remote sensing

MDPI

Article

Enhancing Transferability with Intra-Class Transformations and Inter-Class Nonlinear Fusion on SAR Images

Xichen Huang [1], Zhengzhi Lu [2] and Bo Peng [2,*]

[1] School of Electronic Science, National University of Defense Technology (NUDT), Changsha 410073, China; huangxichen18@nudt.edu.cn
[2] Test Center, National University of Defense Technology (NUDT), Xi'an 710100, China; luzhengzhi23@nudt.edu.cn
* Correspondence: ppbbo@nudt.edu.cn

Abstract: Recent research has revealed that the deep neural network (DNN)-based synthetic-aperture radar (SAR) automatic target recognition (ATR) techniques are vulnerable to adversarial examples, which poses significant security risks for their deployment in real-world systems. At the same time, the adversarial examples often exhibit transferability across DNN models, whereby when they are generated on the surrogate model they can also attack other target models. As the significant property in black-box scenarios, transferability has been enhanced by various methods, among which input transformations have demonstrated excellent effectiveness. However, we find that existing transformations suffer from limited enhancement of transferability due to the unique imaging mechanism and scattering characteristics of SAR images. To overcome this issue, we propose a novel method called intra-class transformations and inter-class nonlinear fusion attack (ITINFA). It enhances transferability from two perspectives: intra-class single image transformations and inter-class multiple images fusion. The intra-class transformations module utilizes a series of diverse transformations that align with the intrinsic characteristics of SAR images to obtain a more stable gradient update direction and prevent the adversarial examples from overfitting the surrogate model. The inter-class fusion strategy incorporates the information from other categories in a nonlinear manner, effectively enhances the feature fusion effect, and guides the misclassification of adversarial examples. Extensive experiments on the MSTAR dataset and SEN1-2 dataset demonstrate that ITINFA exhibits significantly better transferability than the existing transfer-based methods, with the average transfer attack success rate increases exceeding 8% for single models and over 4% for ensemble models.

Keywords: synthetic-aperture radar; automatic target recognition; adversarial attack; deep neural network; transferability

check for updates

Citation: Huang, X.; Lu, Z.; Peng, B. Enhancing Transferability with Intra-Class Transformations and Inter-Class Nonlinear Fusion on SAR Images. *Remote Sens.* **2024**, *16*, 2539. https://doi.org/10.3390/rs16142539

Academic Editor: Benoit Vozel

Received: 21 May 2024
Revised: 22 June 2024
Accepted: 27 June 2024
Published: 10 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Synthetic-aperture radar (SAR), as a sensor utilizing microwaves for imaging, possesses the capability to conduct all-weather day-and-night imaging without being influenced by factors such as lighting and weather conditions. Through imaging processing algorithms, high-resolution images can be obtained, enabling the accurate extraction of feature information for target recognition [1]. Therefore, SAR holds significant application value in both military and civilian domains [2]. In recent years, with the rapid development of deep learning, deep neural networks (DNNs) have made significant breakthroughs in SAR automatic target recognition (ATR). With the end-to-end training method and powerful feature extraction capability, it has dramatically surpassed the traditional recognition methods in terms of recognition efficiency and accuracy [3–9].

However, the internal working process and decision-making mechanisms of DNN models are not transparent, and their reliability and robustness still need improvement.

Recent research has shown that DNN models are vulnerable to adversarial examples, which are crafted by adding subtle yet carefully designed perturbations to the input images [10,11]. While appearing virtually indistinguishable from the original images to the human eyes, they can cause the DNN models to produce incorrect results. The existence of adversarial examples reveals the vulnerability inherent in DNN models, which poses significant security risks for their deployment and application in radar intelligent recognition systems [12]. Consequently, this issue has garnered widespread attention and has become a hot topic in the SAR-ATR field.

Current research on SAR adversarial attacks primarily focuses on white-box scenarios, where the attacker has full access to the internal information of the target model, including its structures, parameters, and gradients [13–16]. However, in the real world, the opponent's SAR systems and deep recognition models are typically unknown, making it impractical to employ white-box attacks directly. Consequently, it is imperative to conduct attacks in black-box settings, which do not necessitate any information about the target model and align with the SAR non-cooperative scenarios. One of the significant properties in black-box scenarios is the transferability of adversarial examples, whereby the adversarial examples generated on the surrogate model can also attack other models. Nevertheless, black-box attacks are unable to achieve comparable performance to white-box attacks, as the generated adversarial examples tend to overfit the surrogate model and have limited transferability on the target model. For instance, as contrasted in Figure 1, all attack algorithms achieve nearly 100% attack success rates in white-box settings. However, in black-box scenarios, the majority of attack algorithms exhibit transfer attack success rates between 60% and 70%, a significant decrease compared to their white-box performance. Consequently, enhancing the transferability of SAR adversarial examples is a significant and challenging problem.

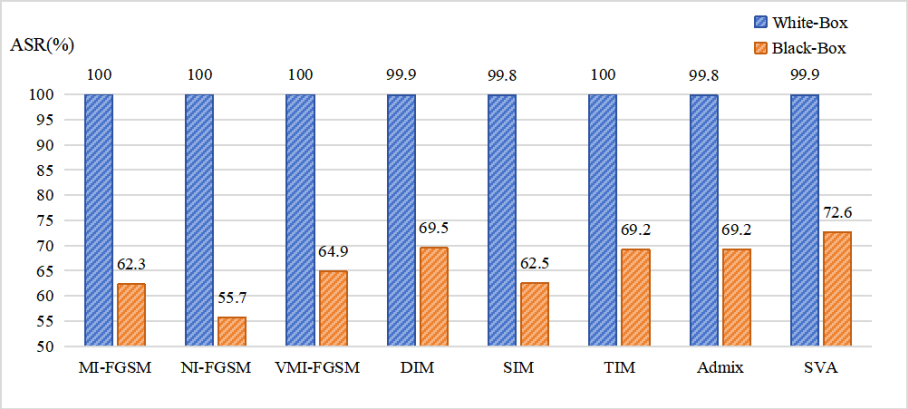


Figure 1. Comparisons of the attack performance of existing algorithms in white-box and black-box scenarios. The data are obtained from single-model experiments with VGG11 utilized as the surrogate model.

Various methods have been proposed to enhance transferability, among which input transformations stand out as a prevailing strategy. The input transformation-based attacks aim to apply a series of transformations to enhance the input diversity and alleviate overfitting to the surrogate model. These transformations include resizing and padding [17], scaling [18], translation [19], and so on. We proceed in this direction and discover that existing transformations typically neglect the unique characteristics of SAR images, such as the multiplicative speckle noise generated by the coherent principle, which results in limited diversity and insufficient transformations for SAR images. Moreover, Admix [20] introduces images from other categories to achieve transformation. However, due to the high contrast and sharp edges between different regions in SAR images, its linear fusion strategy results in insufficient feature fusion and the disruption of original information.

To tackle these problems, we have explored two special strategies via creating diverse input patterns. First, from the perspective of intra-class transformation, we argue that enhancing the input diversity and considering some transformations that align with the characteristics of SAR images can achieve a superior transformation effect and enrich the gradient information, thereby stabilizing the update direction to avoid falling into local optima and alleviating the overfitting to the surrogate model. Second, from the perspective of inter-class transformation, we believe that nonlinear fusion is more suitable for SAR images than a linear mixing strategy, thereby reducing damage to the original features and achieving a better fusion effect.

In this article, we propose a novel method called intra-class transformations and inter-class nonlinear fusion attack (ITINFA) to enhance transferability. At first, we employ a series of transformations on a single image to increase the input diversity. Subsequently, we nonlinearly integrate the information from other categories into the transformed images to guide the adversarial examples to cross the decision boundary. Finally, we utilize the various transformed and mixed images to calculate the average gradient and combine the momentum iterative process to generate more transferable SAR adversarial examples. Extensive experiments on the MSTAR and SEN1-2 datasets demonstrate that ITINFA outperforms the state-of-the-art transfer-based attacks, dramatically enhancing the transferability in black-box scenarios while preserving comparable attack performance in white-box scenarios. The main contributions are summarized as follows:

1. We propose a novel method called ITINFA for generating more transferable adversarial examples to fool the DNN-based SAR image classifiers in black-box scenarios.
2. To improve the limited diversity and insufficient transformations of existing methods commonly utilized for optical images, we fully consider the unique characteristics of SAR imagery and apply diverse intra-class transformations to obtain a more stable update direction, which helps to avoid falling into local optima and alleviates overfitting to the surrogate model.
3. To overcome the insufficient fusion and damage of existing linear mixing, we devise a nonlinear fusion strategy according to the high-contrast characteristic inherent in SAR images, which is beneficial to reduce the background clutter information and avoid excessive disruption of the original SAR image, thereby incorporating information from other categories more effectively.
4. Extensive evaluations based on the MSTAR and SEN1-2 datasets show that the proposed ITINFA significantly improves the transferability of SAR adversarial examples on various cutting-edge DNN models.

The remainder of this article is organized as follows. Section 2 briefly reviews the related works. Section 3 introduces the details of our proposed ITINFA. Section 4 shows experimental results on the MSTAR and SEN1-2 datasets. Finally, we present the discussion and conclusions in Sections 5 and 6.

2. Related Works

In this section, we first briefly introduce adversarial attacks in the image classification domain and the transferability of adversarial examples. Secondly, we provide a review of input transformation-based attacks, the most commonly employed approach to enhance the transferability of adversarial examples. Finally, we review the related works about adversarial attacks in SAR-ATR.

2.1. Adversarial Attacks and Transferability

Supposing x denotes the input image and y^{true} represents the ground-truth label. $\mathcal{F}_\theta$ denotes the classifier with parameters θ which can accurately predict the category of the input image:

$$\mathcal{F}_\theta(x) = y^{true}. \tag{1}$$

Adversarial attacks aim to find a subtle adversarial perturbation δ that can induce misclassification while remaining imperceptible to the human eye, which is expressed in the following mathematical form:

$$\mathcal{F}_\theta(x+\delta) \neq y^{true} \quad \text{s.t.} \|\delta\|_p \leq \epsilon, \tag{2}$$

where $\|\delta\|_p \leq \epsilon$ represents the ℓ_p norm constraint that guarantees the adversarial perturbation δ is visually imperceptible. In this article, we mainly focus on the constraint that $p = \infty$, which is widely adopted in various transfer-based attacks.

