

SAaipaa: Optimizing aspect-angles-invariant physical adversarial attacks on SAR target recognition models

Isar Lemeire, Yee Wei Law, Sang-Heon Lee, Will Meakin, and Tat-Jun Chin

Abstract—Synthetic aperture radar (SAR) enables versatile, all-time, all-weather remote sensing. Coupled with automatic target recognition (ATR) leveraging machine learning (ML), SAR is empowering a wide range of Earth observation and surveillance applications. However, the surge of attacks based on adversarial perturbations against the ML algorithms underpinning SAR ATR is prompting the need for systematic research into adversarial perturbation mechanisms. Research in this area began in the digital (image) domain and evolved into the physical (signal) domain, resulting in *physical adversarial attacks* (PAAs) that strategically exploit corner reflectors as attack vectors to evade ML-based ATR. This paper proposes a novel framework called *SAR Aspect-Angles-Invariant Physical Adversarial Attack* (SAaipaa) for physics-based modeling of reflector-actuated adversarial perturbations, which improves on the rigor of prior work. A unique feature of SAaipaa is its ability to remain effective even when the attacker lacks knowledge of the SAR platform’s aspect angles, by deploying at least one reflector in each azimuthal quadrant and optimizing reflector orientations. The resultant physical evasion attacks are efficiently realizable and optimal over the considered range of aspect angles between a SAR platform and a target, achieving state-of-the-art fooling rates ($> 80\%$ for DenseNet-121 and ResNet50) in the white-box setting for a four-reflector configuration. When aspect angles are known to the attacker, an average fooling rate of 99.2% is attainable. In black-box settings, although the attack efficacy of SAaipaa transfers well between some models (e.g., from ResNet50 to DenseNet121), the transferability to some models (e.g., MobileNetV2) can be improved. A useful outcome of using the MSTAR dataset for the experiments in this article, a method for generating bounding boxes for densely sampled azimuthal SAR datasets is introduced, which leverages the inherent geometric properties of SAR imaging to produce reliable object localization across viewing angles.

Index Terms—Synthetic aperture radar, automatic target recognition, physical adversarial attack, adversarial machine learning.

ACRONYMS

AE	adversarial example
AFRL	Air Force Research Laboratory
ASC	attributed scattering center
ATR	automatic target recognition
BO	Bayesian optimization

Isar Lemeire, Yee Wei Law and Sang-Heon Lee are with UniSA STEM, University of South Australia. Will Meakin and Tat-Jun Chin are with the Australian Institute for Machine Learning, The University of Adelaide.

This paper has supplementary material available at <https://www.youtube.com/watch?v=COq-17vVEps>, which demonstrates the *SAR Aspect-Angles-Invariant Physical Adversarial Attack* (SAaipaa).

(Continued)

DARPA	Defense Advanced Research Projects Agency
DCHUN	Deep Convolutional Highway Unit Network
DE	differential evolution
DNN	deep neural network
GO	geometrical optics
KLD	Kullback-Leibler divergence
LF²B-IM	Low-Frequency and Feature Bias Iterative Method
LPF	low-pass filter
MIGAA	Metasurface Interference-Guided Adversarial Attack
ML	machine learning
MSTAR	Moving and Stationary Target Acquisition and Recognition
PAA	physical adversarial attack
PEC	perfect electric conductor
PO	physical optics
PSO	particle swarm optimization
PWFA	Positively Weighted Feature Attack
QD	quadratic demodulation
RCS	radar cross section
RDA	range-Doppler algorithm
SAaipaa	SAR Aspect-Angles-Invariant Physical Adversarial Attack
SAR	synthetic aperture radar
SAR-PeGA	SAR Perturbation Generation Algorithm
SGD	stochastic gradient descent
SMGAA	Scattering Model-Guided Adversarial Attack

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) is a microwave-based active remote-sensing paradigm that improves radar resolution in the azimuth compared to a static radar [1]. Recent years have witnessed the proliferation of space-based SAR systems due to their all-time all-weather smoke-penetrating remote sensing capabilities. For example, as of June 2025, Capella Space is operating 7 SAR satellites [2], while ICEYE is operating 48 SAR satellites [3].

A SAR transmits microwave pulses at one location and receives the corresponding echoes at subsequent locations. The transmitted and received signals are then coherently combined (i.e., combined in-phase) to create images of the illuminated

terrain [4]. A wealth of deep learning techniques can readily be leveraged to automatically recognize targets in SAR images.

From an adversarial perspective, the idea of compromising SAR imagery is compelling because SAR imagery is generally harder than optical imagery to interpret by human vision, and human users rely on algorithms, which are susceptible to cyberattacks, for interpretation. Unlike kinetic and directed energy attacks, the allure of adversarial ML attacks (“adversarial attacks” for short) lies in their stealth and their lack of tendency to escalate into physical conflicts.

More than a decade after Szegedy et al.’s discovery [5], it is now well known that deep neural networks (DNNs) are susceptible to attacks that exploit these networks’ lack of robustness in a wide range of data domains, including SAR. *Evasion attacks* are a class of adversarial attacks that manipulate test samples, creating so-called adversarial examples (AEs), to evade detection or cause a misclassification by a trained model [6]. Against DNN-based SAR ATR models, evasion attacks first emerged in the digital domain [7]–[15], where digital inputs of the targeted ML model are adversarially perturbed; and subsequently escalated to the physical domain [16]–[21]. The physically implemented form of evasion attack, called *physical adversarial attack* (PAA), manipulates objects in the physical environment the trained model gets tested on [22]. Compared to digital adversarial attacks, PAAs are more concerning [23] because the attacker does not need access to the digital inputs to the targeted model; the attacker only needs to be able to apply physical-domain perturbations to the scenes of interest.

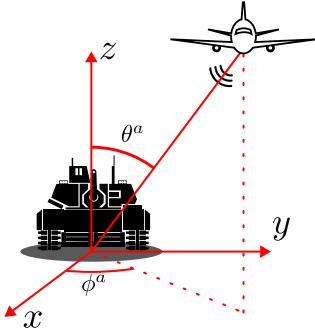


Fig. 1: A target object observed by a SAR system from incidence aspect angle θ^a and azimuth aspect angle ϕ^a . The proposed attack, SAAIPAA, is by design optimal over the ranges of θ^a and ϕ^a covered by the training dataset.

A knowledge gap in the relevant literature is however that no attack formulation so far has modeled reflector-actuated adversarial perturbation effects at different aspect angles. Without this modeling, the potency of reflectors is underutilized. Furthermore, the translation of reflector effects to adversarial perturbations in SAR images has not been adequately captured in the optimization-based formulation of the attacks so far. In this paper, we propose and analyze a new PAA against SAR ATR models, named the *SAR Aspect-Angles-Invariant Physical Adversarial Attack* (SAAIPAA). The attacker launches an evasion attack by optimally deploying simple corner reflectors,

specifically trihedral reflectors, on and near the observed target object. As an improvement to prior work, SAAIPAA considers a *realistic* attacker model in which the aspect angles, as shown in Fig. 1, are unknown, motivating its designation as aspect-angles-invariant. To this end, a loss function is formulated which incorporates rigorous physics-based modeling of trihedral reflector effects as functions of the aspect angles, to capture the translation of reflector effects to adversarial perturbations in SAR images, ensuring attack realism and feasibility. By improving the physics-based modeling of reflector-actuated perturbations, SAAIPAA narrows the gap between digital and physical attacks. SAAIPAA has been evaluated under a wide range of conditions, achieving competitive fooling rates in most experiments.

Through this article, we claim the following contributions:

- 1) We propose SAAIPAA, a novel PAA framework against DNN-based SAR ATR models that improves on prior work by dispensing with the assumption that the SAR platform’s aspect angles are known to the attacker. SAAIPAA generates physical-domain adversarial perturbations by deploying at least one reflector in each azimuthal quadrant and optimizing the orientations of the reflectors. See Sections IV-A-IV-B for details.
- 2) We formulate a loss function, by leveraging rigorous physics-based modeling of SAR observations. The generated adversarial perturbations are defined by the reflectors’ physical properties, ensuring physical feasibility and interpretability. This involves modeling the temporal response and amplitude of the reflected signal as functions of aspect angles, using physical optics (PO) and geometrical optics (GO). The resulting reflected signals are passed through the same measurement and image-focusing operators used in a typical SAR processing chain. This method explicitly simulates the entire physical imaging process rather than approximating non-physics-based perturbations in the image domain using, for example, the attributed scattering center (ASC) model. See Sections IV-C-IV-E for details.
- 3) A novel method for determining the bounding boxes for a densely sampled azimuthal SAR dataset is proposed. This method leverages the inherent properties of SAR images, to produce reliable object localization. See Section V-A for details.
- 4) A comprehensive investigation of optimization strategies is presented, including selection of the optimization algorithm, hyperparameter configuration, and efficiency-efficacy trade-off. See Sections VI-A-VI-B for details.

As a result of our contributions above, SAAIPAA is demonstrably effective under a variety of conditions (see Section VI-C), achieving competitive fooling rates even with limited training data. The adversarial perturbations exhibit strong transferability to unseen samples and models. Attack performance further improves under more favorable assumptions for the attacker, where they have partial/full information about the aspect angles.

II. RELATED WORK

Relevant to our work is the literature on applications of DNNs to SAR ATR and adversarial attacks, both digital and physical, targeting SAR ATR. Adversarial ML research in the visible-light domain is relatively well established, and there is no shortage of survey papers [23,24] that adequately summarize the state of the art of adversarial attacks in this domain. As such, the following discussion focuses solely on the SAR domain.

A. DNNs for SAR ATR

SAR images are typically hard to interpret for humans. When processing a large volume of SAR data is time-critical, a ATR system becomes necessary. Historically, ATR systems relied on traditional model-based or statistical approaches [25]. Over the past decade, the rapid advancement of DNNs has shifted ATR research toward data-driven, ML-based approaches, which significantly outperform traditional approaches [25].

Despite their promising performance, DNNs face key challenges in the SAR domain, including sensitivity to speckle noise [25] and imaging geometry [25,26], as well as a high risk of overfitting due to the limited availability of large high-quality labeled datasets [25]. This scarcity arises from the substantial cost of SAR data acquisition and the confidentiality associated with many operational datasets [25].

To address these limitations, numerous specialized DNNs have been proposed for SAR ATR [26]–[33]. Early studies focused on architectural simplification and regularization to mitigate overfitting under limited training data, as seen in models such as A-ConvNet [33] and the Deep Convolutional Highway Unit Network (DCHUN) [32]. Subsequent research shifted toward exploiting richer spatial and contextual representations through hierarchical and multi-branch feature extraction, incorporating multi-scale, multi-stream, and memory-based modules to capture the scattering behavior of targets across varying imaging geometries [26,27,30,31]. More recent developments extend this trajectory by aiming to capture broader spatial relationships within the scene, by incorporating mechanisms that expand the effective receptive field or aggregate information across distant regions in the image, allowing the network to better represent large-scale structural cues relevant to target shape and orientation [28,29]. In addition to SAR-specific models, generic optical classifiers such as AlexNet [34], DenseNet [35], MobileNet [36], and ResNet [37] have also demonstrated competitive performance on SAR ATR tasks [38].

This paper adopts AConvNet as the primary architecture for training and evaluating AEs. In addition, the generated AEs are tested on standard image classifiers, as outlined in Section VI-A3. These models were chosen as they have been widely employed in prior adversarial machine learning studies against SAR ATR [9,19]–[21], providing a consistent basis for comparison.

B. Digital attacks on SAR ATR

Three major research angles or directions can be observed in the literature: ① generation of realistic AEs, ② computational

efficiency of the generation process, ③ transferability of attacks.

1) *Realistic AEs*: Realistic AEs are those that look natural, where the perturbations are stealthy or imperceptible. The rationale for making AEs realistic is to hamper the detection of artificial perturbations, and is to be differentiated from limiting the ℓ_p -norm of the perturbations, as constraints based on ℓ_p -norm or even structural similarity [39] does not guarantee compliance with the physical laws governing the SAR imaging process. Approaches to generating realistic AEs are either data-driven or model-based.

