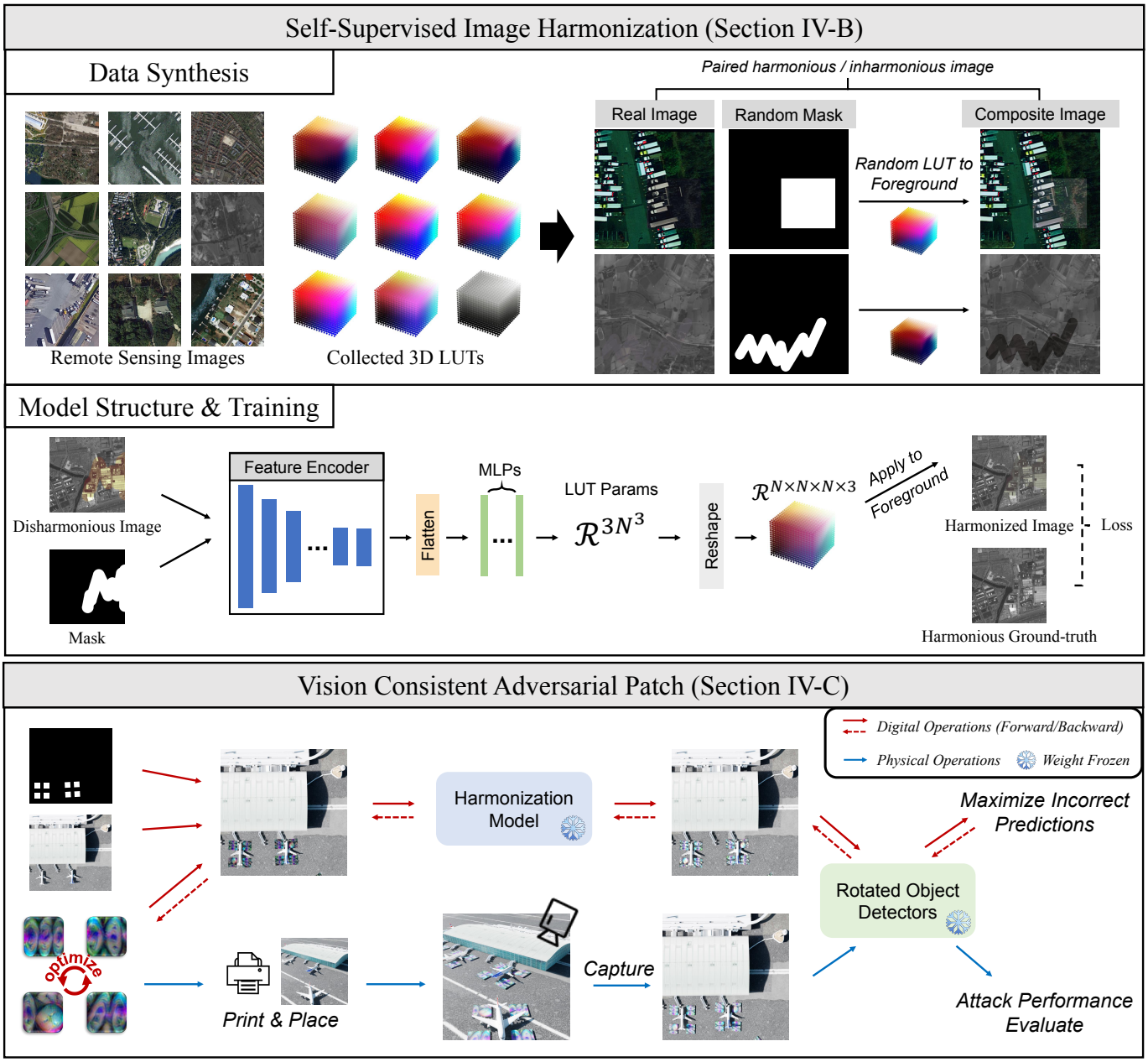

Fig. 3. **The entire pipeline of our method.** Our approach comprises two main components: the self-supervised training of the image harmonization network and the visual consistency optimization of the adversarial patch. In the image harmonization part, we utilize an existing large-scale remote sensing image dataset and collect a set of 3D LUTs. By randomly generating mask regions and applying the LUT to these regions, we create synthetic paired harmonious images (original real image) and composite inharmonious images (after LUT application). Using this synthetic image harmonization dataset, we train the image harmonization network, predicting the parameters of the 3D LUT. In the second step, after completing the training of the harmonization network, we integrate image harmonization into the optimization pipeline of the adversarial attack. This alignment ensures that the optimization goal in the digital space closely mirrors real-world physical scenarios, leading to a significant enhancement in the attack performance of the adversarial patch in the physical world.

B. Self-Supervised Harmonization

1) Harmonization Network Structure. The network structure is depicted in Fig. 3. We draw inspiration from the approach in [26] and predict the parameters of a 3D lookup table (LUT) instead of directly outputting the final harmonized result. This design yields a more lightweight network, facilitating its integration into the adversarial patch optimization phase. Moreover, 3D LUT, being a global color mapping, preserves the content structure within the foreground region, which prevents distorting the patterns in the adversarial patch and thus benefits the attack performance. To predict the 3D LUT, we flatten the encoder features and employ multiple MLPs to predict the final 3D LUT parameters. The parameters are used to generate a 3D LUT $\mathcal{R}^{N\times N\times N\times 3}$, where N represents the number of color value bins, and the number 3 accounts for the resulting pixel values after mapping. For pixel values between two bins, we employ trilinear interpolation for accurate querying [27].

2) Self-Supervised Training Strategy. As there is a lack

of large-scale image harmonization datasets in the remote sensing field, we introduce a self-supervised training strategy to facilitate the training of our harmonization network (see Fig. 3). This strategy involves generating synthetic composite-harmonized image pairs. To simulate a composite image, we must first *define the foreground region* and then *apply color modification*s to differentiate the foreground from the background.

For determining the foreground region, we employ a straightforward method of using random masks. We utilize two types of masks: fixed-form masks and free-form masks. Fixed-form masks consist of predefined shapes (*e.g.*, rectangles, and ellipses), with their parameters (*e.g.*, center position, side length) randomly adjusted to define the foreground region. Additionally, we employ free-form masks, a technique commonly used in image inpainting [41]. Free-form masks offer more diversity and flexibility and resemble iterative drawing brushes, as illustrated in Fig. 3. By using these two random mask generation methods, we simulate the foreground region without relying on semantic information, which facilitates the placement of the adversarial patches in the real world.

For color space modifications, we collect a large number of 3D LUT templates from the Internet. These LUTs simulate various camera imaging mechanisms and color tones present in photography, approximating real composite images with foreground and background under different imaging conditions. This variety ensures the robustness of the trained harmonization network. To construct a comprehensive training dataset, we apply these LUTs to either the foreground region, background region, or both regions.

In summary, our self-supervised training strategy involves randomly selecting mask types, setting mask parameters randomly, choosing 3D LUTs at random, and randomly applying them to the foreground, background, or both.

3) Loss Function. The training of our harmonization network involves optimizing several objectives:

Pixel Regression Loss: This primary loss aims to ensure that the harmonized prediction of the composite image $\widetilde{I}$ closely matches the ground truth I within the foreground region:

$$\mathcal{L}_{pixel} = \mathrm{MSE}(\mathcal{F}(\widetilde{I}) \cdot M, I \cdot M) \tag{4}$$

where MSE denotes mean squared error, $\mathcal{F}$ represents the harmonization network, and M denotes the foreground region mask. This loss function is the main driver for harmonization.

3D LUT Regression Loss: In addition to $\mathcal{L}_{pixel}$, we enforce the predicted 3D LUT parameters to closely match synthetic 3D LUT ground truth, provided we have access to information about the 3D LUT applied to the image:

$$\mathcal{L}_{lut} = \mathrm{MSE}(\widetilde{P}, P_{inv}) \tag{5}$$

where $\widetilde{P}$ represents the predicted 3D LUT parameters, and P_{inv} is the inverted parameters of the originally applied 3D LUT. This loss is applied only when the 3D LUT is applied to either the foreground or the background, as inferring the inverted LUT parameters becomes challenging when both are modified.