FGSM: As the first gradient-based attack, fast gradient sign method (FGSM) [13] calculates the gradient of the loss function and updates the adversarial examples along the direction of the gradient, which is expressed as follows:

$$x^{adv} = x + \epsilon \cdot \text{sign}\big(\nabla_x J(x, y^{true}; \theta)\big), \tag{3}$$

where $\text{sign}(\cdot)$ indicates the sign function and $\nabla_x J(\cdot)$ denotes the gradient of the loss function with respect to the input x. Due to its single-update strategy, FGSM exhibits high efficiency but limited attack performance.

I-FGSM: Building upon the single-step update of FGSM, Kurakin et al. [14] introduced an iterative version named iterative fast gradient sign method (I-FGSM). I-FGSM extends FGSM through multiple iterative update; it is described as follows:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}\Big(\nabla_x J\Big(x_t^{adv}, y^{true}; \theta\Big)\Big), \tag{4}$$

where $x_0^{adv} = x$ and $\alpha = \frac{\epsilon}{n}$, with n being the number of iterations. Research has demonstrated that I-FGSM exhibits more powerful white-box attack performance than FGSM, with poorer transferability due to overfitting the surrogate model.

MI-FGSM: To alleviate the issue of overfitting and improve the transferability, Dong et al. [21] proposed the momentum iterative fast gradient sign method (MI-FGSM). MI-FGSM integrates a momentum term into I-FGSM to stabilize the update direction and prevent it from falling into local optima; it is represented as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J\Big(x_t^{adv}, y^{true}; \theta\Big)}{\big\|\nabla_x J\big(x_t^{adv}, y^{true}; \theta\big)\big\|_1}, \tag{5}$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}\big(g_{t+1}\big), \tag{6}$$

where μ denotes the decay factor and g_t indicates the accumulated gradient information up to the t-th iteration.

NI-FGSM: Lin et al. [18] adopted the Nesterov algorithm to stabilize the gradient update and improve the transferability. The optimization algorithm is called the Nesterov iterative fast gradient sign method (NI-FGSM), which can be expressed as follows:

$$x_t^{nes} = x_t^{adv} + \alpha \cdot \mu \cdot g_t, \tag{7}$$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J\big(x_t^{nes}, y^{true}; \theta\big)}{\|\nabla_x J(x_t^{nes}, y^{true}; \theta)\|_1}. \tag{8}$$

VMI-FGSM: Wang et al. [22] proposed an approach to utilize the gradient variance from previous iterations to stabilize the update direction, and thereby avoid local optima during the search process. Specifically, they approximate the gradient variance by sampling N images in the neighborhood of x, as follows:

$$V(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{x^i} J\Big(x^i, y^{true}; \theta\Big) - \nabla_x J(x, y^{true}; \theta), \tag{9}$$

where the x^i is uniformly sampled in the neighborhood of x. Subsequently, they integrate the gradient variance into the momentum iteration, as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\hat{g}_{t+1} + v_t}{\left\|\hat{g}_{t+1} + v_t\right\|_1}. \tag{10}$$

2.2. Input Transformation-Based Adversarial Attacks

Recent studies [18,21,23] have analogized the optimization process of adversarial examples to the training process of neural networks, comparing the transferability of adversarial examples to the generalization ability of the trained networks. As a result, employing a series of data augmentation techniques to the input images is a standard practice aimed at improving models' generalization ability. Additionally, these strategies have proven effective in enhancing the transferability of adversarial examples.

DIM: As the first input transformation-based attack, DIM [17] employs random resizing and padding on the input images. This involves resizing the images to a random size and subsequently padding them with zeros to revert to their initial size. To strike a balance between the performance of white-box attacks and transferability, a transformation probability p is introduced to determine whether to perform the transformation. The complete process is illustrated as follows:

$$T\left(x_t^{adv}; p\right) = \begin{cases} T\left(x_t^{adv}\right) & \text{with probability } p \\ x_t^{adv} & \text{with probability } 1-p \end{cases}, \tag{11}$$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J\left(T\left(x_t^{adv}; p\right), y^{true}; \theta\right)}{\left\|\nabla_x J\left(T\left(x_t^{adv}; p\right), y^{true}; \theta\right)\right\|_1}, \tag{12}$$

where $T\left(x_t^{adv}; p\right)$ denotes randomly resizing and padding the image with probability p.

SIM: Lin et al. [18] discovered that DNNs exhibit a scale-invariant property, where the outputs remain similar for both the original and scaled versions of images. Leveraging this property, they introduced the scale-invariant method (SIM), which involves scaling images with different factors and averaging the gradients obtained from these scaled images. This process is described as follows:

$$\bar{g}_{t+1} = \frac{1}{m} \sum_{i=0}^{m-1} \nabla_x J\left(\frac{x_t^{adv}}{2^i}, y^{true}; \theta\right), \tag{13}$$

where m is the number of scaled images.

TIM: Dong et al. [19] proposed the translation-invariant method (TIM), which utilizes multiple translated images to calculate the average gradient for optimization. Furthermore, they approximated this process by convolving the gradient with a pre-defined convolutional kernel $\boldsymbol{W}$ to generate adversarial examples, which can be represented as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\boldsymbol{W} * \nabla_x J\left(x_t^{adv}, y^{true}; \theta\right)}{\left\|\boldsymbol{W} * \nabla_x J\left(x_t^{adv}, y^{true}; \theta\right)\right\|_1}. \tag{14}$$

Admix: Zhang et al. [24] proposed an effective data augmentation strategy known as mixup, aimed at enhancing the generalization capabilities of training models. It treats two images equally and mixes them, including their respective labels. This process is expressed as follows:

$$\tilde{x} = \lambda x_i + (1-\lambda) x_j, \qquad \tilde{y} = \lambda y_i + (1-\lambda) y_j, \tag{15}$$

where x_i and x_j are the original images, y_i and y_j are the one-hot label encodings.

Drawing inspiration from this strategy, Wang et al. [20] proposed introducing images from other categories for input transformation. Specifically, they made two key enhance-

ments to the mixing strategy. First, they substituted the equal manner with the master and slave manner, which considered the original image as primary and the introduced images as secondary. Controlling the weights ensured that the introduced images contributed only a small portion. Second, they only mixed the images while retaining the labels, which was able to avoid introducing the gradient of other categories during optimization. The process is expressed as

$$\tilde{x} = \gamma \cdot x + \eta' \cdot x' = \gamma \cdot \left(x + \eta \cdot x'\right), \tag{16}$$

where x is the original image and x' is the image sampled from other categories. And $\eta = \eta'/\gamma$ is employed to ensure that x' only occupies a small portion of admixed image $\tilde{x}$. Subsequently, they calculate the average gradient on a series of admixed images to generate adversarial examples, which can be represented as follows:

$$\overline{g}_{t+1} = \frac{1}{m_1 \cdot m_2} \sum_{x' \in X'} \sum_{i=0}^{m_1-1} \nabla_x J\left(\gamma_i \cdot \left(x_t^{adv} + \eta \cdot x'\right), y^{true}; \theta\right), \tag{17}$$

where m_1 is the number of admixed images controlled by different weight factors γ_i and m_2 is the number of images sampled from other categories.

2.3. Adversarial Attacks in SAR-ATR

In addition to garnering significant attention in computer vision, adversarial attacks have also become a research hotspot in SAR-ATR. Li et al. [25] first pointed out that adversarial examples also existed in SAR images. They empirically analyzed SAR adversarial examples, noting that SAR images containing richer information are more vulnerable to adversarial perturbations. Following this, Huang et al. [26] employed several well-known algorithms to craft adversarial examples, illustrating the vulnerability of SAR target recognition models to such threats. Du et al. [27] designed a generative adversarial network (GAN) to generate SAR adversarial examples that were both efficiently crafted and highly similar to measured SAR images. Subsequently, they further proposed utilizing a U-Net encoding network [28] to replace the traditional optimization process of the C&W algorithm [15], significantly increasing computational speed while maintaining comparable attack performance.

As research in this domain advanced, researchers started to incorporate SAR image characteristics to generate adversarial examples and considered more practical attack scenarios. Peng et al. [29] transformed the input image to the corresponding speckle variant and applied the target mask to confine the perturbation within the target region, aligning with practical physical implementations. Zhang et al. [30] also emphasized that adversarial perturbation confined to the target region would lead to significant performance degradation. Du et al. [31] utilized a generator to generate local universal adversarial perturbations on target regions, achieving comparable performance to global adversarial perturbations. In a fully black-box scenario, where training samples were unavailable, Peng et al. [32] devised a framework for generating universal adversarial perturbations. Experiments on the MSTAR and SARSIM datasets revealed significant effectiveness of their framework for SAR image classification models. Zhang et al. [33] exploited the characteristics of SAR image data to design several objective functions, achieving better performance compared to other baseline approaches. Lin et al. [34] proposed a targeted attack method that extracted the feature maps at shallow layers and constructed the feature-level loss function to generate adversarial examples. Although these methods consider various scenarios in the real world, there is still a long way to go before physically implementing the adversarial perturbations in the signal domain. Furthermore, the transferability remains severely limited when confronted with unknown target models. Therefore, enhancing the transferability of SAR adversarial examples poses a highly challenging task, and this article mainly focuses on developing an algorithm to generate more transferable adversarial examples in black-box settings.

3. Methods

In this section, we provide a detailed description of the proposed ITINFA. Firstly, we elaborate on several intra-class input transformation methods that align with the characteristics of SAR images. They are employed on the single input image to generate diverse transformed images and obtain rich gradient information, thereby alleviating overfitting to the surrogate model. Subsequently, an inter-class nonlinear fusion strategy is employed to incorporate information from other categories into the transformed images, which is beneficial to guide adversarial examples to cross the decision boundary to induce misclassification. Finally, we combine the momentum iterative algorithm with the intra-class transformations and inter-class nonlinear fusion strategies to further enhance the transferability of adversarial examples. The framework of the proposed ITINFA is shown in Figure 2.