Data-driven approaches rely on the same principle behind generative artificial intelligence [40], i.e., learning from existing artifacts to generate new, realistic artifacts, at scale, that reflect the characteristics of the training data [7].

Model-based approaches rely on a physics-based model, such as the ASC model [41], for generating artificial SAR images. A major benefit of incorporating a physics-based model is that it paves way for (but not ensure) a physical implementation, i.e., it helps elevating a digital attack to a physical attack. The ASC model is widely used to provide guidance on where in a SAR image perturbations should be made [10,14,15].

2) *Generation efficiency*: Some attack schemes focus on the computational efficiency of AE generation, for example, by optimizing the generative adversarial networks for AE generation [7,12], or accelerating a traditional digital attack [8].

3) *Transferability*: There is growing impetus for making attacks transferable [9]–[13]. Among the well-known techniques [42], the Low-Frequency and Feature Bias Iterative Method (LF²B-IM) [13] capitalizes on the ① high-dimensional features of a SAR image, accessible from the middle/intermediate layers of a neural network [43]; ② the low-frequency components of a perturbed image to preserve the main structure of the targets in the image; ③ the gradient calculation algorithm of the translation-invariant attack method [44], together with a Gaussian kernel, to extract low-frequency features. The idea of perturbing high-dimensional features originates in the observation that maximizing the distance between images and their AEs in the intermediate feature maps enhances attack transferability [43].

In the same vein, recent attacks [10,11] suppress speckle noise and perturb robust features for transferability. Bernoulli-distributed random masking [45] can suppress speckle noise [11]. A similar method to perturbing robust features is undoing non-robust perturbations through an “attenuator”, which is an encoder-decoder network designed to perturb perturbed images to restore correct classifications or predictions [12]. Another way of accentuating “important” features for transferability can be found in the Positively Weighted Feature Attack (PWFA) [11]. Maximizing the Kullback-Leibler divergence (KLD) between the positively weighted features of the original image and the positively weighted features of the perturbed image enhances transferability [11].

C. (Simulated) physical attacks on SAR ATR

Physical attacks in the optical domain cannot be directly applied to the SAR domain due to differences in the physics

of the sensing process. PAAs targeting SAR ATR systems remain confined to the simulated domain, relying on physical modeling for realism. So far, no PAAs have been physically demonstrated, hence the heading of this subsection. The challenge of evolving digital attacks to the physical domain can be boiled down to the following aspects:

1) *Choice of physical attack vectors*: For practicality, most physical attack vectors are conceived to be passive, i.e., they reflect SAR transmissions and do not produce transmissions. The application of these attack vectors is a form of *passive jamming*, which is the degradation of radar functions by reflecting or absorbing, rather than emitting, electromagnetic waves. The most commonly used passive attack vectors are corner reflectors [9,20,21], as is the case for SAAIPAA. “SAR stickers” [17,20] and triangular reflective materials [18] of uncertain physical properties have also been proposed. Passive but reactive attacks by modulating metasurfaces (e.g., active frequency-selective surfaces, phase-switched screens) in response to received signals have been considered [16,19], but metasurface technologies are still developing.

2) *Placement of physical attack vectors*: The attack vectors are placed in one of three ways: ① only on the target surface [16]–[18], ② only around the target [19,21], ③ both on and around the target [9,20], as is the case for SAAIPAA. In some designs [17,20], placement locations are informed by activation maps, for example, generated with Grad-CAM [46]. By placing attack vectors in the shadow regions, the Scattering Model-Guided Adversarial Attack (SMGAA) is not guaranteed to be physically realizable because shadow regions do not reflect SAR signals.

3) *Mapping perturbations in the physical/signal domain to perturbations in the SAR digital/image domain*: SMGAA [9] maps physical scattering to image-domain perturbations using the ASC model [47,48], but the scattering is generated with a traditional digital attack method instead of a physics-based method. The SAR Perturbation Generation Algorithm (SAR-PeGA) [16] finds phase modulation sequences for generating echos, which are then mapped by the range-Doppler algorithm (RDA) [49] into the image domain. In Zhang et al.’s attack [20], adversarial scatterings are assumed to be strong, allowing a “simple scattering model” to be used; RDA is then used to map the echo signals into pixels. The SAR-PAA attack [21] uses physical optics and the multilevel fast multipole method [50] to determine the radar cross section (RCS) of a target and surrounding scatterers, and generate an image from the RCS using the polar formatting algorithm [51]. SAR-PATT [18] relies on the RaySAR simulator [52], which however requires material property data that is scarcely available. In the Metasurface Interference-Guided Adversarial Attack (MIGAA) [19], phase-switched screens are modulated using rectangular waves [53] and SAR images are formed using RDA. Nevertheless, physical-to-digital mapping of perturbations is not clearly articulated in every attack design [17]. In contrast, the mapping used in SAAIPAA is entirely physics-based.

4) *Optimization formulation*: The most common formulation of an attack is maximizing the loss function of the targeted classifier [9,19,21], which is usually the cross-entropy loss;

this is the same formulation used in SAAIPAA. The optimization problem can alternatively be formulated as maximizing extent of misclassification [17,20], or minimizing a linear combination of negative cross-entropy loss and perturbation [18]. In SAR-PeGA [16], the optimization problem is finding the phase modulation sequence closest to a Universal Adversarial Perturbation [54] (a digital attack) pattern.

5) *Transferability*: Activation maps have been used to guide the placement of physical attack vectors [17,20], following the use of these maps in the RGB domain [55]. Besides activation maps, most physical attacks do not apply specific transferability techniques, although they have been evaluated for transferability.

Digital AEs in the SAR domain have higher fooling rates and are more transferable than those in the optical domain [56], but no such statement can be made for physical AEs because of the significant differences in how these physical examples can be implemented in the SAR domain and in the optical domain.

III. ATTACKER MODEL

The following attacker model specifies assumptions about the goal, capabilities and constraints of the attacker, as well as assumptions about the SAR system targeted by the attacker, for SAAIPAA.

A. Assumptions about the attacker

The attacker seeks to launch an *untargeted* evasion attack, i.e., produce AEs that cause instances of the target class to be misclassified into any other class [6]. Through the untargeted evasion attack, the attacker’s ultimate goal is violating the integrity of the targeted ML model [6].

The following assumptions are made about the attacker’s knowledge:

- The attacker has full knowledge of the ATR ML model. In Sec. VI, this assumption is relaxed for transferability evaluation.
- The attacker does not know the aspect angles from which the SAR system will observe the scene.
- The attacker knows the technical specifications of the SAR system.

Concerning physical implementation, we assume the attacker can place corner reflectors on the ground swath and on the target object, and has the time and resources to determine their deployment locations and orientations.

B. Assumptions about the SAR system

The following assumptions are commonly made in the literature [16,57,58] and are adopted here:

- The SAR system operates in the spotlight mode.
- The SAR system operates in HH mode, i.e., transmitting and measuring received signals in horizontal polarization.
- The SAR system uses quadratic demodulation (QD) to demodulate the signal.
- The SAR system uses RDA [49] as the image formation algorithm.

IV. PROPOSED ATTACK

This section introduces SAAIPAA, starting with the choice and placement of physical attack vectors, followed by the overall optimization formulation, and the mapping of signal-domain perturbations to the image domain. The last part involves details of the SAR imaging process, a physics-based reflection model, backscatter measurement and image formation.

A. Choice and placement of physical attack vectors

SAAIPAA uses trihedral corner reflectors as attack vectors, due to their passive, low-cost nature combined with their ability to produce a bright, localized radar return. The physical perturbation is actuated by m corner reflectors, where each reflector i is parameterized by its position $\vec{p}_i = (x_i, y_i)$, and boresight incidence angle θ_i and azimuth angle ϕ_i . To ensure maximal azimuthal coverage, their azimuth angles are mutually constrained to be uniformly distributed:

$$\phi_i = \phi_1 + (i-1) \frac{2\pi}{m}. \quad (1)$$

Thus, the physical-domain perturbation is parameterized by:

$$\Theta = [x_1, \dots, x_m, y_1, \dots, y_m, \theta_1, \dots, \theta_m, \phi_1]. \quad (2)$$

With $\phi_1 \in [0, \frac{2\pi}{m}]$, and $\forall i : \theta_i \in [0, \frac{\pi}{2}] \wedge x_i \in [-\frac{w}{2}, \frac{w}{2}] \wedge y_i \in [-\frac{h}{2}, \frac{h}{2}]$ where w, h are the scene width and height of the observed scene. Each corner reflector yields a strong return over an azimuthal span of $\frac{\pi}{2}$ [59]. Therefore, $m \geq 4$ ensures at least one corner reflector produces a strong return for any azimuth aspect angle.

B. Objective function

Let $\mathcal{O}$ denote the SAR imaging operator, so $\mathcal{O}(\mathcal{S}, \theta^a, \phi^a)$ is the image of scene $\mathcal{S}$ observed from incidence aspect angle θ^a and azimuth aspect angle ϕ^a , as shown in Fig. 1. SAAIPAA seeks to add a physical perturbation $\tilde{\mathcal{S}}(\Theta)$ parameterized by Θ to the scene $\mathcal{S}$, that causes misclassifications across the entire viewing domain $\theta^a \in [0, \frac{\pi}{2}] \wedge \phi^a \in [0, 2\pi]$. The continuous viewing domain is approximated by a finite set of N SAR observations, with $\{(\theta_n^a, \phi_n^a)\}_{n=1}^N$. The total high-frequency scattering response of the perturbed scene can be approximated as a linear superposition of individual scatterers [41,60]. Accordingly, the perturbed image is approximated by the superposition of the clean image and the perturbation:

$$\mathcal{O}(\mathcal{S} + \tilde{\mathcal{S}}(\Theta), \theta^a, \phi^a) \approx \mathcal{O}(\mathcal{S}, \theta^a, \phi^a) + \mathcal{O}(\tilde{\mathcal{S}}(\Theta), \theta^a, \phi^a). \quad (3)$$

The optimal Θ , for a target class c with label l_c and target model f , is obtained by maximizing the average cross-entropy loss $\mathcal{L}_{CE}$:

$$\begin{aligned} \min_{\Theta} \quad & \frac{1}{N} \sum_{n=1}^N \mathcal{L}_{CE} \left(f \left(\mathcal{O}(\mathcal{S}, \theta_n^a, \phi_n^a) + \mathcal{O}(\tilde{\mathcal{S}}(\Theta), \theta_n^a, \phi_n^a) \right), l_c \right), \\ \text{s.t.} \quad & \phi_1 \in \left[0, \frac{2\pi}{m} \right] \wedge \forall i \in \{1, \dots, m\} : \\ & \theta_i \in \left[0, \frac{\pi}{2} \right] \wedge x_i \in \left[-\frac{w}{2}, \frac{w}{2} \right] \wedge y_i \in \left[-\frac{h}{2}, \frac{h}{2} \right]. \end{aligned} \quad (4)$$

The attack strategy is illustrated in Fig. 2.

C. SAR Imaging process

Given the assumptions specified in Section III-B, creating a SAR image involves the following sequential steps:

- 1) As the SAR platform traverses a predefined flight path, it transmits a sequence of identical signal pulses $E^t(t)$, expressed as a function of fast time t , toward the scene.
- 2) The transmitted signals are reflected by objects within the scene, where each pulse yields a different reflection at a different point in slow time η resulting in a backscattered signal $E^r(t, \eta)$.
- 3) The reflected signal is measured, demodulated using QD, and sampled over fast and slow time, resulting in a two-dimensional matrix.
- 4) The demodulated signal is focused into a SAR image using RDA.

Thus, the image produced by the physically perturbed scene is computed by modeling the SAR imaging process:

$$\mathcal{O}(\tilde{\mathcal{S}}(\Theta), \theta_n^a, \phi_n^a) = \text{RDA} \left(\text{QD} \left(\sum_{i=1}^m E_{i,n}^r(t, \eta) \right) \right), \quad (5)$$

where $E_{i,n}^r(t, \eta)$ is the reflected signal from the i -th corner reflector for the n -th observation, $\text{QD}(\cdot)$ denotes quadratic demodulation operator and $\text{RDA}(\cdot)$ denotes the range-Doppler algorithm operator.