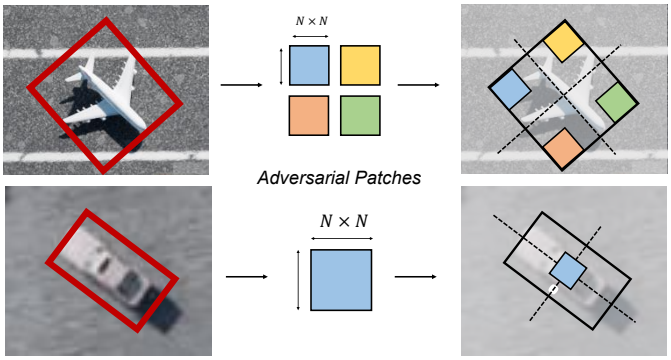

Fig. 4. **Different layouts of the adversarial patches for rotated objects.** Two layouts are considered: corner layout and center layout. The corner layout involves placing the adversarial patch at the four corners of the bounding boxes, while the center layout positions the patch only at the center of the objects. To accommodate practical situations, the center layout is adopted for objects with a relatively plain top surface, whereas the corner layout is employed for objects with non-uniform or curved surfaces.

Boundary Constraint Loss: To prevent the 3D LUT parameters from leaking beyond the RGB color range (normalized into [0, 1]), we introduce a regularization function as follows:

$$\mathcal{L}_{bound} = \mathrm{Avg}(\sum_{p \in \widetilde{P}_{lb}} p^2) + \mathrm{Avg}(\sum_{p \in \widetilde{P}_{hb}} (p-1)^2) \tag{6}$$

where $\widetilde{P}_{lb}$ represents all parameters below the lower bound of the color value, and $\widetilde{P}_{hb}$ represents all parameters above the upper bound. Avg computes the average across all leaked parameters. This loss effectively prevents parameter leakage.

In addition to the above losses, we also include smooth regularization ($\mathcal{L}_{smooth}$) and monotonicity regularization ($\mathcal{L}_{mono}$) to ensure that the learned 3D LUTs are stable and robust. Further details about these two regularization techniques can be found in [27]. The overall loss is a combination of these terms, with weights α, β, γ, and λ assigned to each loss function:

$$\mathcal{L} = \mathcal{L}_{pixel} + \alpha\mathcal{L}_{lut} + \beta\mathcal{L}_{bound} + \gamma\mathcal{L}_{smooth} + \lambda\mathcal{L}_{mono} \tag{7}$$

C. Visual Consistency Adversarial Patch

After completing the training of the harmonization network, we can freeze its weights and seamlessly integrate it into the adversarial patch optimization process. For clarity, we divide the patch optimization into two main parts:

1) Adversarial Example Construction. In this phase, we primarily determine the locations to place adversarial patches ($AdvP$). Previous approaches in remote sensing [20], [21], [32] focused on deceiving horizontal aerial detectors. These methods, leveraging predicted or ground-truth horizontal bounding boxes, applied one adversarial patch per target. These patches were either placed on the top of object targets [20], [21], around the boxes [20], or beneath the targets [32]. In this work, we take a step further by targeting rotated aerial object detectors [42], another popular branch in remote sensing object detection. Additionally, we tuned various layouts for adversarial patches around different types of object targets.

Algorithm 1: The process of the visual consistency optimization for adversarial patch

Input: $Data = \{I, BBox\}_N$, where N is the number of clean images I and their corresponding groundtruth object bounding boxes $BBox$.
Define: $\mathcal{F}(\cdot)$: A pretrained and frozen harmonization network to harmonize the adversarial patch (See Sec. IV-B).
Define: $Det(\cdot)$: A pretrained and frozen aerial detector to provide gradient update direction.
Output: $AdvP$: The optimized adversarial patch.

// Randomly initialize the adversarial patch
$AdvP = Initialization()$
for e *in* $1 : Epochs$ **do**
  for $(I, BBox)$ *in* $Data$ **do**
    // Harmonize the adversarial patches and apply them to the image. (See Sec. IV-A and Fig. 4)
    $I_{dig} = I \overset{R}{\oplus} \mathcal{F}(AdvP)$
    // Calculate detector's loss
    $\mathcal{L}_{det} = Det(I_{dig})$
    // Update the adversarial patch, where η denotes learning rate
    $AdvP \leftarrow AdvP + \eta \cdot \nabla_{AdvP} \mathcal{L}_{det}$
  end
end
Return $AdvP$

We mainly leverage two types of layouts: corner layout and center layout. These layouts are visually illustrated in Fig. 4. Taking the corner layout as an example, we place four adversarial patches at the four corners of the rotated bounding boxes. The side length and aspect ratio of the patches are predefined. In the practical optimization process, we randomly initialize four adversarial patches for each object category. During each iteration, we enumerate all the ground-truth bounding boxes in the input image and place the patches at the corners of the boxes to construct adversarial examples. The center layout follows similar procedures but with only one patch per object.

To account for the diverse shapes of real-world objects, we adapt our layout strategies accordingly. For instance, we opt for the center layout for objects like vehicles and storage tanks, which have relatively smooth top surfaces conducive to placing plain patches. Conversely, for objects like airplanes, with curved top surfaces, we choose the corner layout, situating patches on the four corners of the ground. This strategic variation ensures our patches conform effectively to the specific characteristics of different objects, enhancing their camouflage and attack effectiveness.

After placing the adversarial patches on the objects within the image, we send the resulting adversarial image to the frozen harmonization network to address the digital-to-physical visual inconsistency problem. The harmonized result becomes the final constructed adversarial image that is sent to the following aerial rotated object detector.

2) Adversarial Patch Optimization. With the constructed adversarial image in hand, we feed it to the aerial detector, whose weights are also frozen. The objective of the attack is to deceive the detector into making incorrect predictions. To achieve this, we reverse the sign of the detector's normal training objective. For instance, in the case of the R-CNN series detectors [43], which aims to minimize $\mathcal{L}_{det} = \mathcal{L}_{cls} + \mathcal{L}_{reg}$, where $\mathcal{L}_{cls}$ pertains to object category classification and $\mathcal{L}_{reg}$ deals with bounding box regression, we invert the objective to be $\arg\max_{AdvP} \mathcal{L}_{det}$. Then, we utilize the gradients backpropagated from this inverted objective to update the adversarial patches. To provide a clearer understanding of the process, we present the algorithm outlined in Algorithm 1.

The above pipeline is also readily adaptable to other kinds of detectors. The whole framework enables the generation of digital-to-physical visual consistency adversarial patches against various remote-sensing object detectors, enhancing the vulnerability analysis of the existing detectors.

V. Dataset: 3D Simulated Scenarios

Recent research in the field of physical adversarial attacks for remote sensing [20], [21], [32] has explored various methods for evaluating the physical attack performance of adversarial patches. Zhang *et al.* [21] assessed performance in real-world car parks, while Lian *et al.* [20], [32] employed several airplane models and printed simplified tarmac maps to simulate airport environments. While these approaches provide valuable insights into evaluation methods, they present limitations when addressing the specific challenges in our task. To rigorously evaluate the effectiveness of our method, we require greater control over environmental factors such as illumination and tone. This control is a complex endeavor in both real-world and physically simulated scenarios.

To address this challenge while maintaining data realism, we leverage the graphics rendering engine, Blender, to construct a dataset of six 3D simulated digital scenarios. Leveraging Blender's advanced graphics modeling capabilities, our 3D scenarios achieve a level of environmental realism comparable to real-world scenarios. Additionally, by incorporating various High Dynamic Range Images (HDRIs) within the rendering engine, we can achieve realistic render results and gain control over environmental parameters. This level of control empowers us to conduct a comprehensive evaluation of our solution's effectiveness in mitigating the digital-to-physical visual inconsistency problem.

We present our digitally simulated 3D scenarios in Fig. 5, showcasing a variety of scenes, including terminals with airplanes, roads with vehicles, and factories with tank storage facilities. Constructing these 3D scene models for realistic simulation involves several steps: 1) Initially, we utilize the Google 3D Tiles tool to obtain 3D scans of specific regions selected from Google Maps. However, due to the coarse granularity of the exported 3D scans (*e.g.*, unplanar surfaces), we cannot directly use them as final 3D scene models. Nonetheless, this process provides us with realistic layouts of