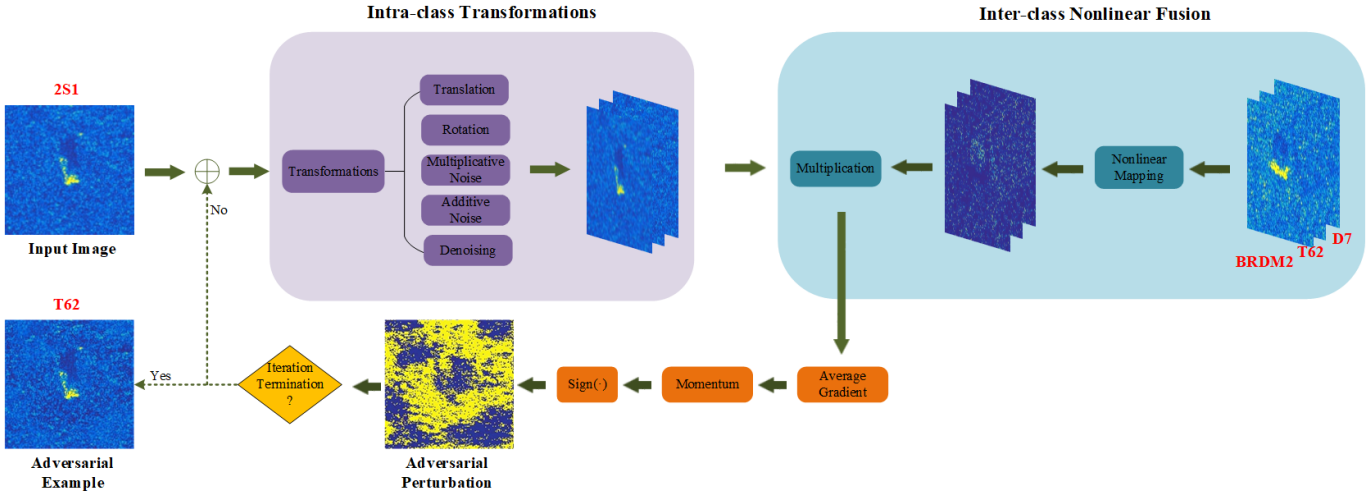


Figure 2. The framework of the proposed ITINFA. For enhanced visual presentation, the Parula color map is utilized to visualize the gray-scale SAR images in this article.

3.1. Intra-Class Transformations

Input transformation-based methods [17–20] have significantly improved the transferability of adversarial examples. They regard generating adversarial examples as the network training process and utilize data augmentation, which is widely employed to improve the generalization ability of the network, to enhance the transferability [23]. However, due to the substantial disparities between SAR images and optical images in characteristics, spectral bands, and imaging mechanisms, these input transformation-based methods still exhibit limited improvement of transferability in SAR adversarial examples. Therefore, it is necessary to consider the transformations that align with the characteristics of SAR images, such as the speckle transformation method proposed in [29], which considers the inherent speckle noise in SAR images and alleviates overfitting to the surrogate model. Moreover, Wang et al. [35] proposed and empirically validated that more diverse transformed images result in better transferability. Motivated by this inspiration, we employ a series of input transformations on a single input image to obtain more diverse transformed images. The employed input transformations are shown in Figure 3.

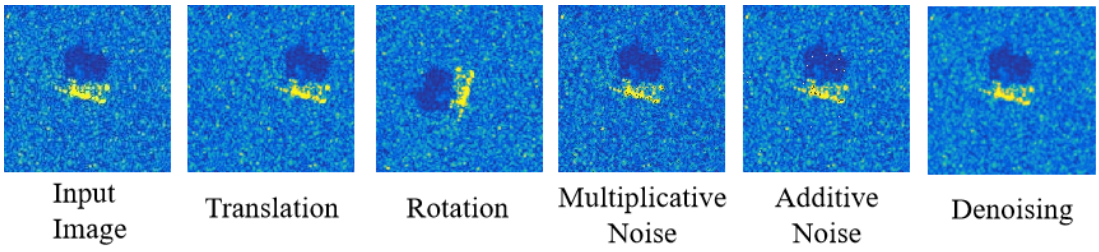


Figure 3. Examples of the input image and transformed images.

3.1.1. Translation

Because convolutional neural networks inherently exhibit a certain level of small-scale translation invariance [36], we consider translation to simulate the positional deviation. Translation can be mathematically expressed as follows:

$$x'_{ij} = x_{i+u,j+v}, \tag{18}$$

where x is the input image of size $m \times n$, (i, j) is the original coordinate and (u, v) is the shift coordinate. In this article, we select a random step for translation ranging from 10 to 30 pixels in the range and azimuth directions, which can guarantee that the target does not exceed the image boundary.

3.1.2. Rotation

As a typical geometric data augmentation technique, rotation is widely employed to transform the input images. In this article, we also utilize rotation to simulate the ideal perspective variation in SAR images, generating more diverse input images. Specifically, we randomly select one angle from 90°, 180°, and 270° to perform the input transformation.

3.1.3. Multiplicative Speckle Noise

Due to the physical scattering concepts of coherent radiation of SAR, multiplicative speckle noise is the unique noise inherent in SAR images [37]. According to the multiplicative noise model, the observed intensity at each resolution cell of the SAR images is formulated as the radar cross-section (RCS) modulating an exponentially distributed multiplicative noise, which can be modeled as follows [36]:

$$I = \sigma \times n, \tag{19}$$

$$p_n(n) = \frac{\mathrm{e}^{-n}}{1 - \mathrm{e}^{-a}}, \tag{20}$$

where I represents the observed intensity and σ denotes the estimated RCS of the illuminated resolution cell. n denotes the multiplicative speckle noise, which follows a truncated exponential distribution with parameter a.

3.1.4. Additive Noise

Most imaging sensors are dominated by additive noise. In addition to the multiplicative noise model, we also consider the additive noise model to diversify the input images further. Specifically, we consider three types of noise in this article: Gaussian noise, salt-and-pepper noise, and uniform noise, and randomly select one type of noise to add to the original SAR images for input transformations.

3.1.5. Denoising

SAR images are often affected by significant amounts of noise, such as Gaussian noise and speckle noise [38]. Therefore, denoising is necessary to enhance the quality of SAR images, enabling better application in downstream tasks such as target detection and

recognition. In this article, we randomly select one of the following denoising methods for input transformations: SAR-block-matching 3D (SAR-BM3D) and discrete cosine transform (DCT). SAR-BM3D [39] is one of the most promising nonlocal despeckling algorithms, achieving a satisfactory balance between speckle reduction and detail preservation. Its basic idea follows the BM3D algorithm [40] for additive white Gaussian noise denoising, but revises the processing steps and considers the unique characteristics of SAR images. DCT [35] is a transformation method that transforms the input image to the frequency domain, eliminates high-frequency components from the image, and restores the image to the spatial domain with the inverse DCT.

3.2. Inter-Class Nonlinear Fusion

Adversarial attacks seek to mislead DNNs, resulting in the misclassification of adversarial examples into wrong categories. Mixing images from other categories during the input transformation process will aid in guiding adversarial examples to cross the decision boundary towards other categories, thereby achieving misclassification. However, existing research has indicated that Admix has limitations due to its linear mixing strategy [41], which incorporates the feature information of the original image insufficiently and without adaptation. At the same time, as sharp edges exist between different regions in SAR images, it may disrupt certain regions in the original SAR image, leading to a suboptimal mixing effect. For instance, when a region in the original image has relatively small pixel values, and the corresponding pixel values in the introduced image are significantly large, even though the weighting coefficient is small, the linear addition may still result in substantial changes in that region and disrupt the feature information of the original image. Furthermore, linear addition may also incorporate background clutter information excessively, which undermines the mixing effect and hinders the misclassification of adversarial examples.

To overcome these problems, we utilize a nonlinear fusion strategy for a better mixing effect. First, we notice that the radar cross-section (RCS) of the target is relatively large, corresponding to regions with higher pixel values in SAR imagery. Therefore, we map the introduced images to a nonlinear space in the form of an exponential function. This exponential mapping expands the information within higher-pixel-value regions while compressing information within lower-pixel-value regions, thereby effectively concentrating the primary information within the target region and mitigating the influence of background clutter. Moreover, we employ multiplication modulation based on the original image, which can alleviate the distortions to the original image features when there exists a substantial difference between the introduced image and the original image within certain regions, thereby integrating information from other categories in a smooth fashion. The complete process is expressed as follows:

$$\tilde{x} = x \odot e^{\frac{x^*}{r}}, \tag{21}$$

where x is the original image and x^* is the randomly sampled image from another category. $\odot$ denotes the element-wise product. r represents the parameter of the exponential function, which is used to modulate the strength of the nonlinear mapping.

Through the nonlinear multiplication strategy, we conduct information fusion more efficiently because it mitigates the disruptive effect on the original image and reduces the background clutter information in the introduced image. Since it is based on pixel-wise multiplication of the original image, the mixed image can preserve the information of the original image more adaptively and avoid excessive disruption while incorporating information from other categories to induce misclassification.

3.3. Intra-Class Transformations and Inter-Class Nonlinear Fusion Attack

Combining the above intra-class transformation and inter-class fusion strategies, we propose a novel input transformation-based attack, namely, intra-class transformations and inter-class nonlinear fusion attack (ITINFA). Firstly, we apply a series of image trans-

formations $T_i(\cdot)$ randomly selected in Section 3.1 to obtain diverse input and prevent the adversarial examples from overfitting the surrogate model. Secondly, we randomly sample several images from other categories and mix them up with the transformed images $T_i(x)$ in a nonlinear manner, which is elaborated on in Section 3.2. The nonlinear multiplication is beneficial to guide adversarial examples to be better classified into other incorrect categories while preserving the information of the original category without significant damage. After the intra-class input transformations and inter-class nonlinear fusion, we calculate the average gradient over the transformed and mixed copies of the input for updates, which can be expressed as

$$\overline{g}_{t+1} = \frac{1}{m_1 * m_2} \sum_{i=0}^{m_1-1} \sum_{j=0}^{m_2-1} \nabla_x J\left(T_i(x) \odot e^{\frac{x_j^*}{r}}, y^{true}, \theta\right), \tag{22}$$

where $T_i(\cdot)$ denotes the input transformations and m_1 is the number of transformations. x_j^* represents the image randomly sampled from another category and m_2 is the number of these sampled images.

Finally, we incorporate momentum for iterative updates to prevent becoming stuck in local optima, thereby boosting transferability more effectively. The pseudo-code of ITINFA is shown in Algorithm 1.