D. Physics-based reflection model

The following shows a derivation of an expression for $E_{i,n}^r(t, \eta)$. For a fixed reflector i and observation n , let us omit the indices (i, n) for brevity. Accordingly, (θ, ϕ) and $\vec{p}$ denote the orientation and position of said corner reflector, while θ^a, ϕ^a denote the aspect angles.

Each transmitted pulse $E^t(t)$ is a linear frequency-modulated waveform, commonly called a *chirp* [1]:

$$E^t(t) = A^t \text{rect} \left(\frac{t}{T} \right) \cos \left(2\pi f_0 t + \pi K t^2 \right), \quad (6)$$

where A^t is the amplitude of the transmitted pulse, T is the pulse duration, f_0 is the center frequency, K is the chirp rate, and rect is the rectangle function. The SAR platform follows a straight path, perpendicular to its line of sight at a distance r_0 , as shown in Fig. 1 and Fig. 3. Its position along the path is given by:

$$\vec{p}^{\text{SAR}}(\eta) = \begin{pmatrix} \cos(\phi^a) & -\sin(\phi^a) & 0 \\ \sin(\phi^a) & \cos(\phi^a) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} r_0 \sin(\theta^a) \\ \eta v \\ r_0 \cos(\theta^a) \end{pmatrix}. \quad (7)$$

Each corner reflector is modelled as a point scatterer for the purpose of modeling the temporal structure of the reflected signal. Consequently, the reflected signal is a delayed and attenuated copy of the transmitted signal [1]:

$$E^r(t, \eta) = \frac{A^r(\eta)}{A^t} E^t(t - \tau(\eta)), \quad (8)$$

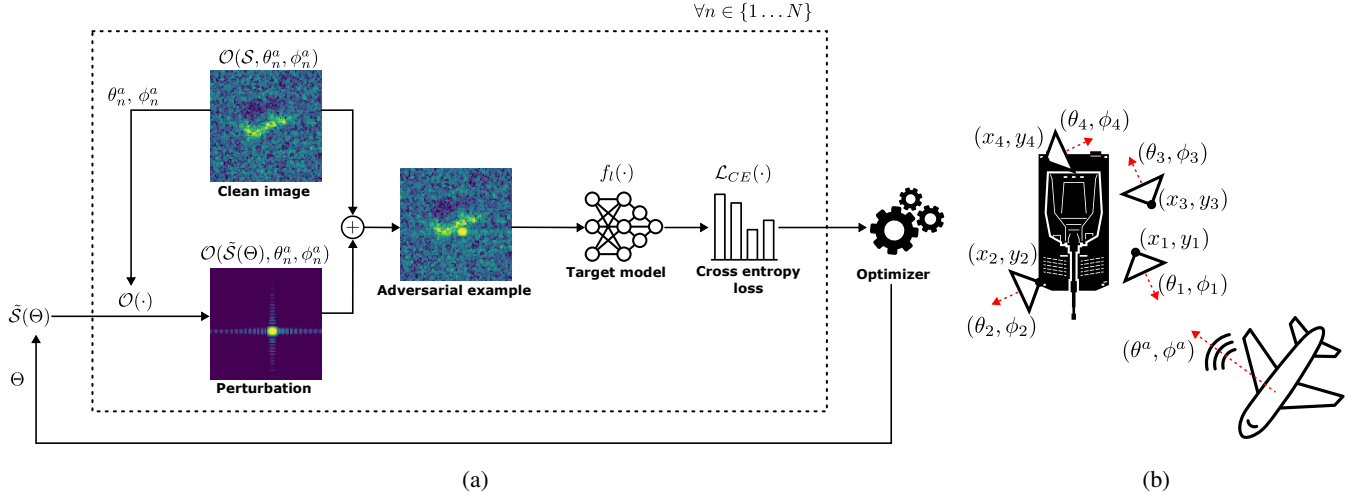


Fig. 2: The strategy of SAAIPAA: (a) Physical-domain adversarial perturbations actuated by m reflectors are optimized over N observations through Eq. (4). (b) Top view of a sample reflector configuration, where $m = 4$. Each i -th reflector is deployed optimally at position (x_i, y_i) with orientation (θ_i, ϕ_i) , in a scene observed by a SAR platform from angles (θ^a, ϕ^a) .

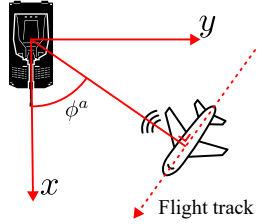


Fig. 3: Top view of a target object (e.g., tank) observed by a SAR system from azimuth aspect angle ϕ^a . The flight track is assumed to be perpendicular to the line of sight.

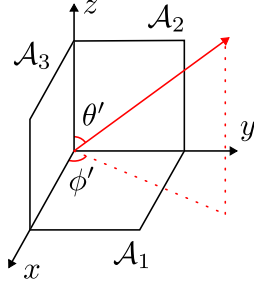


Fig. 4: Numbered plates A_1 , A_2 , A_3 of a square trihedral corner reflector. The coordinate frame is aligned with the plates, and aspect angles θ' , ϕ' are defined in this frame.

where $\tau(\eta) = \frac{2r(\eta)}{v_{\text{light}}}$ is the round trip time, $r(\eta) = \|\vec{p}^{\text{SAR}}(\eta) - \vec{p}\|$ is the distance between the corner reflector and the SAR platform, v_{light} is the speed of light, and $A^r(\eta)$ is the amplitude of the reflected signal.

To quantify $A^r(\eta)$, a method developed by Polycarpou et al. [61], which has been shown to produce predictions closely aligned with experimental measurements, is used. The transmitted signal is approximated as having a constant frequency f_0 , so that the magnitude of the reflected electric fields can be expressed in terms of the far-field spherical components

$E^\theta(\eta)$, $E^\phi(\eta)$ [62]:

$$|A^r(\eta)|^2 \approx |E^\theta(\eta)|^2 + |E^\phi(\eta)|^2. \quad (9)$$

Each spherical component is a summation of fifteen reflection components, each caused by a different reflection path p [61]:

$$E^\theta(\eta) \approx \sum_{p \in \mathcal{P}} E_p^\theta(\eta), \quad E^\phi(\eta) \approx \sum_{p \in \mathcal{P}} E_p^\phi(\eta), \quad (10)$$

$$\mathcal{P} = \{1, 2, 3, 12, 21, 13, 31, 23, 32, 123, 132, 213, 231, 312, 321\},$$

where, assuming A_1 , A_2 , A_3 define the plates of the reflector as illustrated on Fig. 4, $p = 1$ corresponds to a single bounce off surface A_1 , $p = 12$ corresponds to the reflection path with a double bounce, first off surface A_1 then surface A_2 , and $p = 123$ corresponds to the reflection path with a triple bounce in the order of surfaces A_1 , A_2 , and A_3 . The remaining terms follow the same naming convention. Different reflector geometries can be incorporated into the model by varying the polygonal description of A_1 , A_2 , A_3 . In this work, square plates of dimensions $0.3\text{m} \times 0.3\text{m}$ are assumed.

To simplify the analysis, the aspect angles are expressed relative to the corner reflector's boresight [59]:

$$\phi' = \phi^a - \left(\phi - \frac{\pi}{4}\right), \quad \theta' = \theta^a - \left(\theta - \arctan(\sqrt{2})\right), \quad (11)$$

aligning the coordinate frame so that each plate of the trihedral aligns with a coordinate plane, as illustrated in Fig. 4. Only $\phi', \theta' \in [0, \frac{\pi}{2}]$ are considered, as scattering outside this range is negligible [59].

The current density $\vec{J}_p$ on the final reflecting plate A_s , $s \in \{1, 2, 3\}$, of reflection path p is modeled using PO for perfect electric conductor (PEC) surfaces:

$$\vec{J}_p = 2\hat{n}_s \times \vec{H}_p = 2\hat{n}_s \times \frac{A^t}{Z_0} e^{-jk(\hat{k}_p \cdot \vec{r})} \hat{h}_p, \quad (12)$$

where $\vec{H}_p$ is the incident magnetic field, $\hat{n}_s$ is the normal vector of $\mathcal{A}_s$, Z_0 is the intrinsic impedance of free space, $k \approx 2\pi f_0/v_{\text{light}}$ is the phase constant, $\vec{r} = [x, y, z]^T$, $\hat{k}_p$ and $\hat{h}_p$ denote the direction of travel and polarization of $\vec{H}_p$ respectively. The backscattered far-field components caused by the reflection path p are [61,62]:

$$\begin{aligned} E_p^\theta(\eta) &\approx \frac{-jkZ_0N_p^\theta}{4\pi} \frac{e^{-jkr(\eta)}}{r(\eta)}, \\ E_p^\phi(\eta) &\approx \frac{-jkZ_0N_p^\phi}{4\pi} \frac{e^{-jkr(\eta)}}{r(\eta)}, \\ N_p^\theta &= \iint_{\mathcal{A}_p^i} \left(\vec{J}_p \cdot \begin{pmatrix} \cos(\theta') \cos(\phi') \\ \cos(\theta') \sin(\phi') \\ -\sin(\theta') \end{pmatrix} \right) e^{jkL} dA, \\ N_p^\phi &= \iint_{\mathcal{A}_p^i} \left(\vec{J}_p \cdot \begin{pmatrix} -\sin(\phi') \\ \cos(\phi') \\ 0 \end{pmatrix} \right) e^{jkL} dA, \\ L &= x \sin(\theta') \cos(\phi') + y \sin(\theta') \sin(\phi') + z \cos(\theta'), \end{aligned} \quad (13)$$

where $\mathcal{A}_p^i$ is the illuminated area on the final reflecting plate, N_p^θ and N_p^ϕ are the far-field integrals.

Thus, the far-field integrals N_p^θ , N_p^ϕ depend on the incident magnetic wave H_p and the illuminated area of the final reflecting plate $\mathcal{A}_p^i$ within the reflection path p . For single-bounce reflection paths, the reflecting plate is entirely illuminated, i.e., $\mathcal{A}_p^i = \mathcal{A}_s$ for $\theta^a, \phi^a \in [0, \frac{\pi}{2}]$; and the incident magnetic field is identical to the transmitted field, i.e., $H_1 = H_2 = H_3 = H^t$, defined by direction of travel $\hat{k}^t$ and polarization $\hat{h}^t$:

$$\hat{k}^t = - \begin{pmatrix} \sin(\theta') \cos(\phi') \\ \sin(\theta') \sin(\phi') \\ \cos(\theta') \end{pmatrix}, \quad \hat{h}^t = \begin{pmatrix} \cos(\theta') \cos(\phi') \\ \cos(\theta') \sin(\phi') \\ -\sin(\theta') \end{pmatrix}. \quad (14)$$

For double- and triple-bounce reflections, preceding bounces are modeled using GO, where each reflection is treated as specular. This approach determines the propagation and polarization directions, $\hat{k}_p$ and $\hat{h}_p$, of the reflected wave, as well as the illuminated area of the final reflecting surface $\mathcal{A}_p^i$. The illumination areas for multi-bounce interactions are obtained through sequential geometric projection. The corresponding far-field integrals are provided in Section A.

E. Backscatter measurement and image formation

SAR systems typically record a single polarization channel. In this work, the system is assumed to operate in HH mode, so measurements are performed in horizontal polarization, denoted by $\mathcal{M}^H$, corresponding to measuring $|E^\phi|$. Afterwards, the measured signal is demodulated using QD, by mixing the received signal with the complex carrier signal $e^{-j2\pi f_0 t}$, and applying a low-pass filter (LPF) to isolate the baseband component [1]:

$$\begin{aligned} \text{QD}(E^r(t, \eta)) &= \text{LPF} \left(e^{-j2\pi f_0 t} \mathcal{M}^H(E^r(t, \eta)) \right), \\ \text{QD}(E^r(t, \eta)) &= |E^\phi(\eta)| \text{rect} \left(\frac{t - \tau(\eta)}{T} \right) \\ &\quad e^{j\pi[2f_0\tau(\eta) - K(t - \tau(\eta))^2]}. \end{aligned} \quad (15)$$

Afterwards, RDA [49] is applied to focus the demodulated signal into a SAR image.

