remote sensing

MDPI

Article

Transferable Targeted Adversarial Attack on Synthetic Aperture Radar (SAR) Image Recognition

Sheng Zheng [1], Dongshen Han [2], Chang Lu [2], Chaowen Hou [1], Yanwen Han [1], Xinhong Hao [1] and Chaoning Zhang [2,*]

[1] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China; 3120215162@bit.edu.cn (S.Z.); 3120240257@bit.edu.cn (C.H.); hanyanwen@bit.edu.cn (Y.H.); haoxinhong@bit.edu.cn (X.H.)

[2] School of Computing, Kyung Hee University, Yongin 17113, Republic of Korea; han-0129@khu.ac.kr (D.H.); luchang@khu.ac.kr (C.L.)

* Correspondence: chaoningzhang@khu.ac.kr

Abstract: Deep learning models have been widely applied to synthetic aperture radar (SAR) target recognition, offering end-to-end feature extraction that significantly enhances recognition performance. However, recent studies show that optical image recognition models are widely vulnerable to adversarial examples, which fool the models by adding imperceptible perturbation to the input. Although the targeted adversarial attack (TAA) has been realized in the white box setup with full access to the SAR model's knowledge, it is less practical in real-world scenarios where white box access to the target model is not allowed. To the best of our knowledge, our work is the first to explore transferable TAA on SAR models. Since contrastive learning (CL) is commonly applied to enhance a model's generalization, we utilize it to improve the generalization of adversarial examples generated on a source model to unseen target models in the black box scenario. Thus, we propose the contrastive learning-based targeted adversarial attack, termed CL-TAA. Extensive experiments demonstrated that our proposed CL-TAA can significantly improve the transferability of adversarial examples to fool the SAR models in the black box scenario.

Keywords: synthetic aperture radar (SAR); targeted adversarial attack (TAA); transferability; contrastive learning; generalization

check for updates

Academic Editors: Zhenming Peng and Zhengzhou Li

Received: 24 October 2024
Revised: 12 December 2024
Accepted: 18 December 2024
Published: 3 January 2025

Citation: Zheng, S.; Han, D.; Lu, C.; Hou, C.; Han, Y.; Hao, X.; Zhang, C. Transferable Targeted Adversarial Attack on Synthetic Aperture Radar (SAR) Image Recognition. *Remote Sens.* **2025**, *17*, 146. https://doi.org/10.3390/rs17010146

Copyright: © 2025 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

With the development of deep learning, it has been widely applied across various fields, including synthetic aperture radar (SAR) target recognition [1,2]. Unlike traditional SAR target recognition [3,4], which typically relies on a two-step process of manual feature extraction followed by the classification, deep learning methods [5–8] offer an end-to-end approach with automatic feature extraction, providing significant advantages in target recognition. Recently, deep convolutional neural network (DCNN) models, such as AlexNet [9], VGGNet [10], ResNet [11], and AMS-CNN [5], have been widely used for SAR target recognition. Convolutional kernels allow models to extract features effectively, thus significantly improving performance. Notably, AMS-CNN combines convolutional kernels with an attention mechanism and multi-stream feature extraction, enabling it to capture features effectively.

While deep learning has surpassed traditional methods and achieved remarkable success in target recognition, multiple studies [12–14] have demonstrated that these models are highly susceptible to adversarial attacks, where even subtle, imperceptible perturbations added to the input can fool the model. In addition, in the targeted adversarial attack (TAA), the model's predicted label can be manipulated to match a specific target label. A non-targeted adversarial attack (non-TAA) is considered successful if the model predicts a label different from the original, assuming the initial prediction was correct. However, for a TAA to succeed, the model must predict the exact target label, making it a more challenging task than a non-TAA. Moreover, transferable TAA is yet more challenging because it lacks white box access to the target models, with the adversarial examples generated on a source model.

Despite numerous research on transferable TAA mainly conducted on optical images, it remains unclear whether SAR target recognition models are vulnerable to targeted adversarial attacks. To address this, we conduct the first comprehensive study investigating whether adversarial examples can fool SAR target recognition models to a specified target class in the black box scenario. Unlike three-channel color images, SAR images are single-channel grayscale. Therefore, it is easier to train adversarial perturbations that overfit on the source model, but this makes it harder for the generated adversarial examples to transfer effectively to other target models. Additionally, SAR images tend to have more noise, further complicating the challenge of achieving a transferable adversarial attack. In other words, what makes transferable TAA particularly challenging is the poor generalization of adversarial examples, which struggle to transfer from the source to unseen target models effectively. Contrastive learning (CL) [15], a widely used pre-training technique, has demonstrated its ability to improve the model's generalization across various tasks by encouraging the anchor sample to attract the positive sample and repel the negative samples. This helps the model capture essential features more effectively and improves its ability to handle hard samples. Since CL can improve the model's generalization to hard samples, we conjecture that it can also benefit the generalization of adversarial examples. Therefore, we incorporate CL into our training process to enhance the transferability of TAA. Specifically, we use the InfoNCE [16], one of CL's most effective loss functions, to optimize adversarial examples. By leveraging the model's logits, which are demonstrated to represent the feature of the image, we calculate the InfoNCE loss based on the logits of anchor, positive, and negative samples.

Overall, with the increasing importance of SAR target recognition, it is essential to understand the robustness of these models. Previous research [5–7] has shown that these models are susceptible to adversarial attacks, primarily in a white box setting where full access to the model's structure and weights is available. In contrast, this work focuses on evaluating the adversarial robustness of SAR target recognition models in a more practical and challenging black box setting. Our contributions are summarized as follows:

- To understand the transferability of SAR target recognition models, to the best of our knowledge, our work conducts a first yet comprehensive investigation on them.
- We identify the challenges of transferable TAA as the limited generalization of adversarial examples and mitigate it by leveraging contrastive learning during the training.
- Extensive experiments verify that our proposed CL-TAA can significantly enhance the performance of transferable TAA.

The rest of the paper is organized as follows: Section 2 discusses related work on adversarial attack and contrastive learning. Section 3 presents the methodology, including classical recognition models, seminal black box attack methods, and our proposed CL-TAA. Section 4 details the experiments, and Section 5 concludes the paper.