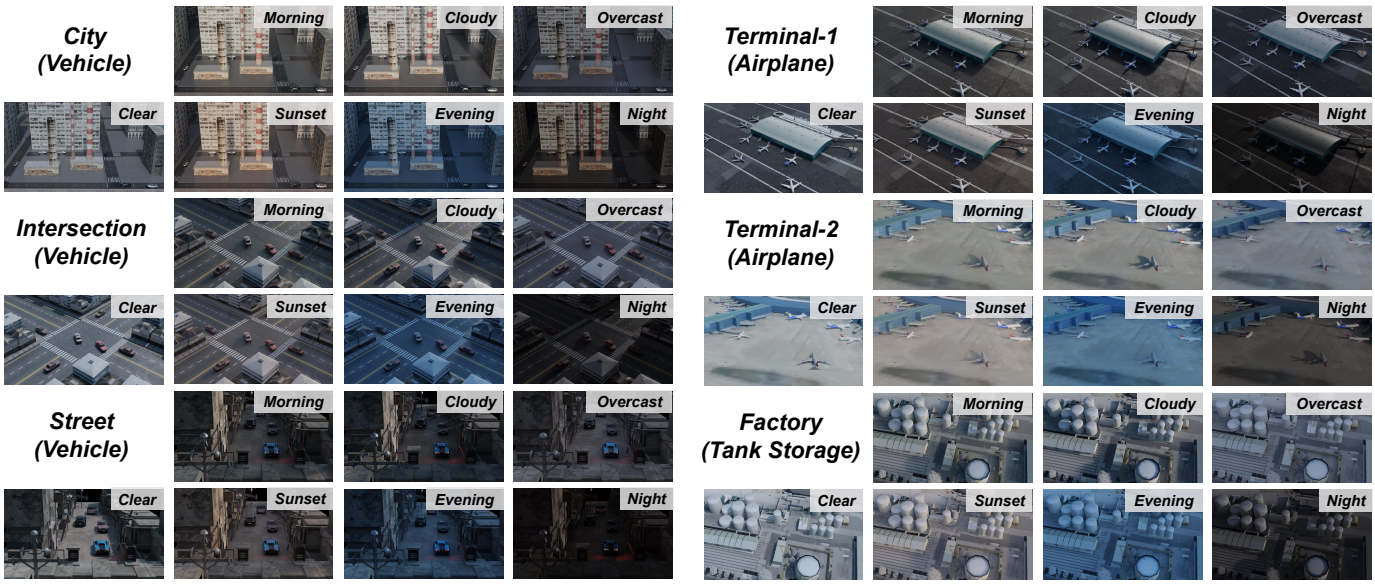

Fig. 5. **Visualization of 3D simulated scenarios.** We present six distinct 3D scenarios generated using a graphics rendering engine, featuring vehicles, airplanes, and tank storage facilities. Environmental factors such as weather, sunshine level, and sun positions can be varied by modifying HDRIs, allowing for diverse image rendering. Here, we showcase one randomly selected rendered image per weather type as an illustrative example. Please zoom in for a better view.

buildings in the scene along with their sizes. 2) Subsequently, we collect numerous finely detailed artist-built 3D elements (*e.g.*, buildings, tank storages) to replace noisy objects in the scans. We apply affine transformations for better alignment with the exported layout and size, resulting in high-quality and realistic 3D scene models. 3) We then gather high-quality digital twins of targeted object categories (*e.g.*, Boeing 747 for planes, BMW X3 for cars) and place them in suitable locations within the 3D scenes (for tank objects, placement is done during phase 2). All objects in the scenarios are ensured to be detectable by rotated object detectors. We utilize the detection results as ground truth bounding boxes and predefine adversarial patch layouts within these boxes.

With the 3D scene model built and the objects placed in it, we add lighting to the scene for rendering. However, simple point lights or area lights may not align with real-world lighting. Hence, for better realism, we've collected 85 HDRIs captured under different environments and categorized into seven distinct weather types (morning, clear, cloudy, overcast, sunset, evening, and night). We leverage these HDRIs as environment maps for our scenarios. By modifying the color strength and using vector mapping operations in Blender, we can vary the sunshine levels and sun positions in the environment, facilitating comprehensive evaluation of our method's performance. With all these operations, we ultimately build realistic scenarios to serve as an ideal platform for evaluating physical adversarial attacks.

VI. Experiments

A. Experimental Setup

Dataset Setup. In our image harmonization training phase, we rely on the extensive DOTA dataset [42], comprising 2,806 aerial images with resolutions ranging from 400 to 13,000 pixels. To synthesize diverse imaging environments, we curate a repository of 2,720 3D Look-Up Tables (LUTs) gathered from the Internet. Utilizing these LUTs, we create a large-scale dataset by randomly selecting an LUT and applying it to a random image from the DOTA dataset. This approach ensures the training of our image harmonization network on a wide variety of visual contexts.

For our adversarial patch attack experiments, we employ our meticulously crafted 3D simulated scenarios. To mimic real-world camera poses within remote sensing imagery, we set the camera pose perpendicular to the ground plane, capturing a 1280×720 top view of the scenario. By leveraging different HDRIs and randomly adjusting parameters such as sunshine levels and sun positions, we create 150 distinct environmental variations for each scenario. This meticulous approach results in a total of 900 (150×6) different scenes, each representing a unique combination of lighting, shadows, and atmospheric conditions. This deliberate variation provides a rich and diverse set of conditions for evaluating the robustness and effectiveness of our method under different environmental contexts.

Implementation Details. For training the image harmonization network, we employ the encoder part of HINet [26] as the encoder, followed by 2 MLPs to predict the LUT's parameters. The input resolution is set to 512×512. We use both fixed-form and free-form masks, each with a 50% probability of being chosen. For applying 3D LUTs, we set a 0.8 probability of applying LUTs to the foreground and a 0.2 probability of applying LUTs to both the foreground and background, to prevent excessive background changes. The color bin N for the 3D LUT is set to 7. We use the AdamW [44] optimizer with a learning rate of $1e^{-5}$, 1K iterations, and a batch size of 8. The weight factors α, β, γ, and λ in Eq. 7 are set to 1, 0.1, 0.1, and 0.1 respectively.

For the optimization of the adversarial patch, the initial

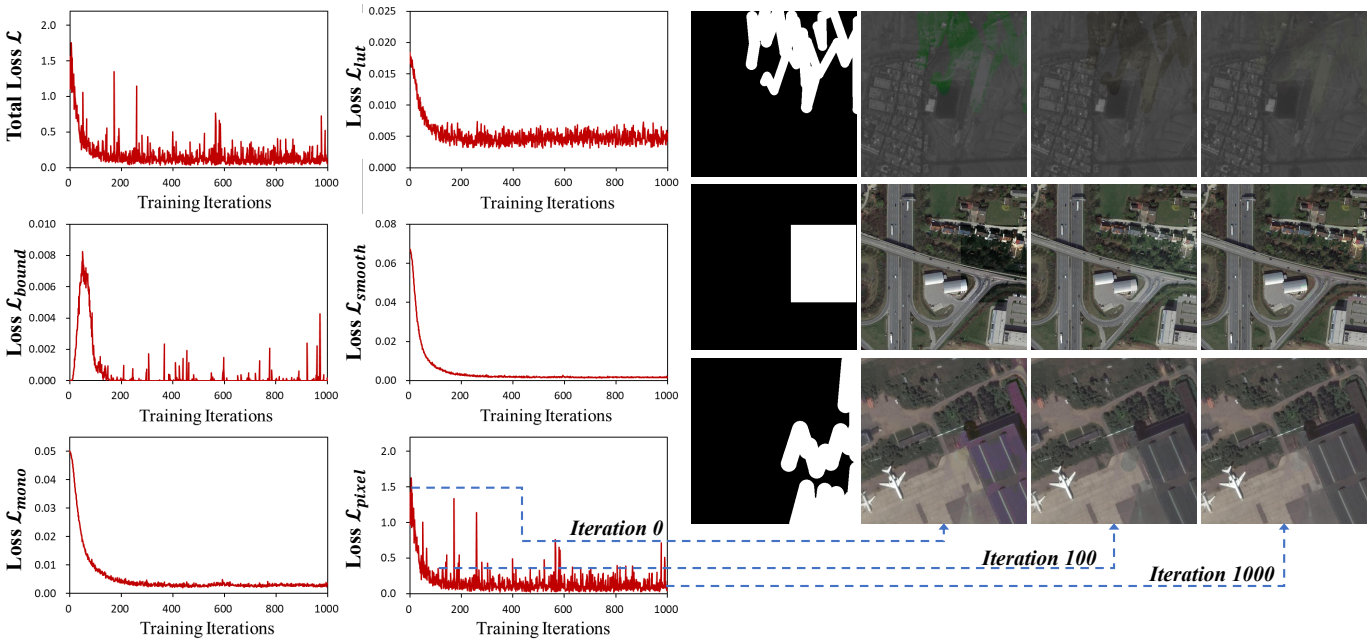

Fig. 6. **Visualization of image harmonization performance and training progress of loss functions.** The graph illustrates the variation of each loss function in Eq. 7 throughout the training process. We also randomly generated some composite images and tested their harmonized results at different iterations. Please zoom in for a better view.

large input images are split into multiple 512×512 patches. To clearly validate the visual difference problem and the effectiveness of our solution, we conduct purely white-box attacks, that is the scenario utilized for optimization mirrors the one employed for testing, ensuring a direct comparison between the two states. We perform 10 optimization epochs using the AdamW optimizer with a learning rate of 0.03. In practical implementations, we further incorporate a safeguarding mechanism to account for the possibility of incorrect color mapping results arising from the image harmonization network. To mitigate such situations, we introduce a probability of 0.5 to determine whether to employ image harmonization, enhancing the overall stability of the results. All experiments are conducted on a single RTX 3090 GPU.