V. DATASET

Short of evaluating the proposed attack on a SAR, a dataset needs to be generated with a simulator or acquired from a third party. In the absence of a simulator capable of simulating a wide mix of material properties, the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset [58] is chosen. Developed in the 1990s by Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory (AFRL), the MSTAR dataset remains widely used, including for the evaluation of PAAs [9,16,18]–[21].

The MSTAR dataset contains labeled high-resolution X-band SAR images of military vehicles and targets, specifically 2S1, BMP-2, BDRM-2, BTR-60, BTR-70, D7, T-62, T-72, ZIL-131, and ZSU-23-4. The HH-polarized samples were captured in spotlight mode, quadratic-demodulated and focused using the RDA, consistent with the assumptions made in Section III-B. The data was acquired at four different incidence angles, but for each incidence angle, the full $[0, 2\pi]$ azimuthal range is densely populated with samples. For a given class, all samples are images of the same physical scene, captured from different aspect angles. A detailed summary of the incidence angles and the number of samples per class per angle is provided in Table I. The high variability of aspect angles makes the MSTAR well suited for evaluating SAAIPAA.

TABLE I: Number of samples per class in the MSTAR dataset, classified by the incidence aspect angle θ^a (degrees).

Class label	75°	73°	60°	45°
2S1	274	299	288	303
BMP-2	195	233	0	0
BTR-60	195	256	0	0
BTR-70	196	233	0	0
D7	274	299	0	0
T-72	196	232	0	0
T-62	273	299	0	0
ZIL-131	274	299	0	0
ZSU-23-4	274	299	406	422
BDRM-2	274	298	420	423

A. Bounding boxes

Each MSTAR sample contains a single target roughly centered in the image, although small misalignments occur across samples. To ensure consistent placement of the corner reflectors, bounding boxes are defined around the target, and the reflectors are shifted to maintain a fixed position relative to the bounding box. Simple methods for defining bounding boxes, such as selecting the brightest rectangle, are unreliable because only the portion of the target facing the radar produces a strong return. While alternative approaches exist [63,64], a novel method was developed that is simple, reliable, and well-suited to the densely sampled, aspect-angles-annotated MSTAR dataset.

The proposed bounding-box method begins with the estimation of the area occupied by each object (representing one class). The dense azimuth sampling of MSTAR allows an

object's area to be inferred across multiple azimuth angles. For each class and incidence angle, the images are first azimuth-aligned by rotating each by the negative of its azimuth aspect angle and then averaged to produce a composite image, as shown in Fig. 5a. The composite image is converted to logarithmic scale and normalized. A bounding box is then constructed by δ -thresholding the image to create a binary mask and fitting a minimum-area rotated rectangle (also known as oriented bounding box), R^{ref} , around the largest contour, as shown in Fig. 5b. The rectangle provides the reference dimensions for the bounding box, which is localized in each image using the procedure discussed below.

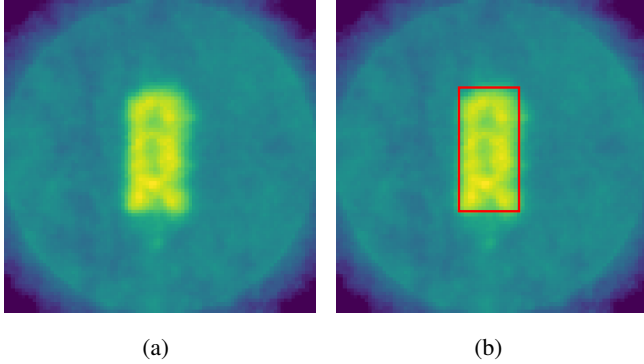


Fig. 5: Composite image created by aligning and averaging all images of the T-62 for $\theta^a = 75^\circ$: (a) without bounding box and (b) with bounding box R^{ref} .

For each sample of the same class and incidence angle θ^a , a rectangle R with the same dimensions as R^{ref} is positioned with its rotation aligned to the sample's azimuth aspect angle ϕ^a . Prior to fitting, the images are preprocessed, specifically logarithmically scaled, thresholded at 0.5, and gamma-corrected at 1.5. To localize R , a loss function is defined, consisting of two components:

- 1) The pixel-based loss $\mathcal{L}_{\text{pixel}}(R)$: This component rewards bright pixels near the bottom of the bounding box, corresponding to the side of the target facing the SAR system. This reflects the inherent property of SAR images, in which surfaces oriented toward the radar produce stronger returns:

$$\mathcal{L}_{\text{pixel}}(R) = -\frac{\sum_{x,y} I(x,y) \mathcal{M}(x,y,R) d(y,R)^\alpha}{|R|}, \quad (16)$$

where $I(x,y)$ is the intensity of the processed image at pixel (x,y) , $\mathcal{M}(x,y,R)$ is the binary mask of R (1 inside the rectangle, 0 outside), $d(y,R)$ is a vertical weighting:

$$d(y,R) = \frac{y - y_{\min}^R}{y_{\max}^R - y_{\min}^R}, \quad (17)$$

$|R|$ is the number of pixels in the rectangle, and α is a hyperparameter.

- 2) The distance-based loss $\mathcal{L}_{\text{dist}}(x^R, y^R)$: This component penalizes displacement from the center of R^{ref} in the

composite image, denoted by $(x^{\text{ref}}, y^{\text{ref}})$:

$$\mathcal{L}_{\text{dist}}(x^R, y^R) = \left(\frac{(x^R - x^{\text{ref}})^2 + (y^R - y^{\text{ref}})^2}{d_{\max}} \right)^\beta, \quad (18)$$

where $d_{\max} = \sqrt{x_{\max}^2 + y_{\max}^2}$ is the maximum displacement allowed, and β is a hyperparameter.

The bounding box is obtained by solving:

$$\begin{aligned} \min_{x^R, y^R} \mathcal{L}_{\text{pixel}}(R) + \lambda \mathcal{L}_{\text{dist}}(x^R, y^R), \\ \text{s.t. } x^R \in [-x_{\max}, x_{\max}] \wedge y^R \in [-y_{\max}, y_{\max}], \end{aligned} \quad (19)$$

where λ is a hyperparameter controlling the relative weight of the distance-based loss. Bounding boxes were generated by this hyperparameter configuration: $\delta = 0.7$, $\alpha = 1.5$, $\beta = 0.5$, and $\lambda = 0.1$. For demonstration, Fig. 6c shows the bounding box for the sample in Fig. 6a, utilizing the weighted mask $\mathcal{M}(x,y,R) d(y,R)^\alpha$ shown in Fig. 6b.

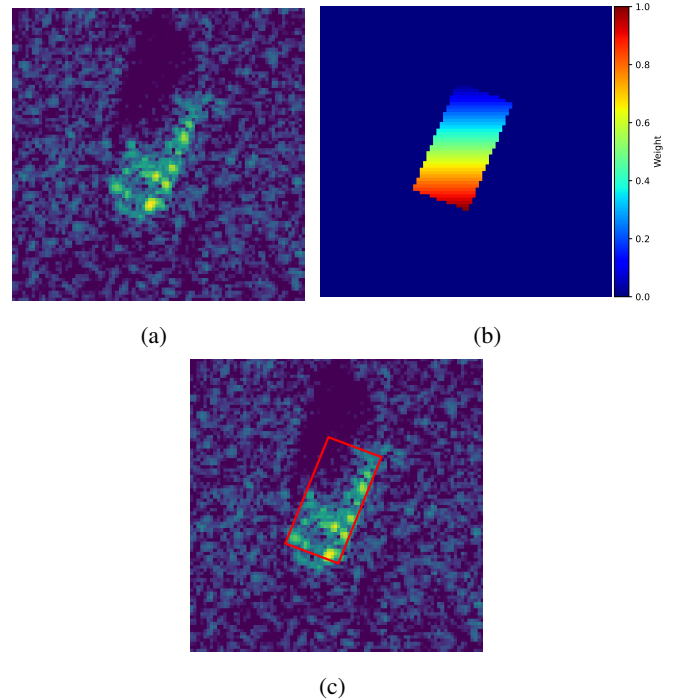


Fig. 6: A SAR image sample of (a) a T-62 observed from $\theta^a = 75^\circ$ and $\phi^a = 200.2^\circ$, (b) the weighted mask $\mathcal{M}(x,y,R) d(y,R)^\alpha$ used to find R , and (c) the bounding box R found by solving (19).

B. Data sampling for evaluating SAAIPAA

SAAIPAA optimizes an adversarial perturbation for each class in the MSTAR dataset. For perturbation optimization, a training set is derived from the dataset using Algorithm 1, with a default azimuth spacing of 10° and a tolerance of 2° . From the remainder of the dataset, a test set is derived to evaluate the optimized perturbation. The test set is derived using Algorithm 1 with an azimuth spacing of 2.5° and a tolerance of 1° .

Algorithm 1 Sampling images for a given class and azimuth spacing

Require: Set of available samples $\mathcal{D}$ for class c , azimuth spacing $\Delta\phi^a$, tolerance ϵ
Ensure: Subset of samples $\mathcal{G}$ uniformly distributed in azimuth for each incidence angle

- 1: $\mathcal{G} \leftarrow \emptyset$ $\triangleright$ Initialize empty set of selected samples
- 2: **for** each incidence angle θ present in $\mathcal{D}$ **do**
- 3: Draw random offset $\phi_0^a \sim U(0, \Delta\phi^a)$ $\triangleright$ Randomize azimuth starting point
- 4: $\phi \leftarrow \phi_0^a$
- 5: **while** $\phi < 2\pi$ **do**
- 6: Select random sample $s \in \mathcal{D}$ such that $\theta_s^a = \theta$ and $|\phi_s^a - \phi| \leq \epsilon$ $\triangleright$ Find sample near target azimuth angle
- 7: $\mathcal{G} \leftarrow \mathcal{G} \cup \{s\}$ $\triangleright$ Add selected sample to subset
- 8: $\phi \leftarrow \phi + \Delta\phi^a$
- 9: **end while**
- 10: **end for**
- 11: **return** $\mathcal{G}$

The test set is deliberately chosen to be densely populated in the azimuthal range, such that it approximates the continuous viewing domain. Meanwhile, the training set is sampled more coarsely to limit the computational cost of training. The train-test gap in fooling rate is evaluated in Section VI-C2.

VI. EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the proposed SAAIPAA. Different optimization algorithms were compared. Comprehensive experiments were conducted to evaluate the attack performance of SAAIPAA under various conditions, such as different numbers of reflectors, limited training data, and transferability to unseen models. Additionally, SAAIPAA was evaluated under the assumption that the attacker has partial (as opposed to no) knowledge of the aspect angles.

A. Experimental setup

As part of the experimental setup, the SAR system parameters and evaluation metrics were specified. ATR models to evaluate SAAIPAA against were developed. Candidate optimization algorithms for solving (4) were identified.

1) *SAR system specification:* The reflected signal $E^r(t, \eta)$ and image formation depend on the SAR system's technical specifications. Certain parameters, such as range r , center frequency f_0 , bandwidth $B = KT$, pixel ground spacing, and polarization can be extracted from the provided metadata [58]. The remaining parameters, including pulse duration T , platform speed v , pulse repetition frequency, and sample rate were estimated based on typical values reported for comparable SAR platforms [65]–[68], and further refined to ensure that simulated SAR images are properly focused. Table II specifies the values of all system parameters in use.

The amplitude of the transmitted signal, A^t , and the MSTAR images are not directly available, as the images are stored in a relative, arbitrary scale. To account for this unknown

TABLE II: Specification of the simulated SAR system.