2. Related Work

2.1. Adversarial Attack

Deep neural networks are vulnerable to adversarial examples, as revealed in [12,17–20], where even imperceptible perturbations to the input can fool the model into predicting wrong results. With the revelation of this intriguing property, numerous studies [21–25] have been conducted on the model's robustness, where these studies can be classified based on the knowledge of the attacker and the goal of the attacker. From the perspective of the attacker's goal, adversarial attacks can be divided into non-targeted and targeted adversarial attacks [26–33]. Non-targeted adversarial attacks (non-TAAs) aim to fool the model by predicting the wrong label, which is different from the ground truth one [21,22,34]. Targeted attacks, however, pursue a specific output and are considered successful only when the model's output matches a predefined target label [23,24]. Regarding the attacker's knowledge, adversarial attacks can be divided into white box and black box attacks [35–37]. In the white box setup, the attackers can fully access the model's structure, parameters, and gradient information, enabling them to exploit the models' specific vulnerabilities and craft highly tailored attacks [26]. By contrast, the black box setting, which is more practical for real-world scenarios, limits attackers to only the model's output, leaving them without any information about the model itself [38]. The Fast Gradient Sign Method (FGSM) [23] and Projected Gradient Descent (PGD) [24] method, known for their effectiveness in generating adversarial examples, are widely used for the white box setting. Iterative-FGSM (I-FGSM) [39], as the extension of FGSM, adopts multiple steps to generate the perturbation to enhance the attack's success rate. It should be noted that PGD is also sometimes referred to as I-FGSM [39]. Carlini and Wagner (C&W) attack [40], another more sophisticated white box attack method, generates highly effective adversarial examples with minimal perturbations. Although these methods are effective in a white box setting, their performance is poor in a black box scenario. Therefore, multiple works [13,14,41] have been proposed to improve the transferability of adversarial examples with the black box setup. Momentum-Iterative FGSM (MI-FGSM) [13] incorporates momentum to update the adversarial examples and significantly improve its transferability across different models. Similarly, Translation-Invariant FGSM (TI-FGSM) [14] introduces translation invariance to generate more transferable adversarial examples, further enhancing their effectiveness. Diverse-Input FGSM (DI-FGSM) [41] applies random transformations to the input, such as scaling and cropping, to improve the generalization of the adversarial examples across different models. In our work, we focus on the more challenging targeted adversarial attack under the black box setting for SAR target recognition models.

2.2. Contrastive Learning

Contrastive learning (CL) [15,42], the milestone technology for self-supervised learning, is used to learn the augmentation-invariant representation. The core idea of CL is to pull together positive pairs while pushing apart negative pairs, enabling models to learn effective representations. In [43], it leverages a set of source labels via data augmentation, allowing the network to capture the effective features of images. Unlike the method mentioned above, which utilizes the parametric form, another work [44] adopts a non-parameter by using the apparent visual similarity among categories. In addition, it introduces a memory bank to store representations of individual instances, allowing a more efficient comparison of each sample against a wide array of negatives. Similar to using a memory bank, MoCo [45] introduces a momentum encoder with a queue and moving-averaged dynamic dictionary, which enables the maintenance of a large and consistent dictionary of negative samples. Besides utilizing the memory bank, leveraging in-batch negative samples can also avoid additional storage, simplifying the training pro-

cess. SimCLR [46] sample negative examples implicitly using the other sample and its augmented views in the same batch, effectively learning the meaningful representation. Another work [47] enhances the feature representation of contrastive models by exploring the importance of local and global feature alignments, resulting in the good capture of fine-grained and high-level information. Given the strong generalization capabilities of CL, it has been successfully applied in various tasks, such as image classification [43], object detection [45], segmentation [46], and 3D representation [16]. In this work, we utilize CL techniques to improve the generalization of adversarial examples.

3. Methodology

3.1. Background and Related Models

Transferable targeted adversarial attack. We first introduce the targeted adversarial attack for the classification task and define f_t as the target model. Let D represent the labeled image dataset with an image label pair $(x, y) \in D$, where x represents the image and y is the corresponding ground truth label. The image x has the dimensions $H \times W \times C$, with H being the height, W the width, and C the number of channels. Let δ denote the perturbation; the optimization goal is to generate the adversarial example $x_{adv} = x_{clean} + \delta$, which can fool the model f_t to predict a specific label y_t. Given the target model is a black box, the purpose of a targeted transfer attack is to generate the adversarial example with a white box source model f_s. Instead of generating an adversarial example by directly optimizing x_{adv} to minimize the output of the target model and the target label, transferable targeted attack focuses on minimizing the output of the source model and target label to create adversarial examples, which can be formulated as follows:

$$\delta^* = \max_{\delta \in \mathbb{S}} L_{CE}(f_s(x_{clean} + \delta), y_t). \tag{1}$$

$\delta \in \mathbb{S}$ is used to constrain the maximum perturbation magnitude, ensuring that the perturbation is imperceptible. L_{CE} is the cross-entropy loss function. After generating adversarial examples x_{adv} by using the source model, it can directly attack the target model with the goal of $f_t(x_{adv}) = y_t$. To enhance the transferability of targeted adversarial attacks by generating adversarial examples, we propose to improve it, as detailed in Section 3.2.

Advances in transferable attack. Transferable attack involves iterative optimization on a single image based on the source model without requiring model training or additional data. As mentioned in Section 2.1, attacks can be classified into non-targeted and targeted depending on whether the target label is specific. We first introduce the classical methods for transferable targeted attacks. The Iterative Fast Gradient Sign Method (I-FGSM) [48], widely used for transferable targeted attacks, is an extension of the Fast Gradient Sign Method (FGSM) [23] for generating adversarial examples by employing multiple steps times with small step sizes instead of using a single step. After each step, the pixel values are clipped to ensure they remain within an ϵ-neighborhood of the original image. The attack optimization of I-FGSM can be expressed as follows:

$$\boldsymbol{x}'_0 = \boldsymbol{x}, \;\; \boldsymbol{x}'_{i+1} = \boldsymbol{x}'_i - \alpha \cdot \operatorname{sign}(\nabla_x J(\boldsymbol{x}'_i, y_t)), \tag{2}$$

where $\boldsymbol{x}'_i$ is the perturbed image of the i-th iteration, and $J(\cdot, \cdot)$ is the loss function. To ensure that the perturbation is imperceptible, the maximum allowed magnitude is limited, i.e., it satisfies $\|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon$.

Momentum Iterative–FGSM (MI-FGSM) [13], built upon the I-FGSM, is proposed to utilize a momentum iterative gradient-based method to improve the transferability of generated adversarial examples. The momentum method [49] is a technique that accumulates a velocity vector in the gradient direction of the loss function across iterations

to accelerate the gradient descent. The previously accumulated gradients are facilitated to barrel through narrow valleys to stabilize update directions and reduce the chances of becoming stuck in poor local minima [50,51]. The attack optimization of MI-FGSM can be expressed as follows:

$$g_{i+1} = \mu \cdot g_i + \frac{\nabla_x J(x'_i, y_t)}{\|\nabla_x J(x'_i, y_t)\|_1}, \; x'_{i+1} = x'_i - \alpha \cdot \; \text{sign}(g_i), \tag{3}$$

where g_i and μ refer to accumulated gradients at the i-th and decay factor, respectively. Another similar technique that uses Nesterov instead of momentum to enhance the performance of transferable attacks is explored in [52].