Algorithm 1 Intra-class Transformations and Inter-class Nonlinear Fusion Attack

Input: A surrogate model $\mathcal{F}_\theta$; The input image x with ground-truth label y^{true}; The maximum perturbation budget ϵ; The number of iterations N; The decay factor μ; The number of transformations m_1; The number of images from other categories m_2

Output: An adversarial example x^{adv}

$\alpha = \epsilon/N; g_0 = 0; \overline{g}_0 = 0; x_0^{adv} = x$
for $t = 0$ to $N-1$ **do**
  Construct a set of m_1 randomly transformed images according to Section 3.1;
  Mix up each transformed image with m_2 images from other categories according to Section 3.2;
  Calculate the average gradient according to Equation (22)
  Update the enhanced momentum: $g_{t+1} = \mu \cdot g_t + \frac{\overline{g}_{t+1}}{\|\overline{g}_{t+1}\|_1}$
  Update the adversarial example: $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})$
end for
return $x^{adv} = x_N^{adv}$

4. Experiments

4.1. Datasets

4.1.1. MSTAR Dataset

The MSTAR dataset [42] is a benchmark dataset for SAR-ATR, which contains a series of X-band SAR images with ten different classes of ground targets at different azimuth and elevation angles. The optical images and their corresponding SAR images are depicted in Figure 4. In the experiments, we select the standard operation condition (SOC) that contains 2747 images collected at a 17° depression angle as the training dataset and 2426 images collected at a 15° depression angle as the test dataset. The detailed information of each category is shown in Table 1.

Table 1. Details of the MSTAR dataset and SEN1-2 dataset used in this article.

MSTAR			SEN1-2		
Class	**Training Set**	**Test Set**	**Class**	**Training Set**	**Test Set**
2S1	299	274	s1_18	500	300
BMP2	233	196	s1_19	500	300
BRDM2	298	274	s1_20	500	300
BTR60	256	195	s1_32	500	300
BTR70	233	196	s1_38	500	300
D7	299	274	s1_45	500	300
T62	299	273	s1_49	500	300
T72	232	196	s1_59	500	300
ZIL131	299	274	s1_61	500	300
ZSU234	299	274	s1_78	500	300
Total	2747	2426	Total	5000	3000

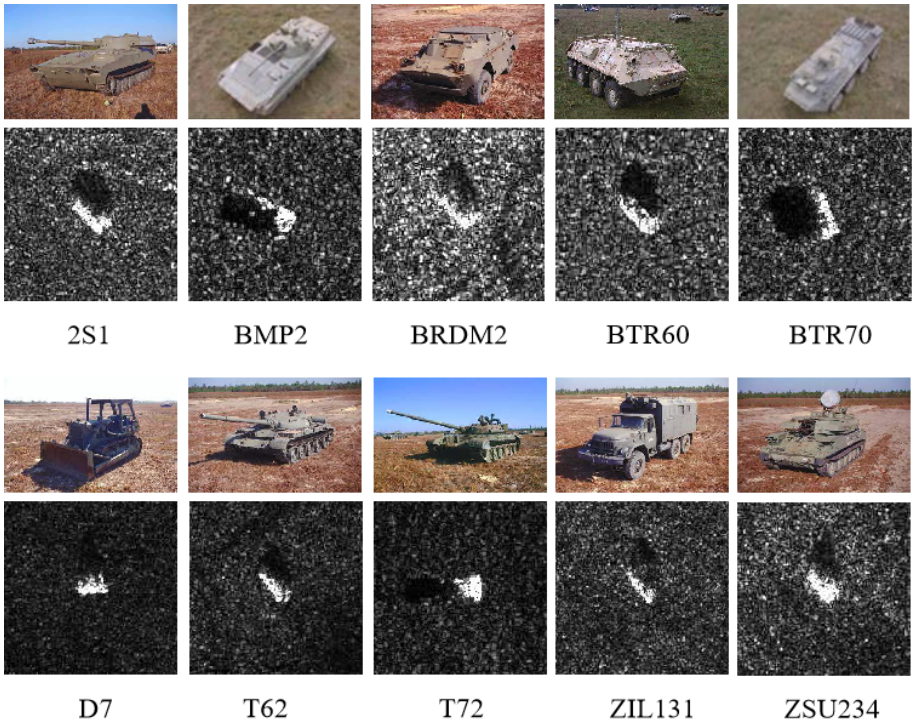


Figure 4. Examples in the MSTAR dataset: (**top**) optical images and (**bottom**) the corresponding SAR images.

4.1.2. SEN1-2 Dataset

The SEN1-2 dataset [43] is an SAR–optical image pair dataset. It comprises four subsets, each corresponding to a different season—spring, summer, autumn, and winter—and contains 282,384 pairs of corresponding image patches. In this article, following the selection strategy adopted in [25], we select the summer subset, which consists of two parts: s1 and s2. The s1 part contains SAR images, while the s2 part contains corresponding optical images. There are 49 folders within s1, each containing SAR images from the same area and representing the same category. After careful consideration, we selected ten distinct categories for our experiments, striving to ensure maximum difference among different categories while maintaining similarity within each category, thus supporting the tasks of target recognition and adversarial attack. Within each category, we selected

500 images for training and 300 images for testing, as summarized in Table 1. Figure 5 illustrates the examples of the SAR images alongside their corresponding optical images.

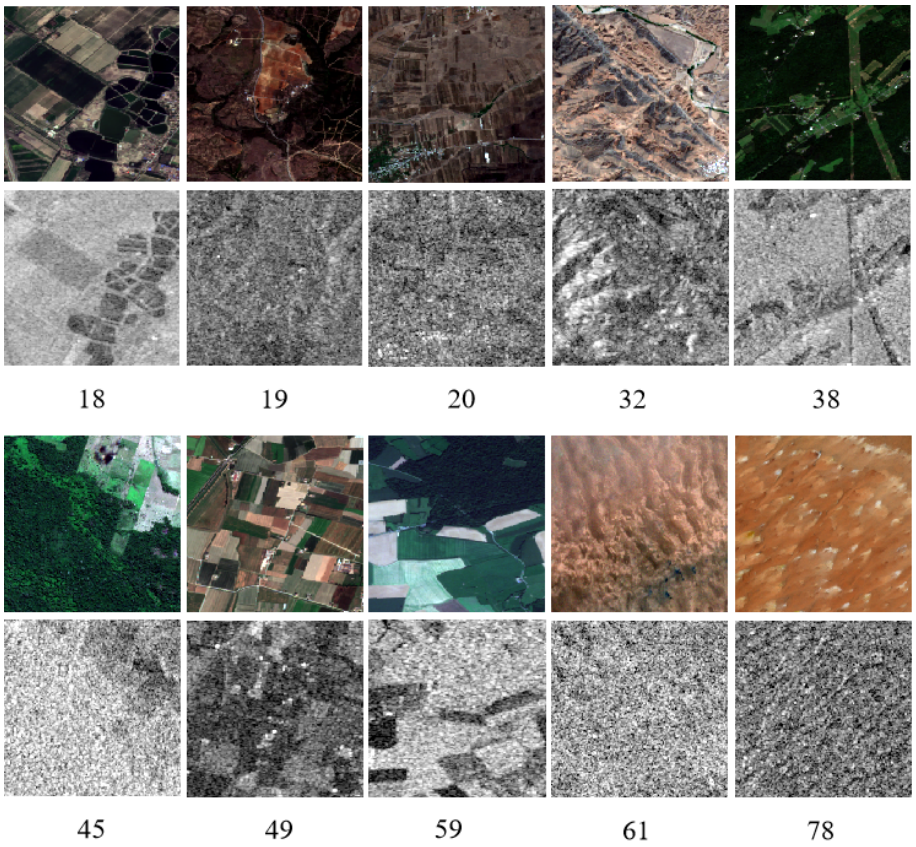


Figure 5. Examples in the SEN1-2 dataset: (**top**) optical images and (**bottom**) the corresponding SAR images.

4.2. Experimental Setup

4.2.1. Models

We conduct the experiments on eight classical DNN models. Among them, AlexNet [44], VGG11 [45], ResNet50 [46], ResNeXt50 [47], and DenseNet121 [48] are widely utilized as feature extraction networks and have achieved excellent target recognition performance. In addition, SqueezeNet [49], ShuffleNetV2 [50], and MobileNetV2 [51] are lightweight DNN architectures that reduce the model parameters and sizes while achieving comparable performance in various target recognition tasks. All of these models are well trained on the MSTAR and SEN1-2 datasets to evaluate the transferability of adversarial examples, and the specific information and recognition accuracies are shown in Table 2.

During the training process, we adopt the preprocessing strategy proposed in [3]. At first, the single-channel gray-scale SAR images are normalized to [0, 1] to accelerate the convergence of the loss function. Subsequently, we randomly crop 88 $\times$ 88 patches of training images for data augmentation and centrally crop 88 $\times$ 88 patches of test images to evaluate the recognition performance. After that, we resize these images to 224 $\times$ 224 and input them into the DNN models. Furthermore, the cross-entropy loss function and the Adam optimizer [52] are employed to train the models, with a learning rate of 0.0001, batch size of 20, and 500 training epochs.

Table 2. The information and recognition accuracies of the DNN models.

Model	# params.	FLOPs ($\times 10^9$)	MSTAR Acc. (%)	SEN1-2 Acc. (%)
AlexNet [44]	57,029,258	0.66	97.57	93.37
VGG11 [45]	128,806,154	7.55	98.69	95.21
ResNet50 [46]	23,522,314	4.05	97.89	96.79
ResNeXt50 [47]	22,994,186	4.21	98.01	96.43
DenseNet121 [48]	6,957,898	2.82	98.31	98.03
SqueezeNet [49]	726,474	0.25	96.66	95.29
ShuffleNetV2 [50]	1,263,446	0.14	97.95	94.15
MobileNetV2 [51]	2,236,138	0.31	96.93	96.73

4.2.2. Baselines

In the experiments, we compare our proposed ITINFA with several classic transfer-based attack algorithms, namely, MI-FGSM [21], NI-FGSM [18], and VMI-FGSM [22]. Additionally, we also compare with several input transformation-based methods, namely, DIM [17], SIM [18], TIM [19], Admix [20], and SVA [29]. The maximum perturbation budget ϵ for all these algorithms is set to 16/255, with the number of iterations T set to 10, step size α set to 16/2550, and decay factor μ set to 1. We set the number of sampled examples in the neighborhood to 5 and the upper bound to 1.5 for VMI-FGSM. We adopt transformation probability $p = 0.5$ for DIM and the number of scaled copies $m = 30$ for SIM. For TIM, we adopt a Gaussian kernel with a size of 5×5. For Admix, we randomly sample $m = 3$ images from other categories with the strength of 0.2 and scale 30 images for each admixed image. For SVA, we adopt a median filter with a kernel size of 5×5 and tail $\beta = 1.5$ of the truncated exponential distribution. We set $m_1 = 30$, $m_2 = 3$, and $r = 0.8$ for our proposed ITINFA. For fair comparisons, all of these methods are integrated into momentum iteration.