Victim Detectors & Evaluation Metrics. We take the critical aerial detection task in remote sensing as an example evaluation target, and evaluate our method against various state-of-the-art aerial rotated object detectors, including ROITransformer [40], Rotated RetinaNet [45], Oriented R-CNN [46], and Oriented RepPoints [47].

We leverage MSE to evaluate the performance of the harmonization network. To assess attack performance, we report the average precision (AP) and the average recall (AR) of the detectors before and after the attack. We also leverage LPIPS [48] to assess the visual consistency of the adversarial patch appearance between the digital space and physical space.

B. Image Harmonization Performance

In our approach, unlike other tasks with existing datasets for supervision, the training of our harmonization network relies on purely synthetic images. In each training iteration, the input composite image is generated by randomly sampling a remote-sensing image and one or two LUTs. Given the vast number of remote sensing images and the collected 3D LUTs, samples in each iteration are entirely new. Therefore, we can directly utilize the $\mathcal{L}_{pixel}$ loss within the training phase, which measures the disparity between the harmonized image and the ground truth, as a metric of harmonization performance.

The evolution of our harmonization network's performance, along with the training progress of the loss functions in Eq. 7, is illustrated in Fig. 6. As training iterations progress, the $\mathcal{L}_{pixel}$ loss gradually converges and remains generally stable. To visually assess the harmonization performance, we generated several composite images and fed them into the harmonization network at different iteration stages. From the visualizations, it is evident that the composite images are successfully harmonized, aligning the color spaces of the foreground and background effectively.

C. Results on Different Physical Environments

1) Quantitative Results: To elucidate the impact of the digital-to-physical visual inconsistency problem, we conducted comprehensive experiments, as outlined in Table I and Table II. All experiments were set up in a purely white-box manner, where the scenes used for digitally optimizing adversarial patches were identical to those where the adversarial patches were physically placed. We compared our visual consistency method with the commonly used normal approach (the one elucidated in Fig. 1). Additionally, we considered the commonly used EOT performance enhancement strategy to assess our method's compatibility. We applied transformations such as RandomBrightness and RandomContrast as part of EOT.

To comprehensively assess the effectiveness of our method, we conducted comparisons using different detectors and

TABLE I

RESULTS ON DIFFERENT PHYSICAL ENVIRONMENTS. THE ATTACK IS SET UP IN A PURELY WHITE-BOX MANNER (*i.e.*, THE TARGET SCENE IS THE SAME AS THE SCENE USED FOR OPTIMIZATION). "NORMAL" DENOTES THE COMMON METHOD, WHILE "CONSISTENT" DENOTES OUR VISUAL CONSISTENCY OPTIMIZATION METHOD. THE REPORTED VALUES REPRESENT THE AVERAGE PRECISION (AP) OF THE TARGET DETECTOR. THE BEST VALUE IS SET TO BOLD AND HIGHLIGHTED IN RED FOR BETTER VISIBILITY. PLEASE ZOOM IN FOR A CLOSER VIEW.

ROITransformer	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.909	0.909	0.909	0.909	1	1	0.977	0.977	0.815	0.815	0.994	0.994	0.909	0.909	0.904	0.904	0.909	0.909
Normal	0	0.631	0.091	0.364	0	0.545	0.014	0.293	0.024	0.414	0.011	0.334	0.091	0.156	0.091	0.359	0.049	0.415
Consistent	0.182	**0.630**	0	**0.273**	0	**0.364**	0.373	**0.288**	0.253	**0.407**	0.268	**0.251**	0.154	**0.154**	0.154	**0.354**	0.208	**0.287**
Normal+EOT	0.030	**0.631**	0.091	0.364	0	0.455	0.091	0.361	0.091	0.499	0.053	0.326	0.091	0.161	0.091	0.358	0.057	0.401
Consistent+EOT	0.181	**0.631**	0.091	**0.273**	0.091	**0.364**	0.368	**0.289**	0.310	**0.406**	0.280	**0.244**	0.156	**0.158**	0.158	**0.356**	0.215	**0.297**
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.908	0.908	0.721	0.721	0.909	0.909	0.909	0.909	0.727	0.727	0.908	0.908	0.993	0.993	0.636	0.636	0.994	0.994
Normal	0.159	0.786	0.091	0.522	0.253	0.704	0	**0.137**	0	**0.182**	0	0.234	0.002	0.536	0.014	**0.364**	0.003	**0.545**
Consistent	0.246	**0.615**	0.156	**0.521**	0.264	**0.701**	0.015	0.138	0	**0.182**	0	**0.182**	0	**0.518**	0.091	**0.364**	0.036	**0.545**
Normal+EOT	0.159	0.700	0.156	**0.425**	0.258	0.703	0	**0.149**	0.091	0.182	0	0.256	0.003	0.454	0.007	**0.363**	0.003	**0.545**
Consistent+EOT	0.322	**0.698**	0.156	**0.425**	0.259	**0.691**	0.016	0.155	0	**0.179**	0	**0.246**	0.018	**0.452**	0.091	0.440	0.091	0.628

Oriented R-CNN	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.909	0.909	0.818	0.818	1	1	0.976	0.976	0.818	0.818	0.982	0.982	0.909	0.909	0.905	0.905	0.909	0.909
Normal	0.250	0.894	0	0.545	0.091	**0.364**	0.118	0.383	0.091	0.540	0.117	0.446	0.148	0.248	0.091	0.443	0.083	**0.413**
Consistent	0.420	**0.878**	0.182	**0.542**	0.091	**0.364**	0.521	**0.320**	0.350	**0.447**	0.409	**0.355**	0.235	**0.227**	0.242	**0.441**	0.279	0.474
Normal+EOT	0.237	0.896	0	0.545	0	**0.364**	0.119	**0.227**	0.091	0.534	0.091	**0.447**	0.140	0.317	0.144	0.527	0.106	**0.405**
Consistent+EOT	0.430	**0.876**	0.091	**0.364**	0.091	**0.364**	0.523	0.237	0.347	**0.448**	0.422	**0.447**	0.233	**0.158**	0.244	**0.441**	0.348	0.481
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.909	0.909	0.725	0.725	0.909	0.909	0.909	0.909	0.632	0.632	0.906	0.906	0.991	0.991	0.636	0.636	0.987	0.987
Normal	0.174	0.618	0.091	0.351	0.182	0.542	0.067	0.172	0	0.175	0	0.264	0.005	**0.411**	0.019	0.329	0.091	0.533
Consistent	0.264	**0.612**	0.091	**0.264**	0.182	**0.541**	0.091	**0.169**	0.013	**0.174**	0	**0.160**	0.012	0.444	0.063	**0.328**	0.004	**0.448**
Normal+EOT	0.171	0.629	0.091	0.354	0.182	0.644	0.091	**0.166**	0.036	0.171	0	0.182	0.028	**0.412**	0.019	0.324	0.125	0.711
Consistent+EOT	0.266	**0.529**	0.091	**0.267**	0.182	**0.364**	0.091	**0.166**	0.091	**0.091**	0	**0.155**	0.004	0.436	0.022	**0.319**	0.182	**0.632**

Oriented RepPoints	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.909	0.909	0.885	0.885	0.909	0.909	0.976	0.976	0.834	0.834	0.971	0.971	0.886	0.886	0.860	0.860	0.887	0.887
Normal	0.004	**0.715**	0.008	0.807	0.002	0.892	0.176	0.240	0.009	0.464	0.009	0.404	0.013	0.073	0.012	0.271	0.014	0.166
Consistent	0.112	0.759	0.002	**0.712**	0.013	**0.788**	0.360	**0.152**	0.112	**0.304**	0.084	**0.148**	0.056	**0.060**	0.044	**0.260**	0.061	**0.150**
Normal+EOT	0..003	**0.634**	0.002	0.800	0.002	0.884	0.152	0.196	0.010	0.440	0.008	0.209	0.012	0.067	0.012	0.271	0.014	0.166
Consistent+EOT	0.094	0.658	0.005	**0.714**	0.023	**0.786**	0.385	**0.176**	0.171	**0.337**	0.091	**0.127**	0.052	**0.065**	0.047	**0.258**	0.059	**0.141**
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.908	0.908	0.799	0.799	0.903	0.903	0.908	0.908	0.725	0.725	0.902	0.902	0.909	0.909	0.683	0.683	0.996	0.996
Normal	0.440	0.787	0.201	0.588	0.285	0.661	0.265	**0.552**	0.100	0.416	0.091	0.492	0.176	0.577	0.102	**0.418**	0.331	0.649
Consistent	0.620	**0.768**	0.291	**0.479**	0.423	**0.652**	0.426	0.588	0.100	**0.310**	0.091	**0.398**	0.274	**0.575**	0.560	0.681	0.378	**0.648**
Normal+EOT	0.431	0.786	0.201	0.488	0.279	0.661	0.275	**0.548**	0.098	0.419	0.100	0.584	0.101	0.612	0.425	0.664	0.261	**0.646**
Consistent+EOT	0.618	**0.771**	0.306	**0.481**	0.427	**0.658**	0.448	0.575	0.098	**0.252**	0.091	**0.425**	0.251	**0.589**	0.107	**0.477**	0.108	0.689