Translation Invariant–FGSM (TI-FGSM) [14] is proposed to boost the transferability of adversarial examples for the target model. It addresses the issue of adversarial examples overfitting to a specific model by optimizing them using randomly translated input images, where it draws inspiration from data augmentation techniques that are typically used to prevent overfitting during model training. Since calculating gradients for multiple translated images is computationally intensive, the author proposes an efficient approach: instead of repeatedly translating the images, it applies a convolutional kernel to the original image to compute locally smoothed gradients. The attack optimization of TI-FGSM can be expressed as follows:

$$x'_{i+1} = x'_i - \alpha \cdot \; \text{sign}(W * \nabla_x J(x'_i, y_t)), \tag{4}$$

where W represents the convolutional kernel used in the TI-FGSM. It has been demonstrated in [53] that the transferability of TI-FGSM benefits from using a smaller kernel size to optimize adversarial examples.

Diverse-Input FGSM (DI-FGSM) [41], similar to TI-FGSM, is also inspired by data augmentation [9,10,54], where DI-FGSM improves the transferability of adversarial examples by creating a diverse input pattern through resizing, cropping, and rotating. Unlike the TI-FGSM, which adopts the fixed augmentation parameter over iterations, DI-FGSM performs different augmentations on the input data to increase the diversity of input images. The attack optimization of TI-FGSM can be expressed as follows:

$$x'_{i+1} = x'_i - \alpha \cdot \; \text{sign}(\nabla_x J(T(x'_i, p), y_t)), \tag{5}$$

where $T(\cdot, \cdot)$ is the stochastic transformation function. $T(x'_i, p)$ can be formulated as follows:

$$T(x'_i; p) = \begin{cases} T(x'_i) & \text{with probability } p \\ x'_i & \text{with probability } 1 - p. \end{cases} \tag{6}$$

$T(\cdot)$ is implemented by resizing the input images to a random size and applying random padding, where zeros are added around the images in a randomized way [55].

Models of deep neural network. AlexNet [9] is a milestone in improving classification accuracy for neural networks. Unlike traditional machine learning algorithms [56], which rely on multiple stages to extract the feature, AlexNet proposes a neural network that performs automatic, end-to-end feature extraction composed of five convolutional layers and three fully connected layers. The success of AlexNet is largely attributed to its innovative structure, which includes using the dropout layer, shifting from the Sigmoid function to ReLU for the activation layer, and replacing average pooling with max pooling. Since AlexNet is excellent at extracting the features of examples, it has also been applied to the SAR automatic target recognition task [57].

VGGNet [10], proposed by the Visual Geometry Group (VGG) of Oxford University, is a convolutional neural network algorithm based on AlexNet. It demonstrated that

deeper neural networks can improve performance to some extent. One key difference from AlexNet is that VGGNet removes the LRN layer, as it was found to have a minimal impact on network performance. Additionally, instead of using larger 5×5 convolutional kernels, VGGNet opts for smaller 3×3 kernels. This choice maintains the same receptive field but allows for more linear variations, and it also reduces the number of parameters in the convolutional layer by about 45%. VGGNet has also been applied to the SAR target recognition and achieves competitive results [58].

ResNet [11], a seminal model for deep neural networks (DNNs), is proposed to address the issue of hard training when increasing the depth of neural networks. Unlike the AlexNet and VGGNet with a more shallow layer, the ResNet model can reach 152 layers. While deeper neural networks theoretically excel at capturing complex and detailed image features, they become increasingly difficult to train for gradient vanishing and explosion. ResNet mitigates the issue by using residual blocks with shortcut connections, where the low-level feature map z (a certain mapping in the network)) is directly used for the input of the high level.

EfficientNet [59] proposes to improve image classification performance by applying scale strategy to all dimensions of CNN, including depth, width, and resolution. The scaling strategy relies on a compound coefficient that adjusts depth, width, and resolution uniformly, ensuring a balanced expansion of the network while optimizing performance. Additionally, EfficientNet scales proportionally in all dimensions based on predefined constants, with the computational resources increasing, maintaining its efficiency and effectiveness. This combination of a scalable design and systematic scaling methodology has set new benchmarks in CNN performance, demonstrating its adaptability and high performance across diverse tasks.

MobileNet [60] is a lightweight CNN model using depth-wise separable convolutions. The implementation of split a standard convolution into two operations: depth-wise convolution for spatial filtering and point-wise (1×1) convolution for combining features. This design dramatically lowers computational complexity and reduces model size without compromising accuracy. MobileNet enables two adjustable hyperparameters for enhanced flexibility. The width multiplier uniformly reduces the number of channels in each layer, while the resolution multiplier decreases the input image resolution, thus lowering computational demands. These parameters allow the model to be tailored to specific application requirements, making it well-suited for resource-constrained environments.

The AMS-CNN (attention-based multi-stream CNN), as introduced in [5], represents the pioneering work being carried out to leverage CNNs for SAR image recognition. The model is structured into three key components: the convolutional module, the stream module, and the classification module. The convolutional module, consisting of convolutional layers, pooling layers, a spatial attention layer, and a channel attention layer, aims to extract low-level features. By using these simple yet effective structures, the AMS-CNN considerably improves the performance of SAR image recognition while reducing the number of parameters compared with those in ResNet.

ConvNeXt [61] is a CNN developed in recent years designed to bridge the performance gap between traditional CNN and Vision transformers. It integrates innovative design principles inspired by vision transformers, such as large kernel sizes, layer normalization, and inverted bottleneck blocks, while maintaining the simplicity and efficiency of traditional ConvNets. It utilizes GELU activation functions and employs larger convolutional kernels to expand the receptive field, which mimickes the global self-attention capabilities of transformers. Layer normalization is used in place of batch normalization, and the architecture adopts a simplified design with fewer activation and normalization layers for streamlined efficiency.

3.2. Our Proposed Method

Performing an adversarial attack on a white box model is easier because the attackers can access the model's structure and weights. Without access to this information, a transferable adversarial attack relies on a source model, where the adversarial example generated on the source model directly applies to the black box models. In this work, we attribute the challenge of this task to the poor generalization of the generated adversarial examples. CL [15] is a widely used pre-training technique for improving the model's generalization capability before training the model for the target task. Inspired by its success in improving the model's generalization capability, we aim to apply it to the generation process of adversarial examples so that they can transfer better to unseen models.