4.2.3. Metrics

The transferability is evaluated by the attack success rate, which is defined as the ratio of the number of examples misclassified after the attack to the total number of correctly classified examples. The formula is expressed as follows:

$$ASR = \frac{\sum_{i=1}^{N} \mathbb{I}\left(\operatorname{argmax} \mathcal{F}_\theta\left(x_i^{adv}\right) \neq y^{true}\right)}{N}, \tag{23}$$

where $\mathbb{I}(\cdot)$ is the indication function and N is the total number of correctly classified examples.

In addition, the misclassification confidence [33] is utilized to further evaluate the attack performance and transferability of various attack algorithms. This is obtained from the softmax output of the target model and denotes the confidence probability that the adversarial example is misclassified into the wrong category in a successful attack.

Adversarial attacks require not only deceiving the DNNs but also ensuring that the generated adversarial perturbations are sufficiently subtle and imperceptible to the human eye. In the experiments, the ℓ_2 norm and structural similarity (SSIM) are employed to evaluate the imperceptibility of adversarial examples. The ℓ_2 norm [53] is a widely employed metric that measures the Euclidean distance between the adversarial example and the original image, while the SSIM [54] is utilized to measure the similarity of these two images.

Finally, we calculate the average time consumption for generating a single adversarial example, serving as a metric to assess the computational efficiency of the attack algorithms.

4.2.4. Environment

In this article, the Python programming language (v3.11.8) and the Pytorch deep learning framework (v2.2.2) [55] are used to implement all the experiments and evaluate the performance of various adversarial attack algorithms. All the experiments are supported by four NVIDIA RTX 6000 Ada Generation GPUs and powered by an Intel Xeon Silver 4310 CPU.

4.3. Single-Model Experiments

We first conduct experiments on a single model to evaluate the transferability of various attack algorithms. The adversarial examples are crafted on a single surrogate model and subsequently employed to attack the target models. If the surrogate model is identical to the target model, it denotes a white-box attack scenario; otherwise, it is classified as a black-box attack. Figure 6 summarizes the attack success rates of adversarial examples generated by each algorithm when the surrogate models are AlexNet, VGG, ResNet, ResNeXt, DenseNet, SqueezeNet, ShuffleNet, and MobileNet, respectively.

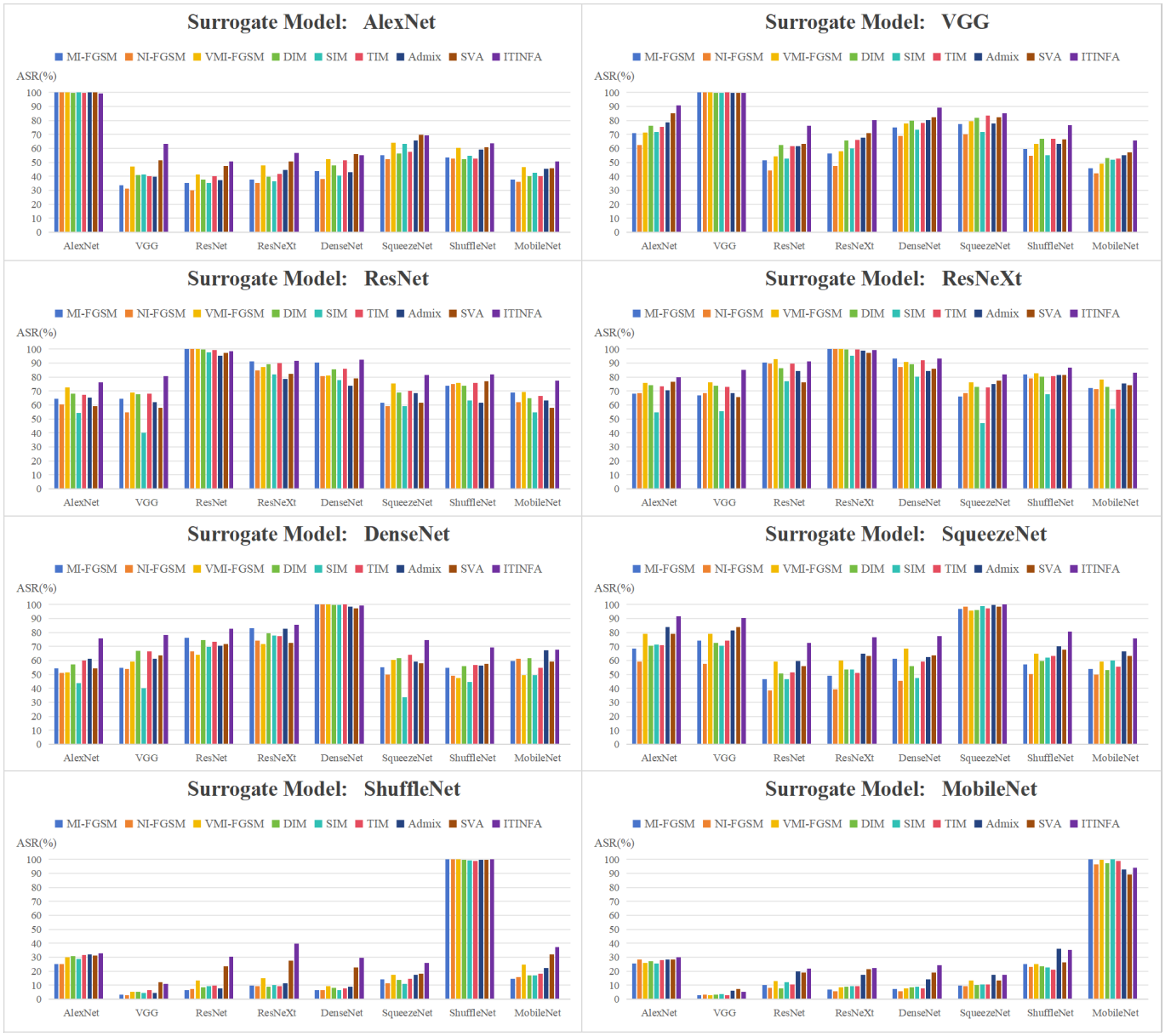


Figure 6. The attack success rates of eight models on the adversarial examples generated on each surrogate model.

It is evident that the attack success rates of white-box attacks for all the surrogate models notably surpass those of black-box attacks. Furthermore, for the AlexNet, VGG, and ShuffleNet models, the white-box attack success rates for all the employed algorithms nearly reach 100%. As for the black-box attack performance, it is discernible that diverse surrogate models exhibit significant discrepancies in the transferability of the generated adversarial examples. For instance, when ResNet and ResNeXt are employed as surrogate models,

they achieve notable transferability among different target models, with the attack success rates consistently surpassing 50%. However, when ShuffleNet and MobileNet are utilized as surrogate models, the transferability is notably limited, with attack success rates scarcely exceeding 20%. Different attack algorithms also exhibit discrepancies in transferability across different target models. In comparison, NI-FGSM generally demonstrates inferior transferability on most models compared to MI-FGSM, which may suggest that Nesterov's accelerated gradient is less effective than momentum on SAR images. Additionally, it can be observed that among several input transformation-based methods, SIM shows the poorest transferability on most target models. This indicates that the scaling transformation may not be suitable for SAR images with sparse characteristics. At the same time, both DIM and TIM exhibit similar transferability on most models, suggesting that the contributions of resizing and translation in improving transferability are comparable for SAR images. Compared to these baselines, our proposed ITINFA consistently achieves the best transferability across eight surrogate models. This consistent and superior attack performance indicates that ITINFA can effectively enhance transferability across various target models. At the same time, it is worth noting that although ITINFA exhibits a slightly lower white-box attack success rate on specific models (e.g., AlexNet, ResNet, MobileNet) compared to some baseline methods, it still ensures a satisfactory level of white-box attack effectiveness, and the slight reduction in white-box attack performance as a trade-off for enhanced transferability is a normal phenomenon.

To better present the experimental results of transferability, Tables 3 and 4, respectively, demonstrate the average transfer attack success rates on the MSTAR and SEN1-2 datasets when each model serves as the surrogate model and generates adversarial examples to attack the other seven target models. The experimental results indicate that regardless of which model acts as the surrogate model, our proposed ITINFA consistently achieves the best transferability. Specifically, ITINFA outperforms the best baseline method with a clear margin of 9.0% on the MSTAR dataset and 8.1% on the SEN1-2 dataset, further underscoring that ITINFA can significantly enhance the transferability across different architectures.

Table 3. The transfer attack success rates on the MSTAR dataset. Each datum represents the average attack success rate obtained using the surrogate model specified in that column to attack the other seven target models.

Attack	AlexNet	VGG	ResNet	ResNeXt	DenseNet	SqueezeNet	ShuffleNet	MobileNet	Average
MI-FGSM	42.4	62.3	73.5	77.1	62.6	59.2	11.4	12.5	50.1
NI-FGSM	39.3	55.7	68.1	76.1	57.9	49.1	11.3	11.9	46.2
VMI-FGSM	51.4	64.9	75.9	81.9	57.7	67.2	16.5	13.7	53.7
DIM	45.1	69.5	74.1	78.6	65.4	59.5	13.2	12.7	52.3
SIM	44.9	62.5	61.7	63.9	51.5	58.9	12.5	13.3	46.2
TIM	46.3	69.2	74.9	78.9	64.7	60.8	13.9	12.8	52.7
Admix	47.8	69.2	67.4	77.2	65.5	69.9	14.9	19.8	54.0
SVA	54.5	72.6	67.9	76.8	62.5	68.1	23.7	19.3	55.7
ITINFA	58.6	80.6	83.2	85.9	76.3	81.0	29.5	22.3	64.7

In addition, we also calculate the average misclassification confidence, ℓ_2 norm, SSIM, and time consumption on the MSTAR dataset to further conduct a comprehensive evaluation of the proposed algorithm. VGG is utilized as the surrogate model to attack the other seven target models, and the experimental results are reported in Table 5. It can be observed that ITINFA not only achieves the highest attack success rate but also attains the highest misclassification confidence, indicating that ITINFA is more effective than other baselines. At the same time, ITINFA demonstrates comparable ℓ_2 norm and SSIM values to other algorithms. This indicates that ITINFA maintains a similar level of imperceptibility as

baseline methods, rendering the adversarial perturbations imperceptible to the human eye, especially when dealing with a limited perturbation budget. Moreover, both ITINFA and Admix require a longer execution time compared to other methods, indicating that the incorporation of additional image categories to guide the optimization of adversarial examples is more time-consuming than other transformation approaches. Although ITINFA takes a relatively longer time compared to other algorithms, it remains within an acceptable range, ensuring an average time consumption of less than 1 s for generating a single adversarial example. It is also worth mentioning that the computational efficiency of ITINFA is variable and positively correlates with the number of transformations and introduced samples. Reducing the number of transformations or introduced samples can significantly enhance computational efficiency at the cost of a slight decrease in attack success rate.