Variable	Value
Range	{4500m, 5000m}
Platform speed	50 m/s
Center frequency	9.6 Ghz
Bandwidth	591 Mhz
Pulse duration	5 μ s
Sample rate	500 Mhz
Pulse repetition rate	1200 Hz
Ground sample distance	0.3m $\times$ 0.3m
Polarization	HH

scaling, the amplitude of the transmitted pulse A^t was chosen so that a corner reflector of dimensions $0.3\text{m} \times 0.3\text{m} \times 0.3\text{m}$, observed from its boresight, yields a peak intensity equal to the average maximum pixel intensity across all target classes at their respective brightest aspect angles. This choice yields physically plausible results. Any required rescaling for practical deployment can be achieved by scaling the physical size of the reflectors, and thus does not affect the attack framework or associated qualitative conclusions.

2) *Evaluation metrics:* Fooling rate [54] or attack success rate [69] is originally called error rate [70]. The fooling rate of an attack $\mathcal{A}$ against a model f applied to dataset $\mathcal{D}$ is often defined informally [71,72], but below, the fooling rate specific to class c is formally the proportion of data in $\mathcal{D}$ that are correctly classified by f in the absence of $\mathcal{A}$ as l_c , but are misclassified by f in the presence of $\mathcal{A}$:

$$\frac{\sum_{\mathbf{X} \in \mathcal{D}} \mathbb{I}\{\hat{y}(\mathbf{X}) = l_c \wedge \hat{y}(\mathcal{A}(\mathbf{X})) \neq l_c\}}{\sum_{\mathbf{X} \in \mathcal{D}} \mathbb{I}\{\hat{y}(\mathbf{X}) = l_c\}}, \quad (20)$$

where $\mathbb{I}$ is the indicator function, $\hat{y}(\mathbf{X})$ is the predicted label for image $\mathbf{X}$.

For each object class, the fooling rate is computed using Eq. (20). The fooling rates for all classes are then averaged to produce the *average fooling rate*, which is the main metric for evaluating attack efficacy in this article. In the MSTAR dataset, there is one scene per object class, and multiple images or observations per scene, so the average fooling rate is equivalently an average taken over all scenes.

3) *ATR models:* SAAIPAA was evaluated against AConvNet [33], a model specifically designed for SAR ATR with a strong performance [33]. To assess transferability across architectures, four additional widely used convolutional networks were implemented: AlexNet [34], DenseNet-121 [35], MobileNetV2 [36], and ResNet50 [37]. These models were chosen to represent a diverse set of architectures, and their extensive usage in prior adversarial machine learning studies targeting SAR ATR [9,19]–[21].

All models were trained with standard SAR data augmentations (sliding-window translation, random rotation, scaling, and additive random noise) to emulate realistic SAR imaging variability and prevent overfitting [73,74]. A 70/20/10 train/validation/test split was used. Training was performed using stochastic gradient descent (SGD), with a batch size of 32, 0.001 learning rate, and 0.9 momentum for 100 epochs. Table III shows the test accuracies.

TABLE III: A summary of target ATR models.

Model	Test accuracy
AConvNet	99.6%
AlexNet	99.7%
DenseNet-121	99.3%
MobileNetV2	98.8%
ResNet50	99.2%

4) *Optimization algorithms*: Three optimization algorithms were investigated. Two evolutionary algorithms, specifically differential evolution (DE) [75] and particle swarm optimization (PSO) [76], were selected based on their effectiveness in adversarial optimization tasks against SAR ATR models [19]–[21] and their ability to navigate nonconvex, discontinuous loss surfaces. For diversity, Bayesian optimization (BO) [77] was included as a model-based alternative, motivated by its potential for sample-efficient search given the low dimensionality of the search space (e.g., 13 variables for 4 reflectors).

B. Finetuning optimization

The first set of experiments focused on finetuning the optimization procedure for maximizing the average fooling rate. Specifically, optimization algorithms were compared, and for the top-performing optimization algorithm, the impact of hyperparameter variation on optimization performance was studied. The impact of the choice of optimization variables (which angles of the corner reflectors to fix, and which angles to optimize) on attack performance was also studied. These experiments established the baseline configuration adopted in the remainder of this work.

1) *Varying optimization algorithm*: PSO, DE, and BO were investigated. Each algorithm was run until the loss function converged, ensuring that differences in performance were not due to early termination. Table IV summarizes the configuration of each optimizer and the corresponding average fooling rates. DE achieved the highest fooling rate and the lowest final loss, as shown in Fig. 7. BO underperformed, potentially due to the unsatisfactory fit of a Gaussian process to the objective function. In this case, the global exploration strategy of the metaheuristics is also potentially more effective at evading local optima than BO’s exploration/exploitation trade-off.

2) *Varying optimizer hyperparameters*: The application of DE was further finetuned with a hyperparameter study. Mutation and recombination probabilities were varied, while other hyperparameters were kept identical to the earlier experiments. Table V summarizes the configurations and results. The configuration with 0.8 mutation probability and 0.9 recombination probability achieved the best average fooling rate, and the lowest loss as shown in Fig. 8. The small variation across configurations indicates that DE is relatively robust. The benefit of high mutation and recombination rates indicates a rugged loss landscape favoring larger exploratory steps.

3) *Varying optimization variables*: As an attack vector, each corner reflector is parameterized by its boresight incidence angle, θ , and its boresight azimuth angle, ϕ . Experiments were conducted to assess whether fixing θ and/or ϕ improves convergence by reducing dimensionality and hence

TABLE IV: Summary of hyperparameters and average fooling rates for various optimizers.

Optimizer	Hyperparameters	Average fooling rate
BO	Nr. of initial points = 150 Max iterations = 700 Smoothness of kernel = 2.5 Exploration–exploitation trade-off = 0.1	52.1%
DE	Population size = 40 Max iterations = 60 Mutation = 0.5 Recombination = 0.7 Crowding for crossover = 20 Mutation probability = 0.8 Tournament selection size = 3	60.8%
PSO	Nr. of particles = 40 Max iterations = 60 Cognitive learning rate = 0.6 Social learning rate = 1.0 Inertia weight = 0.8	58.7%

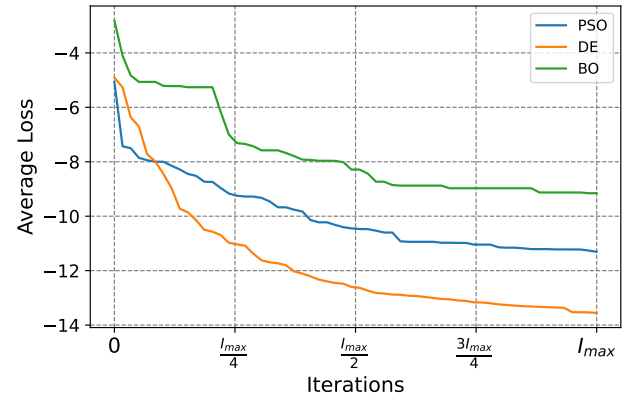


Fig. 7: Average loss per iteration for each optimizer, where $I_{\max}$ is the maximum number of iterations. The lowest loss is registered by Differential Evolution (DE).

search space, without sacrificing attack performance. For the DE-based optimizer, configuration 9 from Table V was used. Four corner reflectors were used, ensuring at least one is visible from any azimuth. Four choices of optimization variables were investigated:

- Configuration 1: θ_i and ϕ_i ($i = 1, \dots, 4$) are free variables.
- Configuration 2: The incidence angles, θ_i , are fixed to $\arctan \sqrt{2}$, while ϕ_i are free.
- Configuration 3: The azimuth angles, ϕ_i , are fixed as per Eq (1) (starting with $\phi_1 = 0$), while θ_i are free.
- Configuration 4: θ_i are fixed as per configuration 2, while ϕ are fixed as per configuration 3.

Table VI records the resultant average fooling rates. Fig. 9 shows the reduction of loss over iterations.

As Table VI shows, fixing the incidence angle θ substantially reduced fooling rates, highlighting its critical role in determining backscattered amplitude. A dataset with greater incidence-angle variation would allow a more thorough analy-

TABLE V: Results for different mutation and recombination configurations.

Parameter configuration	Mutation probability	Recombination probability	Average fooling rate
1	0.3	0.5	61.4%
2	0.5	0.5	62.9%
3	0.8	0.5	62.2%
4	0.3	0.7	59.1%
5	0.5	0.7	61.9%
6	0.8	0.7	60.3%
7	0.3	0.9	61.6%
8	0.5	0.9	63.1%
9	0.8	0.9	65.8%

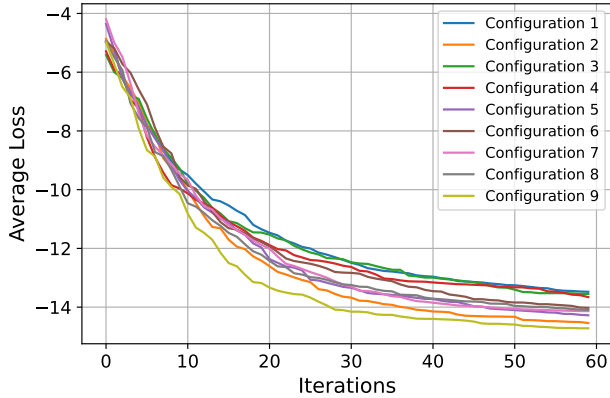


Fig. 8: Average loss per iteration during training for the hyperparameter configurations in Table V. Configuration 9 converges to the lowest value.

sis of this dependency. Similarly, fixing ϕ degraded performance by constraining the optimizer’s ability to distribute reflectors across azimuth subsets. While fixing orientation slightly accelerated convergence, it led to higher loss and lower fooling rates, rendering dimensionality reduction an unfavorable approach to efficiency-efficacy trade-off.

Concluding this subsection on optimization finetuning, using DE with 0.8 mutation probability and 0.9 recombination probability, while allowing the reflector’s orientation to be optimized yields the best attack performance, achieving an average fooling rate of 65.8%. Consequently, all subsequent experiments use this configuration.

C. Evaluating attack performance

Experiments were performed to access the average fooling rate under various conditions.

1) *Visualizations*: Table VII shows the result of optimizing Θ (defined in Equation (2)) for the scene visualized in Fig. 10 specific for the 2S1 class, when four corner reflectors were used. An average fooling rate of 72.9% was achieved on the training set and 71.0% was achieved on the test set.

Fig. 10a to Fig. 10c visualize how one reflector’s brightness waxes and wanes with the azimuth aspect angle. Fig. 10c to Fig. 10d visualize how as one reflector goes out of range, the other reflector takes its place. Fig. 10d to Fig. 10f visualize how for a fixed azimuth aspect angle, the SAR-facing reflector

TABLE VI: Results for different choice of optimization variables.

Configuration	Fixed θ ’s	Fixed ϕ ’s	Average fooling rate
1	False	False	65.8%
2	True	False	57.6%
3	False	True	60.1%
4	True	True	56.8%

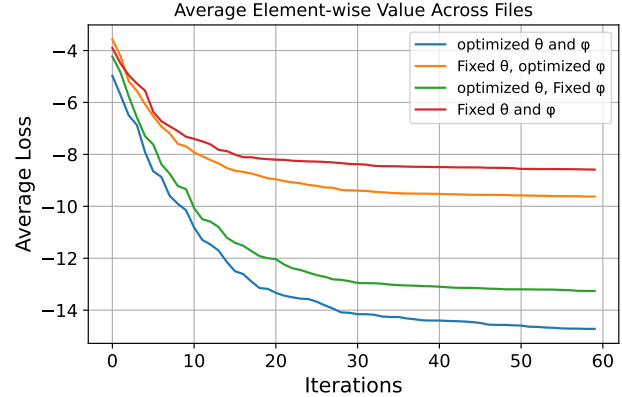


Fig. 9: Average loss per iteration for the different choice of optimization variables in Table VI. Optimizing both θ ’s and ϕ ’s yields the lowest loss, followed by optimizing θ ’s only, optimizing ϕ ’s only, and finally keeping all angles fixed.

remains visible over a range of incidence aspect angles. A supplementary video showing the full viewing sequence is provided online [78]. Taken together, these examples demonstrate how a single perturbation achieves continuous visibility across the full viewing domain, while also highlighting the variations in image appearance as a function of the aspect angles.