The illustration of the difference between traditional CL and our proposed CL-TAA is shown in Figure 1. Contrastive learning is widely used in unsupervised learning to increase the model's generalization to the unseen hard sample. In essence, contrastive learning is sophisticated in improving the model's generalization by comparing the similarity between key and positive features and the differences between key and negative features. We conjecture that contrastive learning can also enhance the generalization of adversarial examples by comparing the similarity between key and positive features and the differences between key and negative features. Unlike previous works [46], which utilize CL for visual representation to enhance the model's generalization of the hard examples, our work specifically focuses on improving the generalization of adversarial examples. In other words, we aim to utilize CL to generate adversarial examples trained on a source model to attack a range of different black box models effectively. Another key distinction in our approach compared to previous studies is that we incorporate CL during the training process to optimize adversarial examples. Earlier works usually employed CL in the pre-training phase. Typically, CL is used in self-supervised settings, where models are trained on unlabeled data and then fine-tuned for downstream tasks. In our case, CL serves as a technique to improve the generalization of adversarial examples rather than for pre-training.

CL-based TAA Method. In the classical CL, three types of samples are typically involved: the anchor sample, positive sample, and negative samples, where these samples typically utilize the feature from the encoder. The anchor sample is sample of interest, and the positive samples are variations or augmentations derived from the anchor sample, while the negative samples are other randomly selected samples. We adopt the same practice in our work to generate the targeted adversarial examples, and what makes it different from the classical CL is the choice of these samples.

The common consensus is that DNNs are for feature extraction where logit values indicate the presence of the feature in the image, and the highest value in the logit is used for identifying the ground truth classes. The logit refers to the output feature of the model before the final softmax layer [29]. In [29], the authors demonstrate that the logit can be viewed as the features of the images. Since classical DNN models are typically designed without an encoder, we adopt logit for our CL. Specifically, we utilize the logit of the image as the anchor sample because it is the input of interest in this context. It has been demonstrated that the effect of adversarial examples is attributed to the governing of perturbation [29]. To improve the perturbation dominance, we create the positive sample by blending a random clean image with the adversarial example. The InfoNCE [16] loss has been used in multiple works and has become a de facto standard loss for CL; thus, we utilize InfoNCE in our work. Following the notation in [45], we represent the logit of the anchor sample, positive sample, and negative sample as q, k+, and k−, respectively. To eliminate scale discrepancies, these logits are typically L2-normalized. Utilizing these

normalized logits, InfoNCE loss applied in the contrastive learning-based TAA generation can be formulated as follows:

$$L_{infonce} = -log\frac{exp(q \cdot k_+/\tau)}{exp(q \cdot k_+/\tau) + \sum_{i=1}^{K} exp(q \cdot k_-^i/\tau)}, \quad (7)$$

where τ is the temperature controlling the sharpness or smoothness of the similarity distribution between feature vectors and thus affects how the model weighs positive and negative pairs when computing the loss. Unlike the classical CL method that generates the negative samples once and then saves them, we adopt the different negative samples for each iteration to increase the generalization of adversarial samples. In other words, k_- in Equation (7) for each iteration is different for each iteration.

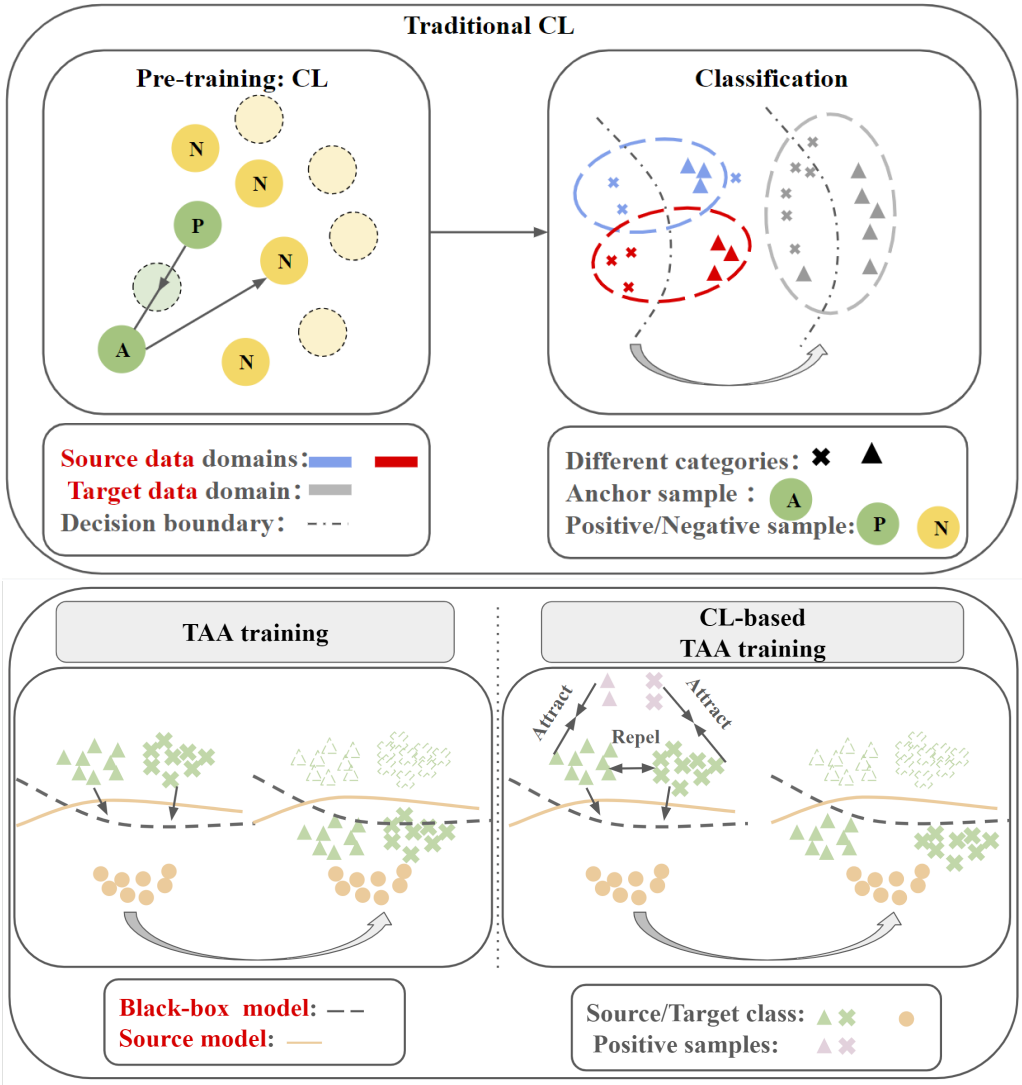


Figure 1. Illustration of traditional contrastive learning (CL) and our proposed contrastive learning-based targeted adversarial attack (CL-TAA). There are two main differences between traditional CL and our method. First, traditional CL aims to improve the model's generalization to hard samples, while our CL-TAA focuses on enhancing the generalization of adversarial examples across different black box models. Another difference is that traditional CL is typically used during the pre-training stage, whereas our CL-TAA is applied during the training.