Table 4. The transfer attack success rates on the SEN1-2 dataset. Each datum represents the average attack success rate obtained using the surrogate model specified in that column to attack the other seven target models.

Attack	AlexNet	VGG	ResNet	ResNeXt	DenseNet	SqueezeNet	ShuffleNet	MobileNet	Average
MI-FGSM	37.5	60.8	45.9	54.4	72.1	45.1	31.1	35.2	47.8
NI-FGSM	39.8	59.7	51.8	55.6	78.5	42.1	34.7	39.1	50.2
VMI-FGSM	49.1	53.1	60.4	64.3	82.3	51.9	38.4	43.1	55.3
DIM	40.9	66.5	55.1	63.8	80.4	56.7	33.7	40.3	54.7
SIM	37.2	53.9	63.5	55.1	74.5	54.6	36.5	36.6	51.5
TIM	40.9	67.7	56.4	65.4	80.3	57.5	34.4	41.3	55.5
Admix	42.2	59.1	61.5	59.9	71.2	45.7	33.5	40.2	51.7
SVA	40.3	67.3	57.1	61.1	73.8	47.9	36.4	38.9	52.9
ITINFA	59.3	67.9	65.7	71.5	83.7	56.8	58.3	45.4	63.6

Table 5. The average value of misclassification confidence, ℓ_2 norm, SSIM, and time consumption.

Attack	Confidence (%)	ℓ_2 Norm	SSIM	Time Consumption (ms)
MI-FGSM	95.2	5.183	0.727	10.2
NI-FGSM	94.7	5.060	0.685	9.5
VMI-FGSM	94.9	5.115	0.733	39.2
DIM	92.1	5.037	0.701	10.7
SIM	96.0	5.257	0.731	68.3
TIM	93.7	5.022	0.703	9.8
Admix	93.6	5.289	0.707	438.9
SVA	96.5	5.067	0.725	33.5
ITINFA	96.9	5.073	0.726	525.1

4.4. Ensemble-Model Experiments

To further validate the effectiveness of our proposed ITINFA, we conduct the ensemble-model experiments proposed in [21], which fuse the logit outputs of various models to generate the adversarial examples. Based on the MSTAR dataset, we construct an ensemble model by integrating four surrogate models, including AlexNet, VGG, ResNet, and ResNeXt, and all the ensemble models are assigned equal weights. Subsequently, we evaluate the transferability by conducting attacks on DenseNet, SqueezeNet, ShuffleNet, and MobileNet. The results are shown in Table 6.

Table 6. The attack success rates and average misclassification confidences of adversarial examples generated on the ensemble model.

Attack	DenseNet	SqueezeNet	ShuffleNet	MobileNet	Average	Confidence
MI-FGSM	67.3	55.4	64.1	53.9	60.18	89.2
NI-FGSM	72.5	64.3	69.0	54.3	65.03	90.5
VMI-FGSM	65.5	61.7	64.3	52.7	61.05	85.6
DIM	75.2	69.9	73.7	60.9	69.93	87.9
SIM	81.5	73.1	80.6	73.1	77.08	89.0
TIM	79.6	67.7	68.9	63.5	69.93	87.2
Admix	90.6	82.2	85.4	73.4	82.90	86.1
SVA	85.5	79.0	81.5	74.9	80.23	87.5
ITINFA	92.3	87.2	90.0	79.7	87.30	90.7

From Table 6, it is observed that adversarial examples generated on the ensemble model exhibit more stable transferability across various target models with different architectures, with the attack success rate consistently remaining within a specific range. Compared with all the baseline methods, ITINFA achieves the average attack success rate of 87.30% and outperforms the best baseline method with a clear margin of 4.40%. Meanwhile, it exhibits the highest misclassification confidence of 90.7%, demonstrating its superior performance in generating transferable adversarial examples on the ensemble model.

4.5. Ablation Studies

In this subsection, to further explore the superior transferability achieved by ITINFA, we conduct ablation studies on the MSTAR dataset to validate that both intra-class transformations and inter-class nonlinear fusion are beneficial to enhance transferability. We set up the following four configurations for ablation studies: (1) the original ITINFA; (2) removing intra-class transformations; (3) removing inter-class nonlinear fusion; (4) removing both intra-class transformations and inter-class nonlinear fusion. We employ AlexNet as the surrogate model to generate adversarial examples and attack the other seven models to evaluate the transferability.

The experimental results are shown in Table 7. It is observed that configuration 1 exhibits the best transferability. However, removing either intra-class transformations module or inter-class nonlinear fusion module decreases transferability, and configuration 4 exhibits the poorest transferability when both modules are removed simultaneously. Therefore, it is evident that both intra-class transformations and inter-class nonlinear fusion are effective in enhancing transferability and can act synergistically to further enhance transferability.

Table 7. The average attack success rates of adversarial examples generated on AlexNet under different configurations.

Attack	VGG	ResNet	ResNeXt	DenseNet	SqueezeNet	ShuffleNet	MobileNet	Average
Configuration 1	63.1	50.7	56.7	55.1	69.4	63.5	50.5	58.4
Configuration 2	42.7	41.9	44.6	44.1	60.5	56.6	41.3	47.4
Configuration 3	51.7	43.0	45.9	45.4	55.1	57.2	45.1	49.1
Configuration 4	32.7	37.5	37.9	40.3	51.7	45.9	35.5	40.2

4.6. Parameter Studies

In this subsection, we explore the impact of three hyper-parameters on the transferability: the number of transformations m_1, the number of randomly sampled images from other categories m_2, and the strength of nonlinear mapping r. All the adversarial examples are generated on VGG and utilized to attack the target models to evaluate transferability.

ITINFA applies m_1 transformations to the input image to obtain the diverse gradient information. To explore the impact of m_1 on the transferability, we conduct experiments with m_1 varying from 5 to 50, m_2 fixed to 3, and r fixed to 0.8. The attack success rates are recorded in Figure 7. It is evident that the white-box attack success rates consistently remain at a level close to 100%, and the transfer attack success rate is lowest when $m_1 = 5$. As m_1 increases, the richness of the gradient information gradually enhances, leading to a gradual improvement in the transferability of adversarial examples. However, once m_1 exceeds 30, the improvement in transferability becomes negligible, while the computational cost increases substantially. Therefore, we select $m_1 = 30$ in the experiments to maintain a favorable balance between transferability and computational cost.

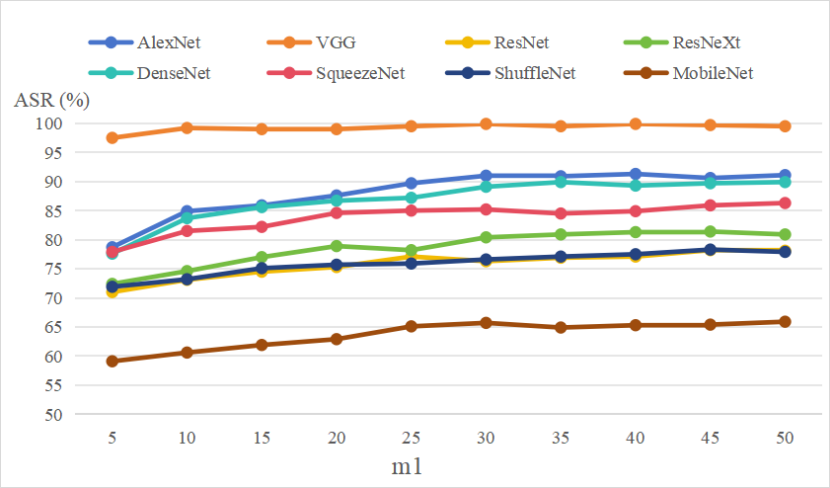


Figure 7. The attack success rates of ITINFA with various numbers of transformations m_1.

In Figure 8, we further report the attack success rates of ITINFA with various values of m_2. We select m_2 from 1 to 5, fix m_1 to 30, and set r to 0.8. When $m_2 \leq 3$, the transferability of adversarial examples on all the target models improves as the value of m_2 increases. This suggests that incorporating the information from other categories more effectively guides the misclassification of adversarial examples. However, when $m_2 > 3$, the improvement in transferability becomes less pronounced and even shows a downward trend. This phenomenon might result from introducing excessive information from other categories, which impairs the transferability of adversarial examples. Additionally, increasing the value of m_2 also brings excessive computational costs. Consequently, $m_2 = 3$ is chosen for our experiments.

In addition, we also explored the impact of nonlinear mapping at different strengths on transferability, and the results are illustrated in Figure 9. In the experiment, r varies from 0.4 to 2.0, m_1 is fixed to 30, and m_2 is fixed to 3. It can be observed that when r is lower than 0.8, the attack success rate increases significantly with the growth in r. However, after r surpasses 1, the attack success rates gradually converge to a stable range and begin to decline slowly. In most of the target models, an optimal attack performance is observed at $r = 0.8$. Consequently, this value is selected as the designated setting for r in our experiments.

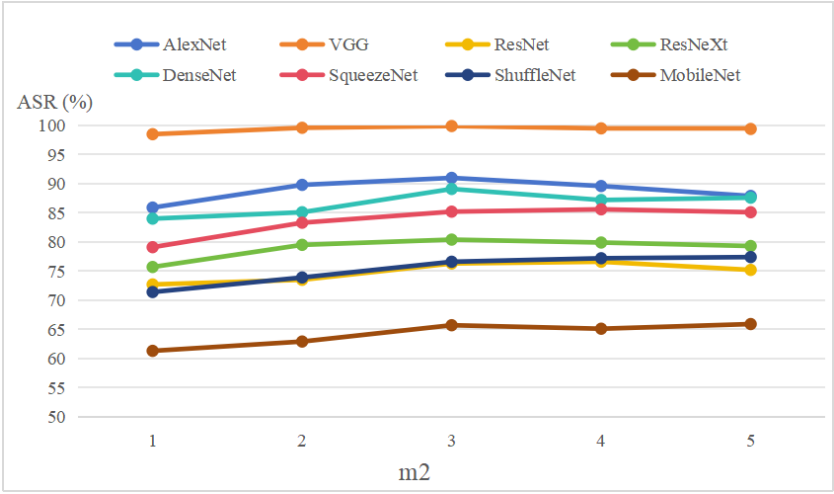


Figure 8. The attack success rates of ITINFA with various numbers of sampled images m_2.