2) *Generalizability to other samples*: A perturbation is effective if it generalizes reliably to observations from unseen aspect angles of the same scene. To evaluate this generalizability, perturbations were trained on datasets sampled with varying azimuth spacings and tested on densely sampled sets with a 2.5° azimuth spacing, yielding the average fooling rates shown in Fig. 11. When trained with an azimuth spacing of 10° , the train–test gap is small, indicating good generalization. This is because small changes in azimuth induce smooth, rotation-like variations in SAR image appearance rather than fundamentally new structures. As a result, a robust ATR model, trained to be invariant to such small rotations, remains vulnerable to perturbations across intermediate angles, explaining the high test fooling rates. As the azimuth spacing of the training set increases, the train–test gap widens. Nevertheless, even with very coarse sampling (azimuth spacing of 90°), a nontrivial fooling rate of 52.5% was obtained. Hence, a coarser training set offers a substantial reduction in computation time with only a slight loss in average fooling rate.

3) *Increasing the number of reflectors*: Attack efficacy can be improved by increasing the number of reflectors per perturbation. Perturbations crafted using 8 corner reflectors, so that 2 are visible from any azimuth aspect angle, achieved an average fooling rate of 88.3%. This represents a significant

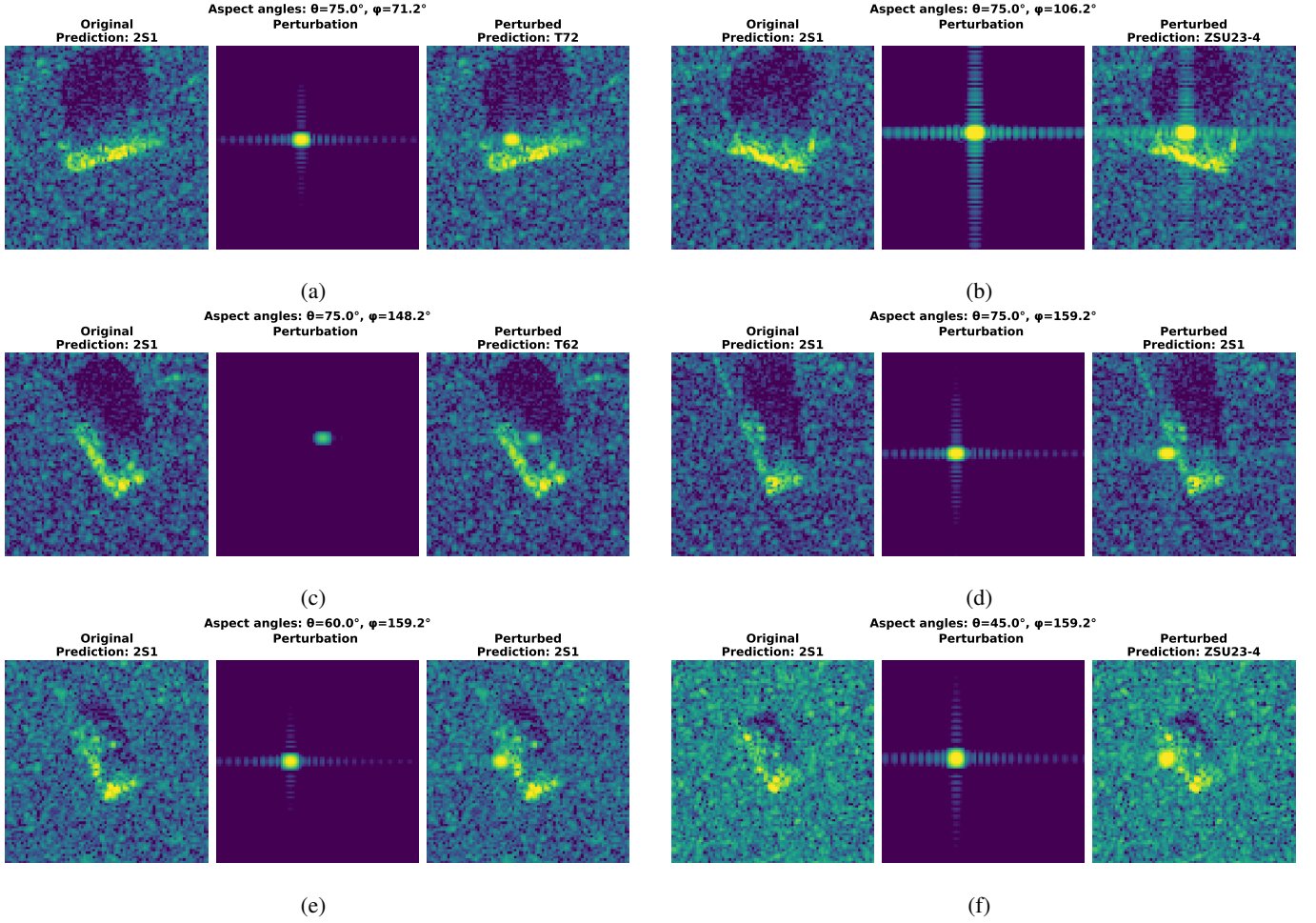


Fig. 10: The original scene, the perturbation parameterized as per Table VII, and the perturbed scene observed from different aspect angles: (a) When $\theta^a = 75^\circ$, $\phi^a = 71.2^\circ$, the second reflector has its boresight at azimuth angle $\phi_2 = 106.3^\circ$, making it visible over the azimuth range $\phi^a \in [61.3^\circ, 151.3^\circ]$. The reflector creates a bright spot, successfully deceiving the target model, while the remaining reflectors are outside their visibility ranges and thus not visible. (b) When $\theta^a = 75^\circ$, $\phi^a = 106.2^\circ$, reflector brightness approaches its peak as the reflector is viewed near its boresight. (c) When $\theta^a = 75^\circ$, $\phi^a = 148.2^\circ$, reflector brightness diminishes as the reflector is observed near the edge of its visibility range. (d) When $\theta^a = 75^\circ$, $\phi^a = 159.2^\circ$ (11° apart from before in azimuth), since each reflector is oriented toward a different quadrant, visibility transitions smoothly from one reflector to the next. Here, the second reflector has dropped out of view while the third has become visible. (e) When $\theta^a = 60^\circ$, $\phi^a = 159.2^\circ$, reflector brightness shows little variation from before. (f) When $\theta^a = 45^\circ$, $\phi^a = 159.2^\circ$, reflector brightness again shows little variation from before.

TABLE VII: Physical properties (position and orientation of the boresight, expressed in incidence angle θ and azimuth angle ϕ) of the adversarial reflectors for a single perturbation.

Reflector	x -position (m)	y -position (m)	θ	ϕ
1	0.31	-2.91	66.3°	16.3°
2	-1.62	-1.80	65.0°	106.3°
3	-1.55	3.18	69.2°	196.3°
4	-0.73	-2.50	75.0°	286.3°

jump compared to an average fooling rate of 65.8% when using 4 reflectors. It is reasonable to expect adding reflectors would improve attack efficacy, at the expense of increased physical complexity, increased cost, and reduced stealth.

4) *Transferability to other models:* Even without full knowledge of the target model, adversarial attacks can succeed, by exploiting the transferability of AEs. In this black-box scenario, the attacker trains perturbations on a surrogate model and applies them to an unknown target model, such that highly transferable perturbations induce misclassifications. To evaluate transferability, the average fooling rates for all surrogate-target model pairs were measured, as summarized in Fig. 12, where diagonal entries correspond to white-box attacks and off-diagonal entries correspond to black-box attacks.

Fig. 12 shows that the perturbations generally transfer well. In some instances, transfer performance even exceeded the white-box baseline, for instance, perturbations trained on AConvNet achieved an average fooling rate of 73.5% when evaluated on AlexNet, likely reflecting differences in model

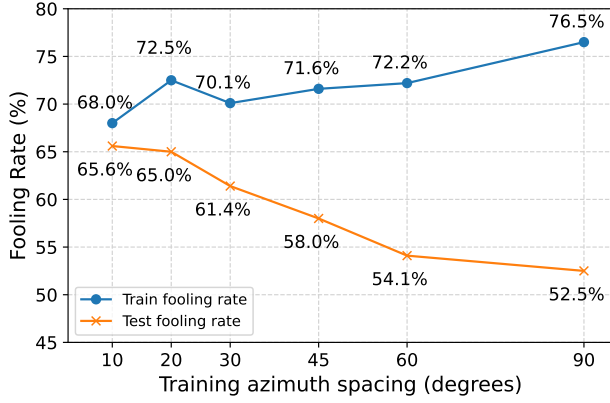


Fig. 11: Average fooling rates achieved on the train and test set using various training azimuth spacing.

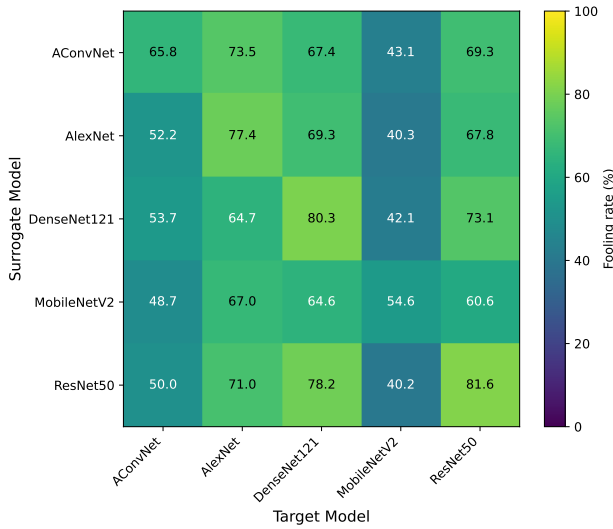


Fig. 12: Average fooling rate when perturbation trained on a surrogate model was tested on a target model.

sensitivity to adversarial attacks. The lowest performance is observed when targeting MobileNetV2, which exhibits reduced average fooling rates in both white- and black-box scenarios. These results highlight the possibility for SAAIPAA to target unknown models. Methods for improving transferability will be investigated in future work.

5) *Partial knowledge of the aspect angles*: So far, the attacker was assumed to have no knowledge of the aspect angles. In a more favorable attacker scenario, where partial information about the aspect angles is available, a more effective attack can be achieved. Suppose the attacker estimates the aspect angles $\hat{\phi}$, $\hat{\theta}$ with a bounded uncertainty Δ , such that:

$$|\hat{\phi} - \phi^a| \leq \Delta, \quad |\hat{\theta} - \theta^a| \leq \Delta. \quad (21)$$

If $\Delta \leq 90^\circ$, a single corner reflector covers the entire potential viewing. The corner reflector's boresight is oriented towards the estimated aspect angles $\phi_1 = \hat{\phi}$, $\theta_1 = \hat{\theta}$. Thus, only its position requires optimization. Training only requires one sample, specifically the sample corresponding to the estimated

aspect angle, substantially reducing the computational cost of training. The perturbation was evaluated over all samples of the class within the angular bounds. For the case $\Delta = 0$, where the attacker had full knowledge of the aspect angles, the AE was evaluated solely on the training sample. This scenario is directly comparable to prior work [9,19]–[21], which only considered fooling rates for fully known aspect angles.

The AEs were trained using DE, using a population size of 40, for 15 iterations. The resulting average fooling rates are summarized in Fig. 13. The average fooling rates were found to increase as the uncertainty decreased, with the case of $\Delta = 0$ yielding rates as high as 99.2%. Even under the largest tested uncertainty of $\Delta = 90^\circ$, corresponding to the full viewing range of the corner reflector, a nontrivial average fooling rate of 47.3% was achieved.

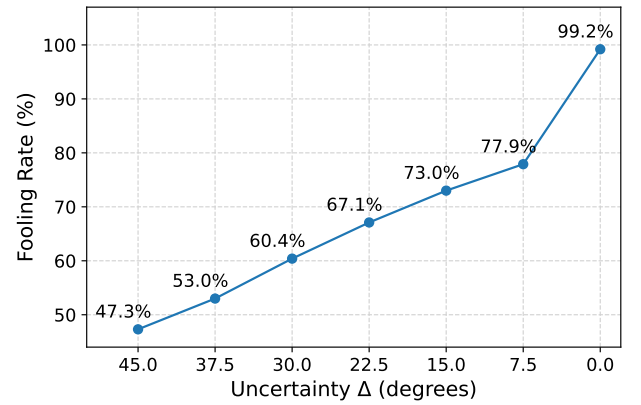


Fig. 13: Average fooling rate per uncertainty Δ .