Practical Implementation. Contrastive learning requires a large number of SAR images for positive and negative samples, while SAR images are limited and expensive to obtain. To address this issue, we use the patches randomly cropped from the to-be-attacked

clean image to generate the positive sample. For the negative samples, we randomly select the negative samples once and also use patches randomly cropped from these samples for each iteration. For these patches, we perform RandomResizedCrop in touchvision without considering the parameter of the random aspect ratio. We term our proposed method as CL-based targeted adversarial attack (CL-TAA); this method drastically reduces the usage of clean images. CL-TAA is summarized in Algorithm 1 and builds on the basic TAA method by incorporating a regularization term, which helps improve the generalization of adversarial examples. Following [62], we set the maximum allowable perturbation magnitude to 16/255, with a step size of 2/255.

Algorithm 1 CL-TAA

Input: Source model f_s; input image x; images for negative samples $X_{neg} = \{x_{neg}^1, x_{neg}^2, \ldots, x_{neg}^M\}$; target label y_t; allowed maximum perturbation magnitude ϵ; total number of iterations T; step size α; hyperparameter of loss function λ; cross-entropy loss function $L_{CE}(\cdot, \cdot)$.

Output: Adversarial example x^* with $\|x^* - x\|_\infty \leq \epsilon$.

$x_0^* = x$;

for $t = 0, 1, 2, ..., T-1$ **do**

$x_{add} = RandomResizedCrop(x)$; $x_{pos} = x_{add} + x_t^*$;

$X_{neg} = RandomResizedCrop(X_{neg})$;

$logit_{pos} = f_s(x_{pos})$;

$logit_{adv} = f_s(x_t^*)$;

$logit_{neg} = f_s(X_{neg})$;

Calculate the loss as

$$L_{total} = L_{CE}(f_s(x_t^*), y_t) + \lambda L_{infonce}(logit_{adv}, logit_{pos}, logit_{neg}); \tag{8}$$

Obtain the gradient by

$$g_t = \nabla_x L_{total}; \tag{9}$$

Update x_{t+1}^* as

$$x_{t+1}^* = Clip_x^\epsilon\{x_t^* - \alpha \cdot \text{sign}(g_t)\}; \tag{10}$$

end for

return $x^* = x_T^*$.

4. Experiments

4.1. Experimental Setup and Dataset

We conduct our experiments on the moving and stationary target acquisition and recognition (MSTAR) dataset [63,64], with a resolution of 0.3 m. The MSTAR dataset, provided by the Air Force Research Laboratory (AFRL) and Defence Advanced Research Projects Agency (DARPA), is widely used for SAR image recognition and includes ten military targets. These SAR images are obtained by radar operating in the X-band under HH polarization conditions, with the imaging depression angle set at two configurations: 15° and 17°. As shown in Figure 2, the MSTAR dataset contains ten different military targets, 2S1, BMP2, BTR60, BTR70, D7, T62, T72, ZIL131, and ZSU234, with their corresponding optimal images also provided.

We crop the MSTAR SAR images as 128 × 128 and obtain a total of 5093, where each image is labeled as one of the ten types of targets. In addition, we normalize the image to the range of $[0, 1]$ for the stable training. To evaluate the proposed method, we adopt the SAR images with different imaging depressions for the training and test sets, with a 17° imaging depression angle for the training set and a 15° imaging depression angle for the test set. The details of ten types of targets are shown in Table 1, with a total number of 2741 for the training set and 2346 for the test set.

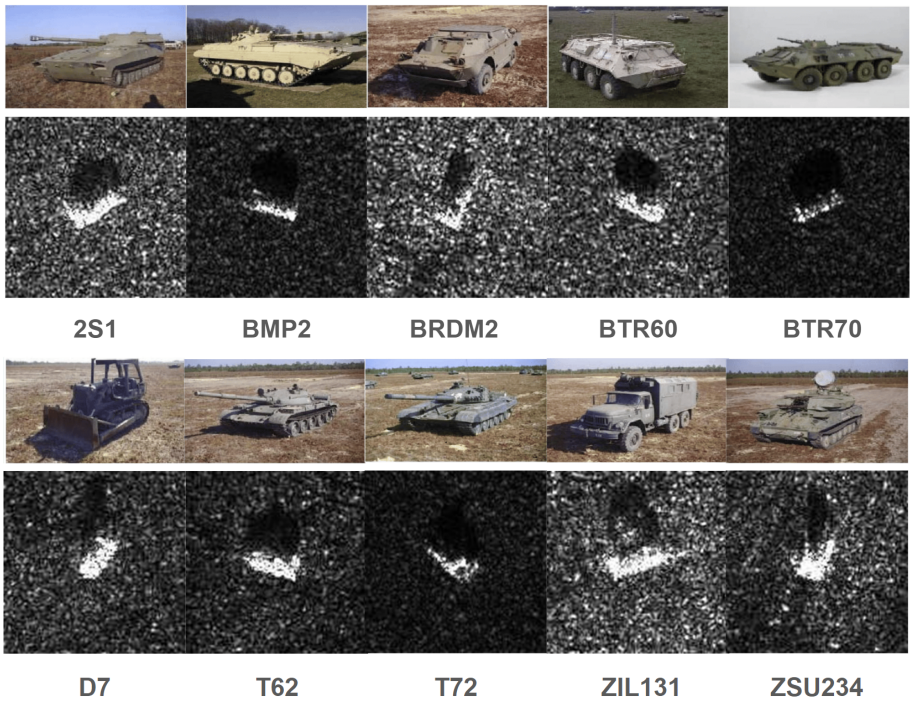


Figure 2. SAR images for ten classes in MSTAR dataset and their corresponding optical images.

Table 1. Sample number of training and test sets for ten types of military targets.