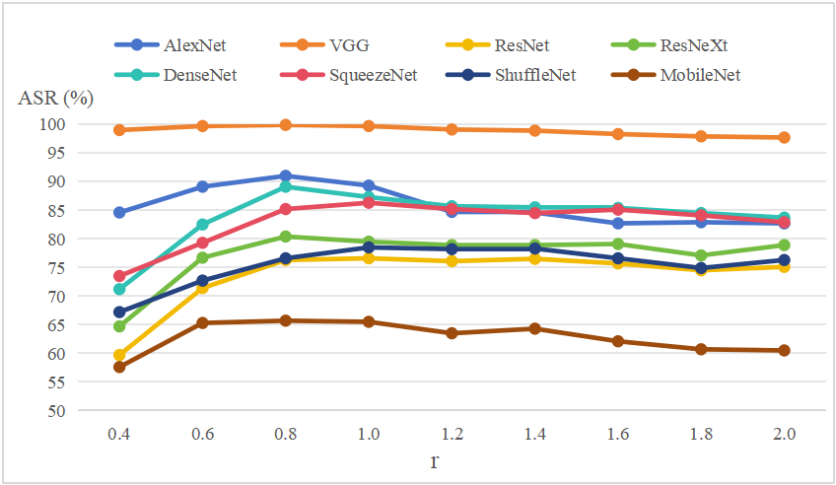


Figure 9. The attack success rates of ITINFA with different strengths of nonlinear mapping r.

4.7. Visualizations

Some results of the adversarial perturbations and adversarial examples generated by the various algorithms are visualized in Figure 10. The first two rows represent the examples from the MSTAR dataset, while the last two rows illustrate the examples from the SEN1-2 dataset. To enhance the visual effect, we utilize the Parula color map to visualize the gray-scale SAR images. It is evident that our proposed ITINFA is capable of achieving satisfactory visual imperceptibility. That is, the adversarial example generated by ITINFA can fool the target model while maintaining high similarity with the original image.

Figure 11 further illustrates an adversarial threat scenario in which a military target tank (T62) is erroneously recognized as a truck (ZIL-131) after the adversarial perturbation generated by ITINFA is added to the original image. We utilize the feature visualization technique GradCAM [56] to visualize the output feature map weights of the last convolutional layer of VGG. It is evident that adding adversarial perturbation results in a significant alteration in the attention regions. When the target category is set to T62, the attention of the model is mainly focused on the target area of the original image, whereas for the adversarial example, it shifts to the background clutter area and relocates back to the target area when the target category is set to ZIL-131. This presentation further enhances the visual understanding of the attack process and indicates the potential of ITINFA for application in real-world scenarios.

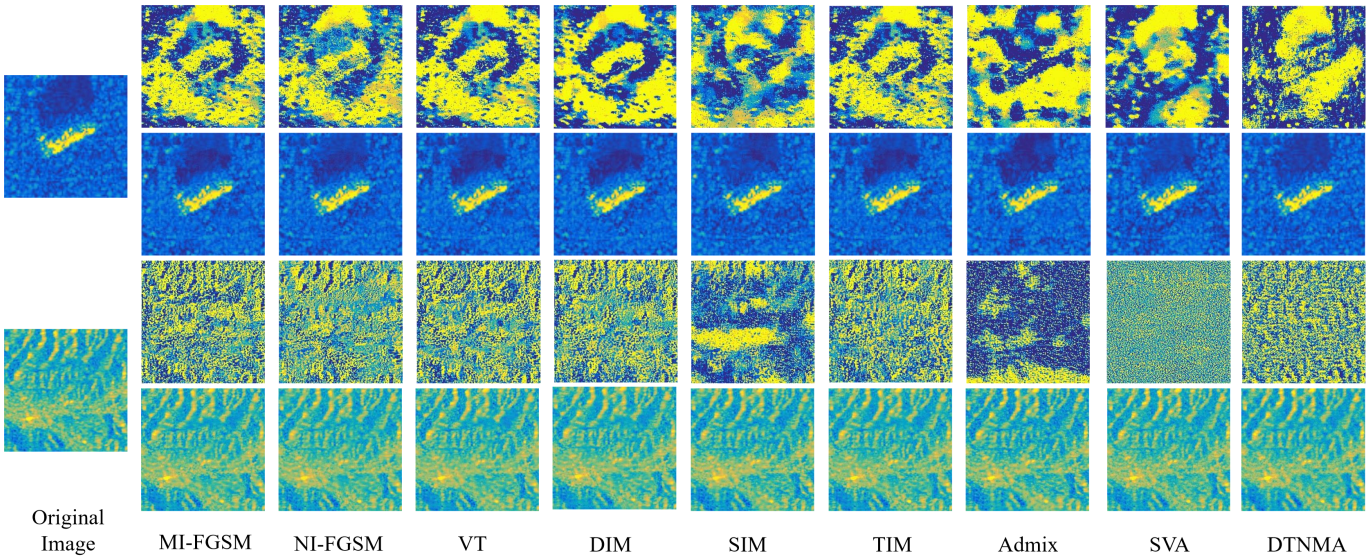


Figure 10. Adversarial perturbations and adversarial examples generated by various algorithms. (**top**) are adversarial perturbations, and (**bottom**) are the corresponding adversarial examples.

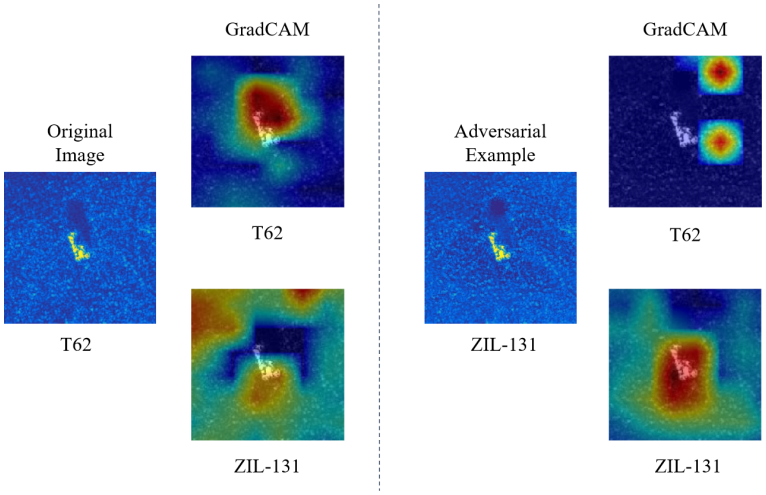


Figure 11. Discrimination maps generated by Grad-CAM on VGG. The maps represent the attention mechanism for the original military target tank (T62) and the misclassified truck (ZIL-131).

5. Discussion

In the field of adversarial attacks on SAR-ATR, due to the highly non-cooperative nature and the secretive characteristics of adversary's SAR systems it is typically challenging to acquire information about target models. Therefore, enhancing the transferability of adversarial examples to unknown target models becomes exceptionally crucial. Input transformation-based attacks represent a prevalent method for boosting transferability today. The essence of these approaches lies in employing data augmentation to process the input image, thereby generating richer gradient information. This facilitates a more generalized optimization direction for updates, avoiding local optima in the optimization process and alleviating the overfitting of adversarial examples on the surrogate model. Therefore, in this article, we continue to explore towards this direction.

The proposed ITINFA comprises two main modules: firstly, from the perspective of the image itself, diverse intra-class transformations are applied to the input SAR image. The transformations include typical techniques used in optical images and methods developed explicitly for the unique properties of SAR imagery. Secondly, from the perspective of the other category images, incorporating information from other categories to better guide

adversarial examples to be misclassified, regardless of the target models. Notably, we employ a nonlinear multiplication for integration rather than linear addition. Experimental results indicate that this modulation approach can achieve better effects. Future research might also consider the effects of other nonlinear mapping methods, such as using a logarithmic function instead of an exponential one. However, due to space limitations, this article will not delve into those. Both the intra-class transformations and inter-class fusion modules have been proven in ablation studies to enhance the transferability of adversarial examples significantly, and they can also be organically integrated with other types of methods to further improve transferability, such as ensemble attacks [57,58] or specialized objective functions [59,60].

Future research in this field will focus on the following aspects. First, enhancing the transferability to unknown black-box models. Given the non-cooperative nature of SAR systems, how to generate highly transferable adversarial examples using a known surrogate model without accessing target models will remain an essential research direction. Second, improving the robustness of adversarial perturbations. Since there may be some degree of position deviation, perspective changes, and other uncontrollable errors when physically implanting adversarial perturbations, ensuring that the generated adversarial examples can still maintain attack capabilities in the face of uncertainties will be a significant challenge. Third, enhancing the physical realizability of adversarial perturbations. Current research primarily focuses on pixel-level perturbations in the digital domain. Although some studies [53,61,62] have linked adversarial perturbations with electromagnetic scattering characteristics of the target by using the attribute scattering center model (ASCM), there is still a long way to go to realize adversarial perturbations physically. Future studies should consider passive interference methods like electromagnetic materials and metasurfaces [63] to alter the target's electromagnetic scattering properties, or using jammers and other active interference methods to add adversarial perturbations to the target's echo signals, thereby physically realizing adversarial attacks in the real world and bridging the gap between theoretical advancements and practical applicability.

6. Conclusions

In this article, we propose a novel adversarial attack method to enhance the transferability of SAR adversarial examples. Initially, we perform diverse intra-class transformations on the input image with a series of data augmentation techniques aligned with the unique characteristics of SAR imagery. This enriches gradient information, guides the optimization direction, and alleviates overfitting to the surrogate model. Subsequently, we innovatively incorporate image information from other categories in a nonlinear manner, which effectively assists the adversarial examples in navigating through the decision boundaries of various models, enhancing their adaptability across different recognition models. Moreover, we integrate a momentum iterative updating algorithm to escape local optima in the optimization process, further enhancing the transferability of the adversarial examples. Comprehensive experimental analysis based on the MSTAR dataset and SEN1-2 dataset has demonstrated that our proposed ITINFA significantly surpasses other baseline methods in terms of transferability while maintaining commendable white-box attack performance.

Author Contributions: Conceptualization, X.H. and Z.L.; methodology, X.H.; software, X.H.; validation, X.H. and Z.L.; formal analysis, X.H. and B.P.; investigation, X.H. and B.P.; resources, B.P.; data curation, X.H.; writing—original draft preparation, X.H.; writing—review and editing, B.P. and Z.L.; visualization, X.H.; supervision, B.P.; project administration, B.P.; funding acquisition, B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Changsha Outstanding Innovative Youth Training Program under Grant kq2107002 and in part by the National Natural Science Foundation of China under Grant 61921001.