Rotated RetinaNet	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.909	0.909	0.818	0.818	1	1	0.976	0.976	0.808	0.808	0.984	0.984	0.909	0.909	0.903	0.903	0.091	0.091
Normal	0.019	0.173	0.013	0.292	0.050	0.377	0.154	0.138	0.110	**0.238**	0.091	0.526	0.399	0.603	0.320	0.698	0.515	**0.782**
Consistent	0.020	**0.133**	0.045	**0.239**	0.091	**0.355**	0.370	**0.135**	0.342	0.401	0.319	**0.498**	0.589	**0.593**	0.512	**0.696**	0.615	0.869
Normal+EOT	0.024	0.162	0.015	0.291	0.024	0.462	0.155	0.110	0.113	0.249	0.091	0.519	0.400	0.600	0.328	**0.698**	0.413	**0.607**
Consistent+EOT	0.028	**0.160**	0.091	**0.172**	0.091	**0.264**	0.364	**0.081**	0.307	**0.223**	0.327	**0.473**	0.592	**0.588**	0.512	**0.698**	0.615	0.865
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.909	0.909	0.808	0.808	0.909	0.909	0.917	0.917	0.597	0.597	0.970	0.970	1	1	0.727	0.727	1	1
Normal	0.809	0.893	0.599	0.776	0.750	**0.897**	0.225	0.718	0.103	0.320	0.122	0.588	0	0.432	0	0.290	0	**0.364**
Consistent	0.850	**0.868**	0.643	**0.773**	0.767	0.901	0.299	**0.697**	0.151	**0.296**	0.146	**0.508**	0.030	**0.425**	0	**0.269**	0	0.455
Normal+EOT	0.807	0.898	0.590	0.779	0.758	**0.899**	0.267	0.674	0.098	0.305	0.150	0.590	0	**0.351**	0	0.335	0	**0.364**
Consistent+EOT	0.851	**0.863**	0.645	**0.762**	0.773	**0.899**	0.381	**0.621**	0.190	**0.264**	0.181	**0.471**	0.017	0.411	0	**0.299**	0	**0.364**

TABLE II

RESULTS ON DIFFERENT PHYSICAL ENVIRONMENTS. THE ATTACK IS SET UP IN A PURELY WHITE-BOX MANNER (*i.e.*, THE TARGET SCENE IS THE SAME AS THE SCENE USED FOR OPTIMIZATION). "NORMAL" DENOTES THE COMMON METHOD, WHILE "CONSISTENT" DENOTES OUR VISUAL CONSISTENCY OPTIMIZATION METHOD. THE REPORTED VALUES REPRESENT THE AVERAGE RECALL (AR) OF THE TARGET DETECTOR. THE BEST VALUE IS SET TO BOLD AND HIGHLIGHTED IN RED FOR BETTER VISIBILITY. PLEASE ZOOM IN FOR A CLOSER VIEW.

ROITransformer	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.998	0.998	0.900	0.900	1	1	1	1	0.861	0.861	1	1	0.992	0.992	0.913	0.913	0.995	0.995
Normal	0.224	0.699	0.041	0.387	0.057	0.501	0.127	**0.309**	0.100	0.455	0.045	0.368	0.066	0.129	0.066	0.361	0.063	0.444
Consistent	0.111	**0.664**	0.040	**0.283**	0.061	**0.360**	0.505	0.375	0.302	**0.414**	0.380	**0.299**	0.227	**0.128**	0.140	**0.344**	0.272	**0.315**
Normal+EOT	0.098	0.661	0.037	0.369	0.054	0.457	0.173	0.400	0.125	0.509	0.045	0.325	0.068	0.163	0.063	0.327	0.063	0.409
Consistent+EOT	0.235	**0.640**	0.051	**0.240**	0.069	**0.333**	0.520	**0.368**	0.368	**0.420**	0.419	**0.249**	0.215	**0.129**	0.167	**0.312**	0.294	**0.314**
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.990	0.990	0.788	0.788	0.992	0.992	0.997	0.997	0.708	0.708	0.961	0.961	1	1	0.7	0.7	1	1
Normal	0.176	0.814	0.090	0.536	0.242	0.709	0	**0.102**	0	0.152	0	0.259	0.013	0.538	0.013	0.394	0.010	0.572
Consistent	0.302	**0.694**	0.106	**0.512**	0.242	**0.705**	0.007	0.109	0	**0.135**	0	**0.208**	0.003	**0.516**	0.016	**0.387**	0.030	**0.539**
Normal+EOT	0.186	0.774	0.114	0.484	0.274	0.766	0	**0.108**	0.004	**0.127**	0	0.202	0.022	**0.497**	0.013	**0.369**	0.020	**0.526**
Consistent+EOT	0.330	**0.734**	0.104	**0.462**	0.240	**0.762**	0.003	0.158	0	**0.127**	0	**0.195**	0.028	**0.497**	0.019	0.434	0.049	0.622

Oriented R-CNN	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.997	0.997	0.894	0.894	1	1	1	1	0.855	0.855	1	1	0.994	0.994	0.924	0.924	0.994	0.994
Normal	0.296	0.937	0.074	0.516	0.102	0.395	0.232	0.405	0.152	0.545	0.182	0.471	0.197	0.257	0.103	0.471	0.160	**0.497**
Consistent	0.504	**0.906**	0.152	**0.514**	0.112	**0.370**	0.643	**0.348**	0.436	**0.470**	0.562	**0.368**	0.294	**0.205**	0.296	**0.463**	0.398	0.505
Normal+EOT	0.305	0.937	0.061	0.578	0.099	**0.304**	0.218	**0.220**	0.166	0.530	0.172	0.464	0.159	0.319	0.152	0.509	0.171	**0.484**
Consistent+EOT	0.516	**0.913**	0.086	**0.359**	0.121	0.380	0.636	0.273	0.459	**0.498**	0.569	**0.452**	0.292	**0.190**	0.295	**0.468**	0.437	0.532
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.984	0.984	0.718	0.718	0.979	0.979	0.973	0.973	0.619	0.619	0.956	0.956	1	1	0.7	0.7	1	1
Normal	0.198	0.630	0.046	0.312	0.158	0.552	0.012	0.168	0	0.150	0	0.254	0.019	**0.409**	0.047	0.394	0.039	0.526
Consistent	0.266	**0.602**	0.062	**0.244**	0.145	**0.535**	0.028	**0.149**	0.004	**0.144**	0	**0.125**	0.028	0.478	0.103	**0.347**	0	**0.434**
Normal+EOT	0.156	0.676	0.048	0.346	0.120	0.556	0.023	0.127	0.008	**0.098**	0	0.171	0.031	**0.409**	0.047	0.378	0.102	0.711
Consistent+EOT	0.288	**0.560**	0.070	**0.278**	0.152	**0.396**	0.030	**0.123**	0.008	0.104	0	**0.118**	0.019	0.587	0.059	**0.347**	0.191	**0.658**

Oriented RepPoints	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.999	0.999	0.928	0.928	0.996	0.996	1	1	0.902	0.902	1	1	0.987	0.987	0.964	0.964	0.987	0.987
Normal	0.134	**0.827**	0.098	0.872	0.085	0.977	0.057	0.348	0.057	0.602	0.060	0.533	0.106	0.293	0.100	0.537	0.130	0.510
Consistent	0.359	0.883	0.077	**0.795**	0.176	**0.895**	0.493	**0.266**	0.402	**0.382**	0.443	**0.244**	0.283	**0.236**	0.213	**0.518**	0.336	**0.473**
Normal+EOT	0.146	**0.759**	0.063	0.862	0.103	0.968	0.061	0.341	0.121	0.566	0.060	0.366	0.109	0.272	0.106	0.536	0.132	0.508
Consistent+EOT	0.343	0.787	0.104	**0.777**	0.196	**0.888**	0.464	**0.302**	0.382	**0.402**	0.421	**0.187**	0.278	**0.255**	0.216	**0.514**	0.342	**0.459**
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.992	0.992	0.818	0.818	0.992	0.992	0.998	0.998	0.762	0.762	0.980	0.980	0.994	0.994	0.719	0.719	1	1
Normal	0.554	0.898	0.240	0.696	0.381	0.783	0.489	0.783	0.144	0.575	0.039	0.618	0.397	0.753	0.209	**0.572**	0.447	**0.789**
Consistent	0.700	**0.868**	0.332	**0.526**	0.516	**0.747**	0.637	**0.780**	0.127	**0.463**	0.079	**0.522**	0.397	**0.731**	0.694	0.750	0.414	**0.789**
Normal+EOT	0.518	0.880	0.268	0.570	0.347	0.779	0.498	0.779	0.119	0.585	0.132	0.719	0.338	0.775	0.600	0.738	0.309	**0.776**
Consistent+EOT	0.696	**0.874**	0.358	**0.564**	0.507	**0.766**	0.647	**0.773**	0.112	**0.369**	0.112	**0.511**	0.303	**0.747**	0.222	**0.613**	0.276	0.851