Training Set		Test Set	
Classes	Number	CLasses	Number
2S1	299	2S1	274
BMP2	233	BMP2	195
BRDM2	298	BRDM2	195
BTR60	256	BTR60	195
BTR70	233	BTR70	196
D7	299	D7	274
T62	299	T62	273
T72	232	T72	196
ZIL131	299	ZIL131	274
ZSU234	299	ZSU234	274

To generate reliable adversarial examples and evaluate the effectiveness of our proposed method, we trained several classical and SAR-based recognition models, including AlexNet, VGG16, ResNet, AMS-CNN, EfficientNet, MobileNet, and ConvNeXt, on the MSTAR dataset. The training and test accuracies of these models are shown in Table 2. We can observe that the test accuracy of all models exceeds 80%, providing a solid foundation for our follow-up experiments.

Table 2. Training and test accuracy of deep learning models (%).

Model	Dataset	Training Accuracy	Testing Accuracy
AlexNet	MSTAR	100	96.78
VGG16	MSTAR	100	97.36
ResNet18	MSTAR	100	96.57
EfficientNet	MSTAR	100	94.84
MobileNet	MSTAR	100	83.80
ConvNeXt	MSTAR	100	88.66
AMS-CNN	MSTAR	100	98.14

4.2. Results on Transferable Targeted Attack

As shown in Section 3.1, we introduced seven different model architectures, including AlexNet, VGG16, ResNet18, AMS-CNN, EfficientNet, MobileNet, and ConvNeXt, for SAR image recognition. Given that AMS-CNN outperforms other models, with an accuracy of 98.14 %, we adopt it as the source model and report its results on all other models. The setup for all these experiments is as follows: the weight of CL loss is set to 0.01, the number of negative samples is 50, the size of the perturbation budget is 16/255, the step size is 2/255, and the number of iterations as 20. We randomly select 1000 images in the MSTAR dataset to conduct the experiments. In addition to evaluating our proposed method on the MSTAR dataset, we also provide the results on the CARABAS-II, the other open-source SAR image dataset. The results of CL-TAA on these two different datasets are detailed in the following tables: Table 3 for MSTAR and Table 4 for CARABAS-II. We can observe that CL-TAA achieves better performance than TAA by a large margin on these two different SAR datasets. Taking the AlexNet model as an example, CL-TAA achieves an accuracy of 17.9% on the MSTAR dataset, compared to 14.9% with I-FGSM. This indicates that our proposed CL loss significantly improves the performance of TAA. We also present the I-FGSM and CL-TAA confusion matrix, shown in Tables 5 and 6, respectively. I-FGSM makes the predicted labels cluster in BRDM2, which leads to decreased performance. In contrast, CL-TAA is more effective at predicting the target label.

Table 3. Quantitative evaluation of black box targeted attack by using the AMS-CNN as the source model on the MSTAR dataset (%).

	Target Models					
Attack	**AlexNet**	**VGG16**	**ResNet18**	**EfficientNet**	**MobileNet**	**ConvNeXt**
FGSM	10.9	10.5	10.3	10.9	11.2	10.9
I-FGSM	14.9	15.5	13.6	13.2	11.7	15.2
CL-TAA	**17.9**	**20.0**	**17.3**	**14.3**	**15.2**	**17.5**

Table 4. Quantitative evaluation of black box targeted attack by using the AMS-CNN as the source model on the CARABAS-II dataset (%).

	Target Models					
Attack	**AlexNet**	**VGG16**	**ResNet18**	**EfficientNet**	**MobileNet**	**ConvNeXt**
FGSM	10.46	10.28	11.25	7.14	14.28	15.87
I-FGSM	14.67	11.90	16.67	10.31	18.42	28.73
CL-TAA	**15.87**	**12.69**	**19.04**	**11.11**	**20.63**	**30.95**

Table 5. Confusion matrix of I-FGSM (%). We present the result using AMS-CNN as the source model.

	Target Labels									
Predicted Labels	**BTR70**	**ZSU234**	**ZIL131**	**BMP2**	**BRDM2**	**BTR60**	**2S1**	**T62**	**T72**	**D7**
BTR70	0.4	0.2	0.4	0.3	0.0	0.8	0.4	0.4	0.2	0.6
ZSU234	1.6	2.3	1.7	1.0	0.8	0.4	1.4	1.2	0.6	0.7
ZIL131	1.1	2.1	0.5	1.5	0.6	0.4	0.7	0.7	0.6	0.5
BMP2	0.3	0.1	0.5	0.6	0.0	0.6	0.2	0.1	0.8	0.3
BRDM2	3.4	2.4	2.6	4.3	4.9	2.5	2.7	2.2	3.7	2.9
BTR60	1.5	0.3	0.3	0.0	0.4	0.6	0.7	1.0	0.6	0.3
2S1	0.7	1.7	2.7	1.0	0.7	2.0	1.9	1.0	1.8	2.1
T62	1.2	0.6	2.4	0.6	1.5	1.7	1.0	2.7	1.2	1.6
T72	0.0	0.3	0.1	0.0	0.0	0.1	0.0	0.1	0.1	0.2
D7	0.2	0.3	0.4	0.9	0.5	0.0	0.5	0.0	0.4	0.9

Table 6. Confusion matrix of CL-TAA (%). We present the results using AMS-CNN as the source model.

	Target Labels									
Predicted Labels	**BTR70**	**ZSU234**	**ZIL131**	**BMP2**	**BRDM2**	**BTR60**	**2S1**	**T62**	**T72**	**D7**
BTR70	1.5	0.3	1.0	1.1	0.1	1.1	1.4	0.8	0.5	0.8
ZSU234	1.6	3.6	2.3	1.3	1.0	0.7	1.7	1.6	1.1	0.3
ZIL131	1.4	1.2	1.3	1.5	1.4	0.9	1.2	1.0	0.7	0.5
BMP2	1.1	0.1	0.5	1.4	0.6	1.0	0.2	0.8	1.7	0.2
BRDM2	1.6	1.3	0.7	1.7	2.7	0.9	0.9	1.1	1.3	1.7
BTR60	1.6	0.2	0.3	0.1	0.3	1.1	1.2	0.7	0.6	0.1
2S1	0.2	1.7	2.7	1.0	1.3	1.6	1.9	1.2	1.9	1.9
T62	0.9	0.6	1.1	0.4	0.7	1.3	0.2	1.9	1.2	1.5
T72	0.1	1.0	1.0	0.7	1.0	0.5	0.5	0.3	0.8	0.2
D7	0.4	0.5	0.8	1.1	0.7	0.1	0.4	0.1	0.3	1.7

Visualization results. Figure 3 presents side-by-side comparisons of attacks across different target models. It shows that our proposed CL-TAA significantly outperforms I-FGSM on various target models. The heatmaps generated by grad-cam [65] for both clean and adversarial images are shown in Figure 4. This indicates that the adversarial perturbation can significantly change the focus areas across different models. We also visualize the logits used in CL, shown in Figure 5. We can observe a clear distinction: the negative sample logits have a very different distribution from the key logits, while the positive sample logits closely resemble the key logits. In addition, we randomly select four adversarial examples for visualization, shown in Figure 6, where $x_{adv} = x_{clean} + x_{perturbation}$. Since the perturbation magnitude is limited to $16/255$, we scale it to $(0, 255)$ for better visualization. We notice that it is challenging to distinguish between the clean images and their adversarial counterparts, suggesting that the adversarial examples generated by CL-TAA are not easily detectable.