VII. CONCLUSION

This work is motivated by the confluence of three technological developments: (1) the proliferation of space-based SAR systems due to their all-time all-weather remote sensing capabilities; (2) the maturing application of ML to SAR-based ATR; and (3) the deluge of discoveries in adversarial ML threatening ML applications, including SAR-based ATR systems. The increasing importance of SAR ATR, combined with the increasing potency of attacks against ML applications, motivates research into novel attack mechanisms, and correspondingly defence mechanisms.

In this paper, we propose and evaluate SAAIPAA, a novel physics-driven framework that produces physically realistic, feasible and interpretable perturbations. Unlike prior PAAs, the SAAIPAA is designed for a restrictive attacker model in which the aspect angles are unknown. To accommodate this constraint we formulate a physics-based loss that models the reflector backscatter and propagates the resulting signals through the full SAR imaging chain as a function of the aspect angles. Empirical results demonstrate that SAAIPAA attains high average fooling rates under diverse conditions. When a single corner reflector is visible at any azimuth aspect angle, the attack achieves an average fooling rate of 65.8%, which rises to 88.3% for two reflectors are per azimuth. The attack remains effective even under severe data scarcity

(52.5% fooling rate with a single training sample per reflector). SAAIPAA also exhibits good transferability to most unseen target models. Under a more favorable attacker scenario, where partial information about the aspect angles is available, the average fooling rate further improves, reaching 99.2% in the best-case setting where the aspect angles are fully known. These results demonstrate that SAAIPAA achieves high average fooling rates even under restrictive attacker models.

Future work includes improving the transferability of SAAIPAA, to increase the effectiveness of attacking unknown (black-box) target models. The outcomes will inform our formulation of defence strategies.

ACKNOWLEDGMENT

This material is based up work supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4082. Isar Lemeire is also supported by the Australian Government through the Research Training Program international (RTPi) Scholarships program. The MSTAR dataset was made available by DARPA and AFRL.

REFERENCES

- [1] E. D. Jansing, *Introduction to Synthetic Aperture Radar: Concepts and Practice*. McGraw-Hill Education, 2021. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9781260458961>
- [2] UP42, “Capella space,” Jun. 2025, last updated: 12 Jun 2025, accessed: 24 Jun 2025. [Online]. Available: <https://docs.up42.com/data/datasets/capella-space>
- [3] ICEYE, “ICEYE launches four new satellites and introduces its new Generation 4 satellite,” press release, 2025, published: 15 Mar 2025, accessed: 24 Jun 2025. [Online]. Available: <https://www.iceye.com/newsroom/press-releases/iceye-launches-four-new-satellites-and-introduces-its-new-generation-4-satellite>
- [4] L. A. A. Harrison, *Introduction to Synthetic Aperture Radar Using Python and MATLAB®*. Artech House, 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9893146>
- [5] 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, 2014.
- [6] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, “Adversarial machine learning: A taxonomy and terminology of attacks and mitigations,” NIST Trustworthy and Responsible AI, NIST AI 100-2e2023, Jan. 2024.
- [7] C. Du and L. Zhang, “Adversarial attack for SAR target recognition based on UNet-generative adversarial network,” *Remote Sensing*, vol. 13, no. 21, 2021.
- [8] C. Du, C. Huo, L. Zhang, B. Chen, and Y. Yuan, “Fast C&W: A fast adversarial attack algorithm to fool SAR target recognition with deep convolutional neural networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [9] B. Peng, B. Peng, J. Zhou, J. Xie, and L. Liu, “Scattering model guided adversarial examples for SAR target recognition: Attack and defense,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
- [10] B. Peng, B. Peng, J. Zhou, J. Xia, and L. Liu, “Speckle-variant attack: Toward transferable adversarial attack to SAR target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [11] Y. Chen, J. Du, Y. Yang, and C. Sun, “Positive weighted feature attack: Toward transferable adversarial attack to SAR target recognition,” in *2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI)*, 2023, pp. 93–98.
- [12] M. Du, Y. Sun, B. Sun, Z. Wu, L. Luo, D. Bi, and M. Du, “TAN: A Transferable Adversarial Network for DNN-Based UAV SAR Automatic Target Recognition Models,” *Drones*, vol. 7, no. 3, 2023.
- [13] B. Peng, B. Peng, J. Zhou, X. Huang, L. Meng, and X. Gao, “Low-frequency features optimization for transferability enhancement in radar target adversarial attack,” in *Artificial Neural Networks and Machine Learning – ICANN 2023*. Cham: Springer Nature Switzerland, 2023, pp. 115–129.
- [14] W. Qin, B. Long, and F. Wang, “SCMA: A scattering center model attack on CNN-SAR target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [15] J. Zhou, S. Feng, H. Sun, L. Zhang, and G. Kuang, “Attributed scattering center guided adversarial attack for DCNN SAR target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
- [16] W. Xia, Z. Liu, and Y. Li, “SAR-PeGA: A Generation Method of Adversarial Examples for SAR Image Target Recognition Network,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 2, pp. 1910–1920, 2023.
- [17] Y. Yu, H. Zou, and F. Zhang, “SAR Sticker: An Adversarial Image Patch that can Deceive SAR ATR Deep Model,” in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 7050–7053.
- [18] B. Luo, H. Cao, J. Cui, X. Lv, J. He, H. Li, and C. Peng, “SAR-PATT: A Physical Adversarial Attack for SAR Image Automatic Target Recognition,” *Remote Sensing*, vol. 17, no. 1, 2025.
- [19] J. Xie, B. Peng, Z. Lu, J. Zhou, and B. Peng, “MIGAA: A Physical Adversarial Attack Method against SAR Recognition Models,” in *2024 9th International Conference on Computer and Communication Systems (ICCCS)*, 2024, pp. 309–314.
- [20] F. Zhang, Y. Yu, F. Ma, and Y. Zhou, “A physically realizable adversarial attack method against SAR target recognition model,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 11 943–11 957, 2024.
- [21] Y. Ma, J. Pei, W. Huo, Y. Zhang, Y. Huang, K. Chen, and J. Yang, “SAR-PAA: A Physically Adversarial Attack Approach Against SAR Intelligent Target Recognition,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, no. 2, pp. 1377–1393, 2025.
- [22] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *ICLR Workshop*, 2017.
- [23] H. Wei, H. Tang, X. Jia, Z. Wang, H. Yu, Z. Li, S. Satoh, L. Van Gool, and Z. Wang, “Physical adversarial attack meets computer vision: A decade survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 9797–9817, 2024.
- [24] K. Nguyen, T. Fernando, C. Fookes, and S. Sridharan, “Physical adversarial attacks for surveillance: A survey,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 12, pp. 17 036–17 056, 2024.
- [25] P. Lang, X. Fu, J. Dong, H. Yang, J. Yin, J. Yang, and M. Martorella, “Recent Advances in Deep-Learning-Based SAR Image Target Detection and Recognition,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 6884–6915, 2025.
- [26] P. Zhao, K. Liu, H. Zou, and X. Zhen, “Multi-Stream Convolutional Neural Network for SAR Automatic Target Recognition,” *Remote Sensing*, vol. 10, no. 9, 2018.
- [27] F. Zhou, L. Wang, X. Bai, and Y. Hui, “SAR ATR of Ground Vehicles Based on LM-BN-CNN,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7282–7293, 2018.
- [28] Y. Xie, W. Dai, Z. Hu, Y. Liu, C. Li, and X. Pu, “A novel convolutional neural network architecture for SAR target recognition,” *Journal of Sensors*, vol. 2019, no. 1, p. 1246548, 2019.
- [29] G. Dong and H. Liu, “Global Receptive-Based Neural Network for Target Recognition in SAR Images,” *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1954–1967, 2021.
- [30] W. Wang, C. Zhang, J. Tian, J. Ou, and J. Li, “A SAR Image Target Recognition Approach via Novel SSF-Net Models,” *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, p. 8859172, 2020.
- [31] R. Shang, J. Wang, L. Jiao, R. Stolkin, B. Hou, and Y. Li, “SAR Targets Classification Based on Deep Memory Convolution Neural Networks and Transfer Parameters,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 8, pp. 2834–2846, 2018.
- [32] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou, “Deep convolutional highway unit network for SAR target classification with limited labeled training data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1091–1095, 2017.
- [33] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, “Target classification using the deep convolutional networks for SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, 2016.
- [34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
- [35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *2017 IEEE Conference on*

- Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 2261–2269.
- [36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [38] J. Li, Z. Yu, L. Yu, P. Cheng, J. Chen, and C. Chi, “A comprehensive survey on SAR ATR in deep-learning era,” *Remote Sensing*, vol. 15, no. 5, 2023.
- [39] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
- [40] Gartner, “Gartner experts answer the top generative AI questions for your enterprise,” Gartner Insights, 2024, accessed 22 Jan 2024. [Online]. Available: <https://www.gartner.com/en/topics/generative-ai>
- [41] L. Potter and R. Moses, “Attributed scattering centers for SAR ATR,” *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 79–91, 1997.
- [42] J. Gu, X. Jia, P. de Jorge, W. Yu, X. Liu, A. Ma, Y. Xun, A. Hu, A. Khakzar, Z. Li, X. Cao, and P. Torr, “A survey on transferability of adversarial examples across deep neural networks,” *Transactions on Machine Learning Research*, 2024. [Online]. Available: <https://openreview.net/forum?id=AYJ3m7Bocl>
- [43] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, “Transferable adversarial perturbations,” in *Computer Vision – ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 471–486.
- [44] Y. Dong, T. Pang, H. Su, and J. Zhu, “Evading defenses to transferable adversarial examples by translation-invariant attacks,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4307–4316.
- [45] X. He, Y. Li, H. Qu, and J. Dong, “Improving transferable adversarial attack via feature-momentum,” *Computers & Security*, vol. 128, p. 103135, 2023.
- [46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.
- [47] Y. Akylidiz and R. L. Moses, “Scattering center model for SAR imagery,” in *SAR Image Analysis, Modeling, and Techniques II*, F. Posa, Ed., vol. 3869, International Society for Optics and Photonics. SPIE, 1999, pp. 76 – 85.
- [48] M. Gerry, L. Potter, I. Gupta, and A. Van Der Merwe, “A parametric model for synthetic aperture radar measurements,” *IEEE Trans. Antennas Propag.*, vol. 47, no. 7, pp. 1179–1188, 1999.
- [49] R. Raney, H. Runge, R. Bamler, I. Cumming, and F. Wong, “Precision SAR processing using chirp scaling,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 786–799, 1994.
- [50] S. Amini and A. Profit, “Multi-level fast multipole solution of the scattering problem,” *Engineering Analysis with Boundary Elements*, vol. 27, no. 5, pp. 547–564, 2003.
- [51] D. Munson, J. O’Brien, and W. Jenkins, “A tomographic formulation of spotlight-mode synthetic aperture radar,” *Proceedings of the IEEE*, vol. 71, no. 8, pp. 917–925, 1983.
- [52] S. Auer, R. Bamler, and P. Reinartz, “RaySAR - 3D SAR simulator: Now open source,” in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016, pp. 6730–6733.
- [53] J. Wang, D. Feng, L. Xu, and W. Hu, “Synthetic aperture radar image modulation using phase-switched screen,” *IEEE Antennas Wireless Propag. Lett.*, vol. 17, no. 5, pp. 911–915, 2018.
- [54] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal Adversarial Perturbations,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2017, pp. 86–94.
- [55] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, “Dual attention suppression attack: Generate adversarial camouflage in physical world,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 8565–8574.
- [56] L. Chen, Z. Xu, Q. Li, J. Peng, S. Wang, and H. Li, “An empirical study of adversarial examples on remote sensing image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7419–7433, 2021.
- [57] H. Li, H. Huang, L. Chen, J. Peng, H. Huang, Z. Cui, X. Mei, and G. Wu, “Adversarial examples for CNN-based SAR image classification: An experience study,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1333–1347, 2021.
- [58] T. D. Ross, S. W. Worrell, V. J. Velten, J. C. Mossing, and M. L. Bryant, “Standard SAR ATR evaluation experiments using the MSTAR public release data set,” in *Algorithms for Synthetic Aperture Radar Imagery V*, E. G. Zelnio, Ed., vol. 3370, International Society for Optics and Photonics. SPIE, 1998, pp. 566 – 573, dataset at <https://www.sdms.af.mil/index.php?collection=mstar>.
- [59] E. F. Knott, *Radar cross section measurements*. Springer New York, NY, 2012.
- [60] J. B. Keller, “Geometrical theory of diffraction,” *J. Opt. Soc. Am.*, vol. 52, no. 2, pp. 116–130, Feb. 1962.
- [61] A. C. Polycarpou, C. A. Balanis, and C. R. Birtcher, “Radar cross section of trihedral corner reflectors using PO and MEC,” *Annales Des Télécommunications*, vol. 50, no. 5, pp. 510–516, May 1995.
- [62] C. A. Balanis, *Advanced Engineering Electromagnetics*, 2nd ed. John Wiley & Sons, 2012.
- [63] X. Lin, B. Zhang, F. Wu, C. Wang, Y. Yang, and H. Chen, “SIVED: A SAR Image Dataset for Vehicle Detection Based on Rotatable Bounding Box,” *Remote Sensing*, vol. 15, no. 11, 2023.
- [64] X. Yang, Q. Zhang, Q. Dong, Z. Han, X. Luo, and D. Wei, “Ship instance segmentation based on rotated bounding boxes for SAR images,” *Remote Sensing*, vol. 15, no. 5, 2023.
- [65] S. Perna, A. Natale, C. Esposito, P. Berardino, G. Palmese, and R. Lanari, “Imaging capabilities of an airborne X-band SAR based on the FMCW technology,” in *Multimodal Sensing: Technologies and Applications*, E. Stella, Ed., vol. 11059, International Society for Optics and Photonics. SPIE, 2019, p. 110590G. [Online]. Available: <https://doi.org/10.1117/12.2527924>
- [66] T. J. Walls, M. L. Wilson, D. Madsen, M. Jensen, S. Sullivan, M. Addario, and I. Hally, “Multi-mission, autonomous, synthetic aperture radar,” in *Radar Sensor Technology XVIII*, K. I. Ranney and A. Doerry, Eds., vol. 9077, International Society for Optics and Photonics. SPIE, 2014, p. 907706. [Online]. Available: <https://doi.org/10.1117/12.2053561>
- [67] R. Horn, M. Jaeger, M. Keller, M. Limbach, A. Nottensteiner, M. Pardini, A. Reigber, and R. Scheiber, “F-SAR - recent upgrades and campaign activities,” in *2017 18th International Radar Symposium (IRS)*, 2017, pp. 1–10.
- [68] G. Alberti, L. Citarella, L. Ciofaniello, R. Fusco, G. Galiero, A. Minoliti, A. Moccia, M. Sacchettino, and G. Salzillo, “Current status of the development of an Italian airborne SAR system (MINISAR),” in *SAR Image Analysis, Modeling, and Techniques VI*, F. Posa, Ed., vol. 5236, International Society for Optics and Photonics. SPIE, 2004, pp. 53 – 59. [Online]. Available: <https://doi.org/10.1117/12.512224>
- [69] C. Zhang, P. Benz, A. Karjauv, J. W. Cho, K. Zhang, and I. S. Kweon, “Investigating top-k white-box and transferable black-box attack,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 064–15 073.
- [70] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *ICLR*, 2015.
- [71] A. Dabouei, S. Soleymani, F. Taherkhani, J. Dawson, and N. M. Nasrabadi, “SmoothFool: An efficient framework for computing smooth adversarial perturbations,” in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 2654–2663.
- [72] N. Akhtar, A. Mian, N. Kardan, and M. Shah, “Advances in adversarial attacks and defenses in computer vision: A survey,” *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.
- [73] J. Ding, B. Chen, H. Liu, and M. Huang, “Convolutional neural network with data augmentation for SAR target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, 2016.
- [74] H. Furukawa, “Deep learning for target classification from SAR imagery: Data augmentation and translation invariance,” arXiv preprint arXiv:1708.07920, 2017.
- [75] R. Storn and K. Price, “Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces,” *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, Dec. 1997.
- [76] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, vol. 4, 1995, pp. 1942–1948.
- [77] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf
- [78] I. Lemeire, “SAAIPAA demo,” <https://youtu.be/COq-17vVEps>, accessed: 8 Oct 2025.

APPENDIX

C. Triple reflections

A. Single reflections

$$\begin{cases} N_1^\theta = 0, \\ N_1^\phi = \frac{2A^t}{Z_0} \cos \theta' \iint_{\mathcal{A}_1^i} e^{2jk(x \cos \phi' + y \sin \phi)} \sin \theta' \, dx \, dy. \\ N_2^\theta = 0, \\ N_2^\phi = \frac{2A^t}{Z_0} \sin \theta' \cos \phi' \iint_{\mathcal{A}_2^i} e^{2jk(y \sin \theta' \sin \phi' + z \cos \theta')} \, dy \, dz. \\ N_3^\theta = 0, \\ N_3^\phi = \frac{2A^t}{Z_0} \sin \theta' \sin \phi' \iint_{\mathcal{A}_3^i} e^{2jk(x \sin \theta' \cos \phi' + z \cos \theta')} \, dx \, dz. \end{cases}$$

B. Double reflections

$$\begin{cases} N_{12}^\theta = -\frac{4A^t}{Z_0} \sin(\theta') \cos(\theta') \sin(\phi') \iint_{\mathcal{A}_{12}^i} e^{2jky \sin(\theta') \sin(\phi')} \, dy \, dz, \\ N_{12}^\phi = -\frac{2A^t}{Z_0} \sin(\theta') \cos(\phi') \iint_{\mathcal{A}_{12}^i} e^{2jky \sin(\theta') \sin(\phi')} \, dy \, dz. \\ N_{13}^\theta = \frac{4A^t}{Z_0} \sin(\theta') \cos(\theta') \cos(\phi') \iint_{\mathcal{A}_{13}^i} e^{2jkx \sin(\theta') \cos(\phi')} \, dx \, dz, \\ N_{13}^\phi = -\frac{2A^t}{Z_0} \sin(\theta') \sin(\phi') \iint_{\mathcal{A}_{13}^i} e^{2jkx \sin(\theta') \cos(\phi')} \, dx \, dz. \\ N_{21}^\theta = -\frac{4A^t}{Z_0} \cos(\theta')^2 \sin(\phi') \cos(\phi') \iint_{\mathcal{A}_{21}^i} e^{2jky \sin(\theta') \sin(\phi')} \, dx \, dy, \\ N_{21}^\phi = -\frac{2A^t}{Z_0} \cos(\theta') \cos(2\phi') \iint_{\mathcal{A}_{21}^i} e^{2jky \sin(\theta') \sin(\phi')} \, dx \, dy. \\ N_{23}^\theta = -\frac{4A^t}{Z_0} \sin(\theta') \cos(\theta') \cos(\phi') \iint_{\mathcal{A}_{23}^i} e^{2jkz \cos(\theta')} \, dx \, dz, \\ N_{23}^\phi = -\frac{2A^t}{Z_0} \sin(\theta') \sin(\phi') \iint_{\mathcal{A}_{23}^i} e^{2jkz \cos(\theta')} \, dx \, dz. \end{cases}$$

$$\begin{cases} N_{31}^\theta = \frac{4A^t}{Z_0} \cos(\theta')^2 \sin(\phi') \cos(\phi') \iint_{\mathcal{A}_{31}^i} e^{2jkx \sin(\theta') \cos(\phi')} \, dx \, dy, \\ N_{31}^\phi = \frac{2A^t}{Z_0} \cos(\theta') \cos(2\phi') \iint_{\mathcal{A}_{31}^i} e^{2jkx \sin(\theta') \cos(\phi')} \, dx \, dy. \\ N_{32}^\theta = \frac{4A^t}{Z_0} \sin(\theta') \cos(\theta') \sin(\phi') \iint_{\mathcal{A}_{32}^i} e^{2jkz \cos(\theta') \sin(\phi')} \, dy \, dz, \\ N_{32}^\phi = \frac{2A^t}{Z_0} \sin(\theta') \cos(\phi') \iint_{\mathcal{A}_{32}^i} e^{2jkz \cos(\theta')} \, dy \, dz. \end{cases}$$

$$\begin{cases} N_{123}^\theta = 0, \\ N_{123}^\phi = \frac{2A^t}{Z_0} \sin(\theta') \sin(\phi') \iint_{\mathcal{A}_{123}^i} \, dx \, dz. \\ N_{132}^\theta = 0, \\ N_{132}^\phi = \frac{2A^t}{Z_0} \sin(\theta') \cos(\phi') \iint_{\mathcal{A}_{132}^i} \, dy \, dz. \\ N_{213}^\theta = 0, \\ N_{213}^\phi = \frac{2A^t}{Z_0} \sin(\theta') \sin(\phi') \iint_{\mathcal{A}_{213}^i} \, dx \, dz. \\ N_{231}^\theta = 0, \\ N_{231}^\phi = \frac{2A^t}{Z_0} \cos(\theta') \iint_{\mathcal{A}_{231}^i} \, dx \, dy. \\ N_{312}^\theta = 0, \\ N_{312}^\phi = \frac{2A^t}{Z_0} \sin(\theta') \cos(\phi') \iint_{\mathcal{A}_{312}^i} \, dy \, dz. \\ N_{321}^\theta = 0, \\ N_{321}^\phi = \frac{2A^t}{Z_0} \cos(\theta') \iint_{\mathcal{A}_{321}^i} \, dx \, dy. \end{cases}$$

Isar Lemeire received the B.Sc. and M.Sc. degrees in Computer Science Engineering from Ghent University, Ghent, Belgium in 2021 and 2023 respectively. He is currently pursuing the Ph.D. degree in Computer Science at the University of South Australia, Adelaide, Australia. His research interests include computer vision, adversarial machine learning, and synthetic aperture radar (SAR) imaging.

Will Meakin received the BSoftwEng(Hons) from the University of South Australia in 2017. He is currently pursuing a Ph.D. in Computer Science at the Australian Institute of Machine Learning. His research interests include computer vision and adversarial machine learning.

Yee Wei Law received the B.Eng., M.Eng. and Ph.D. degrees from University of Southampton, Nanyang Technological University, and University of Twente respectively. Before joining UniSA, he was a Research Fellow at the Department of Electrical and Electronic Engineering, The University of Melbourne. He is currently a Senior Lecturer at UniSA, focusing on interdisciplinary research related to cybersecurity, machine learning and space.

Sang-Heon Lee received the B.Eng. degree in Aeronautical Engineering from InHa University, Korea, the M.Eng.Sc. degree in Mechatronics from the University of New South Wales, Australia, and the Ph.D. degree in Systems Engineering from the Australian National University. He is currently a Professor at University of South Australia specializing in machine learning applications. His research interests include engineering management, machine vision systems, hyperspectral image processing, and the application of machine learning and deep learning in agricultural and medical domains. He has co-authored more than 150 papers in international refereed journals and conference proceedings.

Tat-Jun Chin is Professorial Chair of Sentient Satellites at The University of Adelaide. He received his PhD in Computer Systems Engineering from Monash University in 2007, which was partly supported by the Endeavour Australia-Asia Award, and a Bachelor in Mechatronics Engineering from Universiti Teknologi Malaysia in 2004, where he won the Vice Chancellor's Award. Tat-Jun's research interest lies in optimisation for computer vision and machine learning, and their application to intelligent satellites and space robotics. He has published more than 100 research articles on the subject, and has won several awards for his research, including a CVPR award (2015), a BMVC award (2018), Best of ECCV (2018), three DST Awards (2015, 2017, 2021), an IAPR Award (2019) and an RAL Best Paper Award (2021). He was a Finalist in the Academic of the Year Category at Australian Space Awards 2021.