Rotated RetinaNet	**Terminal-1**						**Terminal-2**						**Factory**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.992	0.992	0.891	0.891	1	1	1	1	0.836	0.836	1	1	0.987	0.987	0.978	0.978	0.987	0.987
Normal	0.129	0.294	0.072	0.387	0.152	0.427	0.332	0.302	0.230	**0.345**	0.208	0.591	0.415	0.683	0.349	0.737	0.540	**0.838**
Consistent	0.186	**0.239**	0.085	**0.342**	0.171	**0.414**	0.518	**0.273**	0.405	0.466	0.505	**0.533**	0.641	**0.629**	0.598	**0.734**	0.696	0.906
Normal+EOT	0.133	0.280	0.077	0.386	0.134	0.519	0.334	0.270	0.264	0.361	0.215	0.562	0.409	0.642	0.384	0.742	0.416	**0.694**
Consistent+EOT	0.179	**0.258**	0.138	**0.282**	0.179	**0.349**	0.516	**0.227**	0.386	**0.325**	0.522	**0.526**	0.648	**0.621**	0.609	**0.741**	0.684	0.903
	City						**Intersection**						**Street**					
Environ / Attack	Bright		Dark		Color		Bright		Dark		Color		Bright		Dark		Color	
	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical	Digital	Physical
Clean	0.998	0.998	0.840	0.840	0.996	0.996	1	1	0.731	0.731	1	1	1	1	0.722	0.722	1	1
Normal	0.908	**0.998**	0.724	0.826	0.865	0.981	0.336	0.843	0.262	0.475	0.235	0.741	0	0.494	0	0.328	0	**0.368**
Consistent	0.940	1	0.736	**0.822**	0.842	**0.973**	0.359	**0.818**	0.302	**0.467**	0.151	**0.614**	0.006	**0.472**	0	**0.303**	0	0.408
Normal+EOT	0.914	**0.998**	0.720	0.822	0.865	0.979	0.344	0.798	0.223	0.467	0.274	0.728	0	**0.391**	0	0.306	0	**0.368**
Consistent+EOT	0.944	1	0.746	**0.820**	0.886	**0.973**	0.445	**0.751**	0.294	**0.433**	0.300	**0.561**	0.016	0.456	0	**0.281**	0	**0.368**

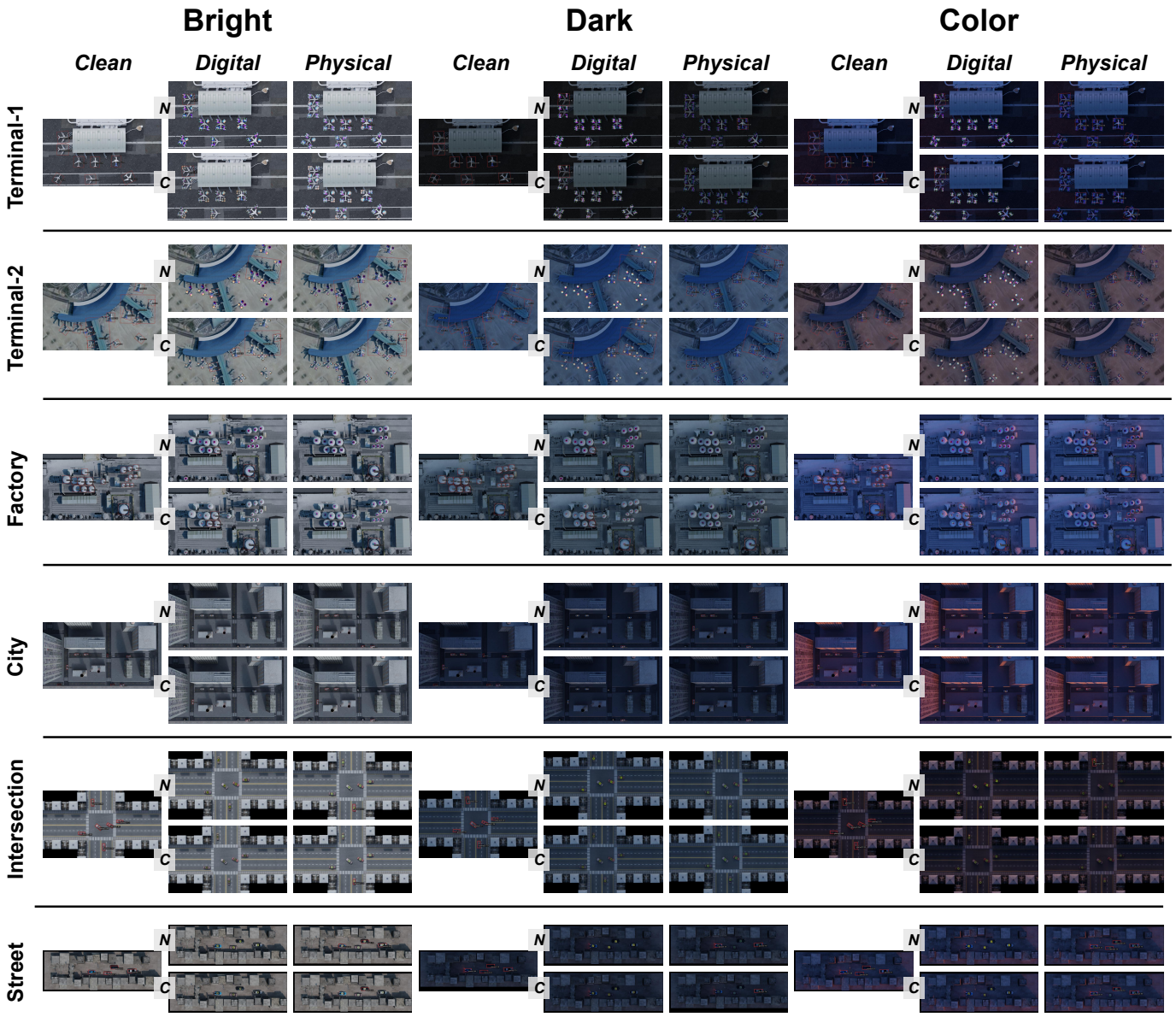

Fig. 7. **Visualization of attack performance of the normal method and our consistent method in different environments.** For clarity, we display the results based on ROITransformer. "Bright", "Dark", and "Color" denote different subparts of divided environmental variations of each scene. "Digital" and "Physical" indicate whether the adversarial example is digitally crafted (*i.e.,* directly pasting the adversarial patch onto the digital image) or projected to the physical world. "N" and "C" are abbreviations of 'Normal" and "Consistent", respectively, denoting the attack method, *i.e.,* whether our method is applied. In the adversarial examples, red bounding boxes indicate whether the detector can detect the objects. Please zoom in for a better view.

scenes. For each scene, we evenly divided its 150 environmental variations into three subparts: Bright, Dark, and Color, with Bright and Dark indicating varying illuminations and Color representing environments with distinct tones (*e.g.*, the yellow or pink hues at sunset). The visualizations of these subparts can be referred to in Fig. 7.

Table I and Table II present the Average Precision (AP) and the Average Recall (AR) of detectors for adversarial examples optimized in the digital space and projected into the physical space. From the AP and AR results, several observations can be made:

- A noticeable gap in attack performance exists between digital adversarial examples and their physical counterparts. In most scenarios, there is a clear increase in AP and AR when adversarial patches are projected into the physical world. This validates the presence of optimization bias in the digital optimization phase of adversarial patches, meaning the digital optimization objective is inconsistent with the ultimate physical attack goal.
- Our visual consistency optimization method helps mitigate this gap. Our method improves the physical attack performance, resulting in lower detection AP and AR values compared to the normal approach. It's worth noting that while our method causes a deterioration in digital attack performance (attributed to the varying appearance of the adversarial patch in each image due to our method considering the background image's environment, adding challenges to patch optimization convergence),

TABLE III
RESULTS ON VISUAL CONSISTENCY. THE REPORTED VALUES REPRESENT THE LPIPS BETWEEN DIGITAL ADVERSARIAL EXAMPLES AND THEIR PHYSICAL COUNTERPARTS. THE BEST VALUE IS SET TO BOLD.