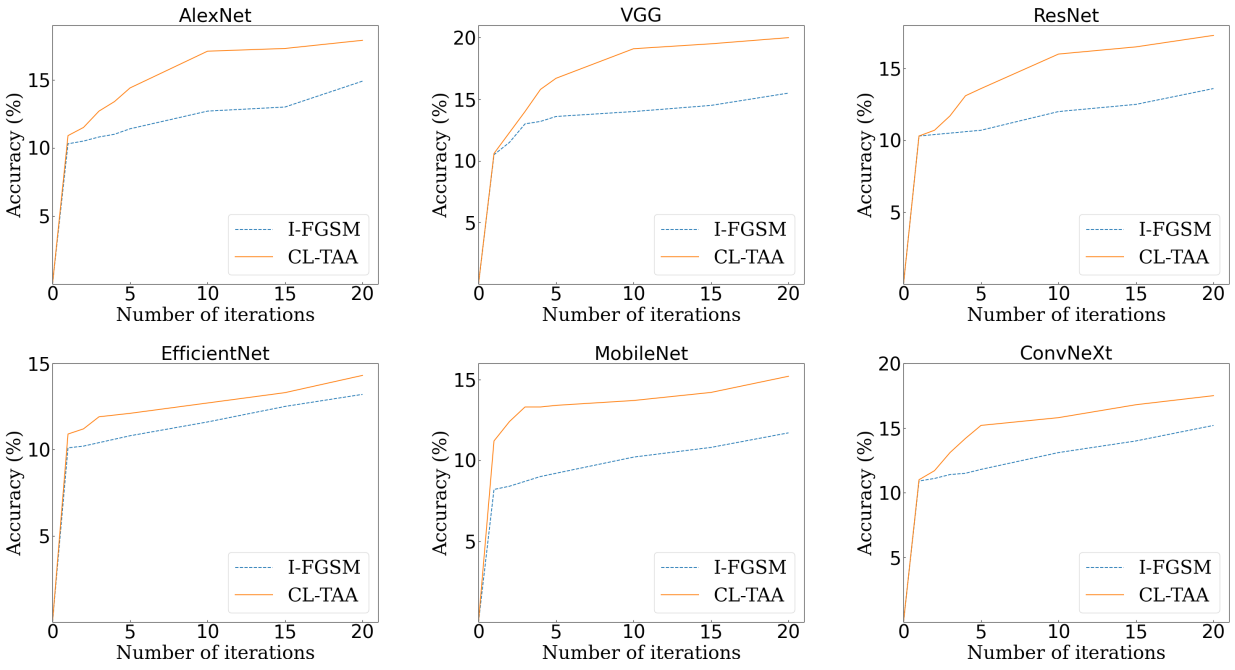


Figure 3. Targeted transfer success rates (%) on different models.

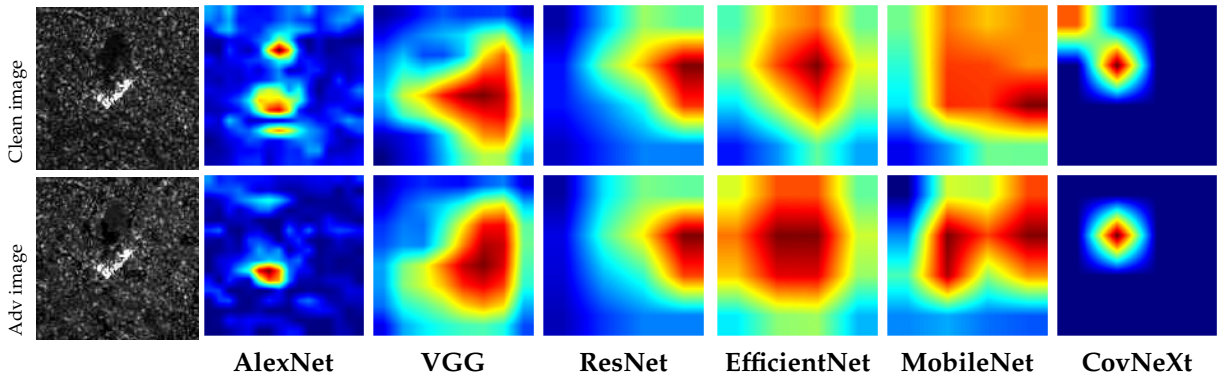


Figure 4. Visualization of the heatmap for different models.

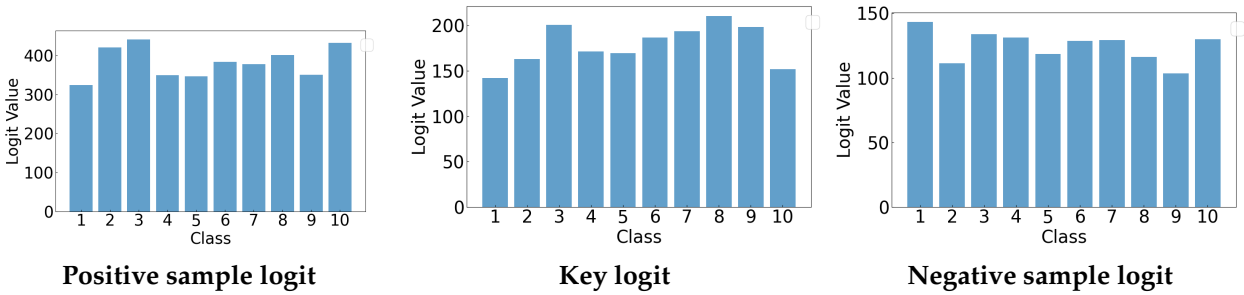


Figure 5. Visualization of logits used for CL-TAA.