Data Availability Statement: The data presented in this study are available in MSTAR at https://download.csdn.net/download/qq_34277608/11537943 (accessed on 13 August 2019). and in SEN1-2 at https://doi.org/10.48550/arXiv.1807.01569 (accessed on 4 July 2018).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chang, S.; Deng, Y.; Zhang, Y.; Zhao, Q.; Wang, R.; Zhang, K. An Advanced Scheme for Range Ambiguity Suppression of Spaceborne SAR Based on Blind Source Separation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
2. Yue, D.; Xu, F.; Frery, A.C.; Jin, Y. Synthetic Aperture Radar Image Statistical Modeling: Part One-Single-Pixel Statistical Models. *IEEE Geosci. Remote Sens. Mag.* **2020**, *9*, 82–114. [CrossRef]
3. Chen, S.; Wang, H.; Xu, F.; Jin, Y. Target Classification Using the Deep Convolutional Networks for SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4806–4817. [CrossRef]
4. Wang, C.; Pei, J.; Wang, Z.; Huang, Y.; Wu, J.; Yang, H.; Yang, J. When Deep Learning Meets Multi-Task Learning in SAR ATR: Simultaneous Target Recognition and Segmentation. *Remote Sens.* **2020**, *12*, 3863. [CrossRef]
5. Zhu, X.; Montazeri, S.; Ali, M.; Hua, Y.; Wang, Y.; Mou, L.; Shi, Y.; Xu, F.; Bamler, R. Deep Learning Meets SAR: Concepts, Models, Pitfalls, and Perspectives. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 143–172. [CrossRef]
6. Pei, J.; Wang, Z.; Sun, X.; Huo, W.; Zhang, Y.; Huang, Y.; Wu, J.; Yang, J. FEF-Net: A Deep Learning Approach to Multiview SAR Image Target Recognition. *Remote Sens.* **2021**, *13*, 3493. [CrossRef]
7. Zeng, Z.; Sun, J.; Xu, C.; Wang, H. Unknown SAR Target Identification Method Based on Feature Extraction Network and KLD–RPA Joint Discrimination. *Remote Sens.* **2021**, *13*, 2901. [CrossRef]
8. Li, J.; Yu, Z.; Yu, L.; Cheng, P.; Chen, J.; Chi, C. A Comprehensive Survey on SAR ATR in Deep-Learning Era. *Remote Sens.* **2023**, *15*, 1454. [CrossRef]
9. Li, B.; Cui, Z.; Wang, H.; Deng, Y.; Ma, J.; Yang, J.; Cao, Z. SAR Incremental Automatic Target Recognition Based on Mutual Information Maximization. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 4005305. [CrossRef]
10. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
11. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [CrossRef]
12. Xu, Y.; Bai, T.; Yu, W.; Chang, S.; Atkinson, P.; Ghamisi, P. AI Security for Geoscience and Remote Sensing: Challenges and Future Trends. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 60–85. [CrossRef]
13. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
14. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial Examples in the Physical World. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
15. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–24 May 2017; pp. 39–57.
16. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Las Vegas, NV, USA, 26–30 June 2016; pp. 2574–2582.
17. Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; Yuille, A.L. Improving Transferability of Adversarial Examples with Input Diversity. In Proceedings of the Computer Vision and Pattern Recognition Conference, Long Beach, CA, USA, 15–20 June 2019.
18. Lin, J.; Song, C.; He, K.; Wang, L.; Hopcroft, J. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
19. Dong, Y.; Pang, T.; Su, H.; Zhu, J. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Long Beach, CA, USA, 15–20 June 2019; pp. 4312–4321.
20. Wang, X.; He, X.; Wang, J.; He, K. Admix: Enhancing the Transferability of Adversarial Attacks. In Proceedings of the International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 16158–16167.
21. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting Adversarial Attacks with Momentum. In Proceedings of the Computer Vision and Pattern Recognition Conference, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9185–9193.
22. Wang, X.; He, K. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 20–25 June 2021; pp. 1924–1933.
23. Ge, Z.; Liu, H.; Wang, X.; Shang, F.; Liu, Y. Boosting Adversarial Transferability by Achieving Flat Local Maxima. In Proceedings of the Conference and Workshop on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 70141–70161.
24. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
25. Li, H.; Huang, H.; Chen, L.; Peng, J.; Huang, H.; Cui, Z.; Mei, X.; Wu, G. Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1333–1347. [CrossRef]

26. Huang, T.; Zhang, Q.; Liu, J.; Hou, R.; Wang, X.; Li, Y. Adversarial Attacks on Deep-Learning-Based SAR Image Target Recognition. *J. Netw. Comput. Appl.* **2020**, *162*, 102632. [CrossRef]
27. Du, C.; Zhang, L. Adversarial Attack for SAR Target Recognition Based on UNet-Generative Adversarial Network. *Remote Sens.* **2021**, *13*, 4358. [CrossRef]
28. Du, C.; Huo, C.; Zhang, L.; Chen, B.; Yuan, Y. Fast C&W: A Fast Adversarial Attack Algorithm to Fool SAR Target Recognition with Deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5.
29. Peng, B.; Peng, B.; Zhou, J.; Xia, J.; Liu, L. Speckle-Variant Attack: Toward Transferable Adversarial Attack to SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5.
30. Zhang, Z.; Gao, X.; Liu, S.; Peng, B.; Wang, Y. Energy-Based Adversarial Example Detection for SAR Images. *Remote Sens.* **2022**, *14*, 5168. [CrossRef]
31. Du, M.; Bi, D.; Du, M.; Xu, X.; Wu, Z. ULAN: A Universal Local Adversarial Network for SAR Target Recognition Based on Layer-Wise Relevance Propagation. *Remote Sens.* **2022**, *15*, 21. [CrossRef]
32. Peng, B.; Peng, B.; Yong, S.; Liu, L. An Empirical Study of Fully Black-Box and Universal Adversarial Attack for SAR Target Recognition. *Remote Sens.* **2022**, *14*, 4017. [CrossRef]
33. Zhang, F.; Meng, T.; Xiang, D.; Ma, F.; Sun, X.; Zhou, Y. Adversarial Deception against SAR Target Recognition Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4507–4520. [CrossRef]
34. Lin, G.; Pan, Z.; Zhou, X.; Duan, Y.; Bai, W.; Zhan, D.; Zhu, L.; Zhao, G.; Li, T. Boosting Adversarial Transferability with Shallow-Feature Attack on SAR Images. *Remote Sens.* **2023**, *15*, 2699. [CrossRef]
35. Wang, X.; Zhang, Z.; Zhang, J. Structure Invariant Transformation for Better Adversarial Transferability. In Proceedings of the International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4607–4619.
36. Ding, J.; Chen, B.; Liu, H.; Huang, M. convolutional neural network with Data Augmentation for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]
37. Argenti, F.; Lapini, A.; Bianchi, T.; Alparone, L. A Tutorial on Speckle Reduction in Synthetic Aperture Radar Images. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–35. [CrossRef]
38. Deledalle, C.A.; Denis, L.; Tupin, F.; Reigber, A.; Jäger, M. NL-SAR: A Unified Nonlocal Framework for Resolution-Preserving (Pol)(In) SAR Denoising. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 2021–2038. [CrossRef]
39. Parrilli, S.; Poderico, M.; Angelino, C.V.; Verdoliva, L. A Nonlocal SAR Image Denoising Algorithm Based on LLMMSE Wavelet Shrinkage. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 606–616. [CrossRef]
40. Dabov, K.; Foi, A.; Katkovnik, V.; Egiazarian, K. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Trans. Image Process.* **2007**, *16*, 2080–2095. [CrossRef] [PubMed]
41. Wang, T.; Ying, Z.; Li, Q.; Lian, Z. Boost Adversarial Transferability by Uniform Scale and Mix Mask Method. *arXiv* **2023**, arXiv:2311.12051.
42. Ross, T.D.; Worrell, S.W.; Velten, V.J.; Mossing, J.C.; Bryant, M.L. Standard SAR ATR Evaluation Experiments Using the MSTAR Public Release Data Set. In Proceedings of the Algorithms for Synthetic Aperture Radar Imagery V, SPIE, Bellingham, WA, USA, 14–17 April 1998; Volume 3370, pp. 566–573.
43. Schmitt, M.; Hughes, L.; Zhu, X. The SEN1-2 Dataset for Deep Learning in Sar-Optical Data Fusion. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2018**, *4*, 141–146. [CrossRef]
44. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep convolutional neural networks. In Proceedings of the Conference and Workshop on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
45. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference, Las Vegas, NV, USA, 26–30 June 2016; pp. 770–778.
47. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
48. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Densely Connected Convolutional Networks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
49. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5 MB Model Size. *arXiv* **2016**, arXiv:1602.07360.
50. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 116–131.
51. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the Computer Vision and Pattern Recognition Conference, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
52. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
53. Peng, B.; Peng, B.; Zhou, J.; Xie, J.; Liu, L. Scattering Model Guided Adversarial Examples for SAR Target Recognition: Attack and Defense. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
54. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

55. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Brooklyn, NY, USA, 2019; pp. 8024–8035.
56. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
57. Liu, Y.; Chen, X.; Liu, C.; Song, D. Delving into Transferable Adversarial Examples and Black-Box Attacks. *arXiv* **2016**, arXiv:1611.02770.
58. Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J.E.; He, K. Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability. In Proceedings of the Computer Vision and Pattern Recognition Conference, New Orleans, LA, USA, 19–24 June 2022; pp. 14983–14992.
59. Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; Yang, Y. Transferable Adversarial Perturbations. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 452–467.
60. Wu, W.; Su, Y.; Chen, X.; Zhao, S.; King, I.; Lyu, M.R.; Tai, Y.W. Boosting the Transferability of Adversarial Samples via Attention. In Proceedings of the Computer Vision and Pattern Recognition Conference, Seattle, WA, USA, 14–19 June 2020; pp. 1161–1170.
61. Qin, W.; Long, B.; Wang, F. SCMA: A Scattering Center Model Attack on CNN-SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5.
62. Zhou, J.; Feng, S.; Sun, H.; Zhang, L.; Kuang, G. Attributed Scattering Center Guided Adversarial Attack for DCNN SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5.
63. Xu, H.; Guan, D.F.; Peng, B.; Liu, Z.; Yong, S.; Liu, Y. Radar One-Dimensional Range Profile Dynamic Jamming Based on Programmable Metasurface. *IEEE Antennas Wirel. Propag. Lett.* **2021**, *20*, 1883–1887. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.