Attack \ Environ	City			Intersection			Street		
	Bright	Dark	Color	Bright	Dark	Color	Bright	Dark	Color
Normal	0.0067	0.0229	0.0078	0.0039	0.0178	0.0086	0.0061	0.0146	0.0109
Consistent	**0.0036**	**0.0179**	**0.0058**	**0.0038**	**0.0130**	**0.0035**	**0.0054**	**0.0131**	**0.0077**
Attack \ Environ	**Terminal-1**			**Terminal-2**			**Factory**		
	Bright	Dark	Color	Bright	Dark	Color	Bright	Dark	Color
Normal	0.1007	0.1299	0.1001	0.0891	0.1137	0.0881	0.0369	0.0704	0.0436
Consistent	**0.0973**	**0.1155**	**0.0873**	**0.0713**	**0.1027**	**0.0708**	**0.0315**	**0.0549**	**0.0289**

TABLE IV
RESULTS ON UNIVERSAL ADVERSARIAL PATCHES. THE ADVERSARIAL PATCHES APPLIED TO TERMINAL-1 AND TERMINAL-2 ARE OPTIMIZED ON EITHER THE DOTA DATASET OR THE RSOD DATASET. THE REPORTED VALUES REPRESENT THE AVERAGE PRECISION (AP) OF THE DETECTOR. THE BEST VALUE IS SET TO BOLD.

Detector \ Environ	DOTA				RSOD			
	Terminal-1		Terminal-2		Terminal-1		Terminal-2	
	Normal	Consistent	Normal	Consistent	Normal	Consistent	Normal	Consistent
ROITransformer	0.818	**0.727**	0.320	**0.286**	0.818	**0.727**	0.234	**0.157**
Oriented R-CNN	**0.727**	**0.727**	0.425	**0.409**	0.818	**0.727**	0.532	**0.429**
Oriented RepPoints	0.896	**0.793**	0.262	**0.173**	0.871	**0.786**	0.316	**0.137**
Rotated RetinaNet	0.709	**0.352**	0.215	**0.182**	0.626	**0.567**	0.222	**0.155**

TABLE V
RESULTS ON BLACK-BOX TRANSFERABILITY. "SURROGATE" DENOTES THE DETECTOR USED FOR OPTIMIZING THE ADVERSARIAL PATCHES, WHILE "TARGET" DENOTES THE DETECTOR FOR EVALUATING THE ATTACK PERFORMANCE OF THE ADVERSARIAL PATCHES. THE REPORTED VALUES REPRESENT THE AVERAGE PRECISION (AP) OF THE TARGET DETECTOR. VALUES WITH A GRAY BACKGROUND INDICATE WHITE-BOX ATTACKS (*i.e.,* THE SURROGATE DETECTOR IS THE SAME AS THE TARGET ONE). THE BEST VALUE IS SET TO BOLD.

Environ	Surrogate \ Target	ROITransformer		Oriented R-CNN		Oriented RepPoints		Rotated RetinaNet	
		Normal	Consistent	Normal	Consistent	Normal	Consistent	Normal	Consistent
Bright	ROITransformer	0.631	**0.630**	0.908	**0.818**	0.908	**0.907**	0.908	**0.907**
	Oriented R-CNN	0.772	**0.720**	0.894	**0.878**	0.904	**0.903**	0.757	**0.735**
	Oriented RepPoints	0.758	**0.731**	0.814	**0.703**	**0.715**	0.759	0.843	**0.809**
	Rotated RetinaNet	0.482	**0.411**	0.616	**0.503**	0.574	**0.521**	0.173	**0.133**
Dark	ROITransformer	0.364	**0.273**	**0.636**	**0.636**	0.818	**0.727**	0.636	**0.633**
	Oriented R-CNN	0.273	**0.182**	0.545	**0.542**	**0.818**	**0.818**	0.798	**0.712**
	Oriented RepPoints	0.700	**0.690**	0.779	**0.706**	0.807	**0.712**	0.811	**0.802**
	Rotated RetinaNet	0.260	**0.258**	0.539	**0.409**	0.726	**0.692**	0.292	**0.239**
Color	ROITransformer	0.545	**0.364**	**0.636**	**0.636**	**0.818**	**0.818**	0.727	**0.726**
	Oriented R-CNN	0.545	**0.358**	**0.364**	**0.364**	**0.909**	**0.909**	0.907	**0.889**
	Oriented RepPoints	0.883	**0.865**	0.855	**0.852**	0.892	**0.788**	0.907	**0.889**
	Rotated RetinaNet	0.441	**0.325**	0.438	**0.345**	0.818	**0.816**	0.377	**0.355**

our method still outperforms the normal approach in physical attacks. This further demonstrates the effectiveness of our method and suggests that digital attack performance doesn't strongly correlate with physical attack

- The commonly used EOT strategy doesn't consistently improve results. Although EOT can enhance physical attack performance in certain situations, it can also degrade performance in others. This phenomenon may be attributed to the purely white-box attack setup, where the image transformations introduced by EOT are not consistent with the target scene, hindering the optimization of adversarial attacks. Nevertheless, our method consistently achieves better performance in most situations, whether or not EOT is applied. Furthermore, our method, even without EOT, often outperforms the normal approach with EOT, further underscoring our method's effectiveness.

2) Qualititative Results: In Fig. 7, we presented visualizations of the adversarial patches, digital adversarial examples, and their physical counterparts for both the normal and our consistent methods. These visualizations clearly demonstrated our method's superior visual consistency in the digital optimization phase, highlighting the substantial improvement over the normal approach.

To delve deeper, we calculated the LPIPS value between

the digital adversarial examples and their physical counterparts to evaluate the level of visual consistency. Using the adversarial patches generated on ROITransformer as an example, the results are displayed in Table III. Our method achieves lower LPIPS values, indicating greater consistency between digital and physical adversarial examples and further verifying the effectiveness of our digital-to-physical visual consistency method.

D. Extended Explorations and Analysis

This subsection delves deeper into the application of our method in crafting universal adversarial patches and its implications for black-box transferability, further uncovering the potential of our approach.

1) Results on Universal Adversarial Patches: In the above experiments, the scenes used to optimize the adversarial patches are the same as the ones where the optimized patches will be physically applied. However, in real-world scenarios, it may be challenging to access or collect images of the target scene, and optimizing adversarial patches from scratch for every new scene might be inconvenient. To address this, we assess the effectiveness of our method on universal adversarial patches—adversarial patches optimized on a large collected dataset and then applied to any new target scene.

Taking the example of Airplane objects, we assess our method's effectiveness on two datasets, DOTA and RSOD [49]. By randomly sampling 50 images from each dataset, we optimize our adversarial patches on these images, respectively. As the size of airplanes varies between images in these datasets, we adaptively adjust the size of the patches according to the ground truth bounding box. The optimization epochs are set to 10. After optimization, we apply the patches to our Terminal-1 and Terminal-2 scenes, testing their performance across all 150 environmental variations of each scene and reporting the detection AP results in Table IV. Our method consistently outperforms the normal one under different detectors and scenes, confirming the effectiveness of our approach.

2) Results on Black-Box Transferability: Expanding the scope of our experiments, we also analyze the impact of our consistent optimization on black-box attack transferability—where adversarial patches optimized for a specific detector are used to attack other detectors not seen during the optimization phase. Conducting experiments on the Terminal-1 Scene, the results in Table V demonstrate that our method achieves better attack performance on black-box detectors than the normal one, even when not surpassing the initial white-box setup (as seen in the results of Oriented RepPoints in the Bright environment). This validates the sustained superiority of our method across different configurations, highlighting its robustness and effectiveness in diverse scenarios.

3) Integrability to Current Attack Methods: As highlighted in Section I, our visual consistency optimization is designed as a *plug-and-play* component. To further confirm its integrability, we incorporate it into recent remote sensing attack methods, namely, APPA [20] (outside variant) and CBA [32].

Conducting experiments on the Terminal-1 scene, the results are displayed in Table VI. With our consistent optimization embedded, we observe improved attack performance for both APPA and CBA in most cases. This validation underscores our method's robust integrability with current attack methods.

TABLE VI
INTEGRABILITY TO CURRENT ATTACK METHODS. THE REPORTED VALUES REPRESENT THE AVERAGE PRECISION (AP) OF THE DETECTOR. THE BEST VALUE IS SET TO BOLD.

Detector	Method	Bright	Dark	Color
ROITransformer	APPA	0.998	0.909	**1**
	APPA+Consistent	**0.921**	**0.818**	**1**
	CBA	**0**	0.182	0.091
	CBA+Consistent	**0**	**0.091**	**0**
Oriented R-CNN	APPA	1	1	1
	APPA+Consistent	**0.909**	**0.818**	**0.955**
	CBA	0.091	0.182	**0.091**
	CBA+Consistent	**0**	**0.091**	**0.091**
Oriented RepPoints	APPA	1	**0.909**	1
	APPA+Consistent	**0.909**	**0.909**	**0.937**
	CBA	0.176	0.545	0.455
	CBA+Consistent	**0**	**0.091**	**0.273**
Rotated RetinaNet	APPA	0.998	0.908	1
	APPA+Consistent	**0.909**	**0.898**	**0.913**
	CBA	0.127	0.182	0.091
	CBA+Consistent	**0.091**	**0**	**0**

TABLE VII
EFFICIENCY EXPLORATION ON A 512 × 512 IMAGE. "PARAMS" REFERS TO THE MODEL PARAMETERS, "MEM" INDICATES THE MAXIMUM MEMORY CONSUMPTION PER OPTIMIZATION ITERATION, AND "TIME" DENOTES THE OPTIMIZATION TIME PER ITERATION.

Metrics	Normal	Consistent
Params (M)	55.13	74.89
FLOPs (G)	51.48	81.71
Mem (M)	1096	1173
Time (s)	0.16	0.19

4) Efficiency Analysis: In Table VII, we provide an efficiency comparison between our consistent optimization and the normal optimization method using a single 512×512 image of Terminal-1 as an example. We assess efficiency based on model parameters (Params), calculation amount (FLOPs), memory overhead (Mem) during optimization, and runtime per iteration (Time).

From the results, we observe that incorporating a harmonization network increases model parameters and computation. However, the increase in memory and time costs is minimal, remaining within affordable limits for personal computers.

VII. Discussion of Limitations and Future Work

In our pursuit of visual consistency optimization and physical attacks against aerial detectors, we have identified several current limitations and promising directions for future research:

- **Improvement of Image Harmonization.** Although our method, as seen in Fig. 7 and Table III, exhibits a closer appearance to the physical counterpart compared

to previous methods, perfect alignment in the digital optimization phase with real-world appearance remains a challenge. Exploring enhanced image harmonization strategies is anticipated to further enhance physical attack performance.

- **Generalization to Other Fields.** While our work primarily addresses remote sensing, the digital-to-physical visual inconsistency issue may extend to other domains, such as natural imagery. Investigating the applicability of our approach in these domains could open avenues for broader applications in adversarial attacks and digital-to-physical transitions.
- **Exploration of Irregular Shape Adversarial Patches.** Current physical attacks in remote sensing rely on plain patches with regular shapes, like rectangles, which may not always suit the complexity of real-world scenarios. Future research could delve into the use of irregularly shaped patches, which could involve techniques like printing or spraying textures directly onto the surfaces of objects. This could offer more flexibility and realism in adversarial patch design.
- **Further Utilization of Simulated 3D Scenarios.** Our work introduces digital 3D simulated scenarios as a valuable platform for evaluating physical attacks in the remote sensing field. These scenarios emulate real-world scenes convincingly due to the complex graphics modeling mechanisms involved. We hope that this simulated dataset can serve as an accessible and beneficial evaluation tool for future research in the realm of remote sensing physical attacks, facilitating the development of more robust and effective attack techniques.

VIII. Conclusion

In this study, we have not only identified but effectively mitigated the critical challenge of "digital-to-physical visual inconsistency" encountered during the adversarial patch optimization phase. Our innovative approach introduces image harmonization into the field of adversarial attacks and incorporates a self-supervised training strategy to address the lack of a large-scale image harmonization dataset. We validate our designs by attacking aerial rotated object detectors, and for the sake of environmental control, we harness the capabilities of a graphics rendering engine to craft a comprehensive dataset of 3D simulated scenarios for thorough evaluation. The extensive array of experiments, meticulously conducted on these digital 3D simulated scenarios, has not only validated the existence of the digital-to-physical visual inconsistency problem but also provided compelling evidence of the efficacy of our proposed solution. This work thus contributes to the growing body of knowledge in adversarial attacks, particularly in bridging the gap between digital and physical adversarial patch optimization.

References

[1] Y. Zhang, F. Xie, J. Chen, and J. Liu, "Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis," *Computers in Biology and Medicine*, p. 106712, 2023.

[2] Y. Zhang, J. Chen, K. Wang, and F. Xie, "Ecl: Class-enhancement contrastive learning for long-tailed skin lesion classification," *arXiv preprint arXiv:2307.04136*, 2023.

[3] J. Chen, K. Chen, H. Chen, Z. Zou, and Z. Shi, "A degraded reconstruction enhancement-based method for tiny ship detection in remote sensing images with a new large-scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[4] J. Chen, K. Chen, H. Chen, W. Li, Z. Zou, and Z. Shi, "Contrastive learning for fine-grained ship classification in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[5] A. Dixit, A. K. Gupta, P. Gupta, S. Srivastava, and A. Garg, "Unfold: 3d u-net, 3d cnn and 3d transformer based hyperspectral image denoising," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[6] C. Liu, R. Zhao, and Z. Shi, "Remote-sensing image captioning based on multilayer aggregated transformer," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[7] C. Liu, R. Zhao, H. Chen, Z. Zou, and Z. Shi, "Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[8] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[12] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie *et al.*, "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems*. Springer, 2018, pp. 195–231.

[13] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, "Frequency domain model augmentation for adversarial attack," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 2022, pp. 549–566.

[14] J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, and Z. Shi, "Diffusion models for imperceptible and transferable adversarial attack," *arXiv preprint arXiv:2305.08192*, 2023.

[15] Y. Zhang, J. Chen, L. Liu, K. Chen, Z. Shi, and Z. Zou, "Generating imperceptible and cross-resolution remote sensing adversarial examples based on implicit neural representations," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[16] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[17] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 1028–1035.

[18] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "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*. Springer, 2020, pp. 665–681.

[19] Y.-C.-T. Hu, B.-H. Kung, D. S. Tan, J.-C. Chen, K.-L. Hua, and W.-H. Cheng, "Naturalistic physical adversarial patch for object detectors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7848–7857.

[20] J. Lian, S. Mei, S. Zhang, and M. Ma, "Benchmarking adversarial patch against aerial detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[21] Y. Zhang, Y. Zhang, J. Qi, K. Bin, H. Wen, X. Tong, and P. Zhong, "Adversarial patch attack on multi-scale object detection for uav remote sensing images," *Remote Sensing*, vol. 14, no. 21, p. 5298, 2022.

[22] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[23] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,"

in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[24] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 2019, pp. 52–68.

[25] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293.

[26] J. Chen, Y. Zhang, Z. Zou, K. Chen, and Z. Shi, "Dense pixel-to-pixel harmonization via continuous image representation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[27] H. Zeng, J. Cai, L. Li, Z. Cao, and L. Zhang, "Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2058–2073, 2020.

[28] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 11, pp. 7778–7796, 2021.

[29] G. W. Meyer, H. E. Rushmeier, M. F. Cohen, D. P. Greenberg, and K. E. Torrance, "An experimental evaluation of computer graphics imagery," *ACM Transactions on Graphics (TOG)*, vol. 5, no. 1, pp. 30–50, 1986.

[30] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[31] S. T. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, and G. Wang, "Connecting the digital and physical world: Improving the robustness of adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 962–969.

[32] J. Lian, X. Wang, Y. Su, M. Ma, and S. Mei, "Cba: Contextual background attack against optical aerial detection in the physical world," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[33] J.-F. Lalonde and A. A. Efros, "Using color compatibility for assessing image realism," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[34] F. Pitie, A. C. Kokaram, and R. Dahyot, "N-dimensional probability density function transfer and its application to color transfer," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1434–1439.

[35] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[36] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Deep image harmonization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3789–3797.

[37] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang, "Dovenet: Deep image harmonization via domain verification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8394–8403.

[38] Z. Ke, C. Sun, L. Zhu, K. Xu, and R. W. Lau, "Harmonizer: Learning to perform white-box image and video harmonization," in *European Conference on Computer Vision*. Springer, 2022, pp. 690–706.

[39] J. Chen, Z. Zou, Y. Zhang, K. Chen, and Z. Shi, "Zero-shot image harmonization with generative model prior," *arXiv preprint arXiv:2307.08182*, 2023.

[40] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[41] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4471–4480.

[42] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3974–3983.

[43] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[44] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[46] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3520–3529.

[47] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1829–1838.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[49] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2486–2498, 2017.