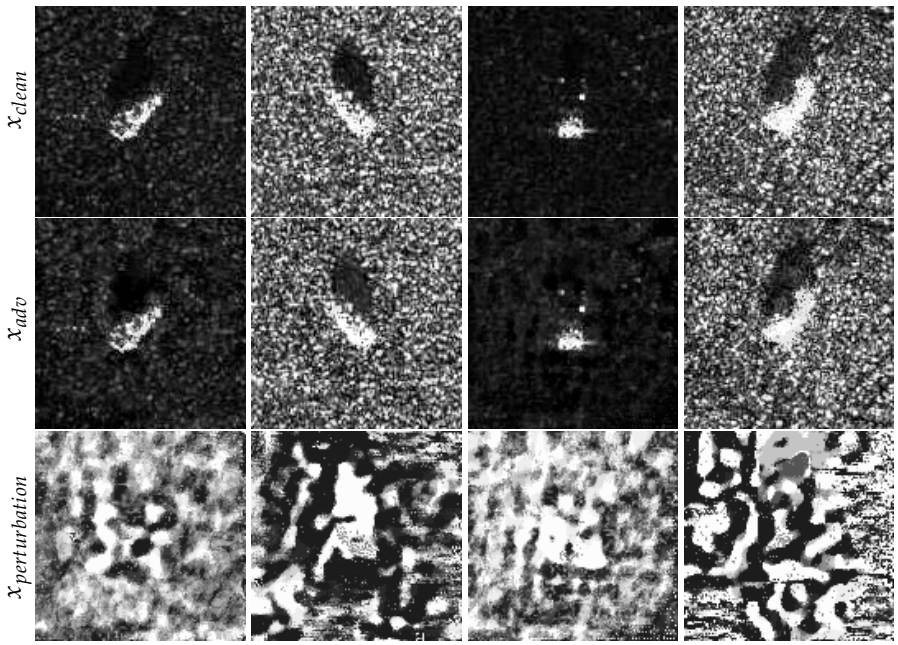


Figure 6. Adversarial examples generated by CL-TAA.

4.3. Comparison with SOTA Techniques

Transferable targeted adversarial attacks (TAAs) have long been considered challenging in computer vision. The difficulty lies in crafting adversarial examples that can both fool a specific model and other unseen models. This challenge is further amplified when the goal is to predict a specific target label rather than mislead the model. As mentioned in Section 2.1, several works [13,14,41,66] have explored ways to improve the transferability

of generated adversarial examples, where the seminal methods include MI-FGSM [13] TI-FGSM [14], and DI-FGSM [41]. Essentially, our proposed CL loss is similar to these three methods because they all regularize the input gradient. We experiment with these three methods, with results shown in Table 7. Surprisingly, we find that the benefits of MI-FGSM and TI-FGSM for improving transferability to black-box models are negligible. For DI-FGSM, we observe a noticeable improvement in black box transferability. While all these three methods are effective for optical images, their performance varies when applied to SAR images, with DI-FGSM yielding the best results. We propose the reason to be that MI-FGSM and TI-FGSM primarily modify the gradients, offering less of an impact on over-fitting for SAR images. Conversely, DI-FGSM directly enhances the diversity of the input SAR images, resulting in reduced over-fitting on the source model. A more detailed investigation of this phenomenon is left for future work. For DI-FGSM, we observe a noticeable improvement in black box transferability, though it is less significant compared to the enhancement achieved by our proposed CL loss.

Table 7. Targeted adversarial attack on MSTAR dataset. We present the results by using the AMS-CNN (%) as the source model.

	Target Models					
Attack	**AlexNet**	**VGG16**	**ResNet18**	**EfficientNet**	**MobileNet**	**ConvNeXt**
I-FGSM	14.9	15.5	13.6	13.2	11.7	15.2
MI-FGSM	15.2	13.1	13.5	11.6	9.2	16.0
TI-FGSM	15.2	14.8	13.4	12.7	12.4	14.9
DI-FGSM	16.8	15.9	13.9	13.2	12.6	15.5
CL-TAA	**17.9**	**20.0**	**17.3**	**14.3**	**15.2**	**17.5**

4.4. Ablation Study

Weight of CL loss. The weight of CL loss is a crucial parameter in determining its contribution to the total loss. Adjusting this weight directly impacts how the CL loss influences the model's training process. A well-balanced weight can enhance performance, while an improper setting may negatively impact it. To identify the optimal weight for maximizing the effectiveness of the CL loss, we conduct an ablation study on this parameter, shown in Table 8. We can observe that the performance of TAA first increases and then decreases with the weight of CL loss increased from 0 to 100. Specifically, the performance gradually increases with the weight from 0 to 0.01 and then decreases from 0.01 to 100. This indicates that a higher weight parameter has a negative impact on the optimization goal, resulting in reduced performance. In our paper, we set the weight of CL loss to 0.01.

Table 8. Effect of CL loss weight. We report the results with different values of regularization parameters λ in the black box setting.

λ	**AlexNet**	**VGG16**	**ResNet18**	**EfficientNet**	**MobileNet**	**ConvNeXt**
0	14.9	15.5	13.6	13.2	11.7	15.2
10^{-2}	**17.9**	**20.0**	**17.3**	**14.3**	**15.2**	**17.5**
1	16.4	18.8	15.7	13.4	13.8	15.7
10^{+2}	15.5	16.9	14.7	12.4	13.3	14.5

Size of negative sample. The negative sample in contrastive learning is designed to push the anchor away, creating a repelling effect, whereas the positive sample attracts the anchor closer. This means that the anchor focuses more on the independent features shared with the positive samples, and the diverse negative sample can ensure the anchor learns a more generalized feature representation. Therefore, we conduct an ablation study on

various negative samples (1, 10, 50, 100), shown in Table 9. We observe that increasing the number of negative samples significantly improve the performance of targeted adversarial attacks (TAAs). It indicates that incorporating diverse negative sample representations benefits TAA training. The variety in negative samples likely helps the adversarial examples to focus on independent features, strengthening their generalization ability. To further increase the diversity of negative sample representations, we adopt the different negative samples for each iteration.

Table 9. Effect of the number of negative samples. We report the results with different numbers of negative samples N under the black box setting.

N	AlexNet	VGG16	ResNet18	EfficientNet	MobileNet	ConvNeXt
1	15.0	16.6	14.9	12.0	12.8	15.5
10	17.6	19.6	16.2	13.1	14.2	16.7
50	17.9	20.0	17.3	14.3	15.2	17.5
100	**18.2**	**20.9**	**17.9**	**14.5**	**15.4**	**17.7**

Temperature. It has been demonstrated that temperature has a significant influence on the performance of contrastive learning [67,68]. An ablation study on the different temperatures was conducted, as shown in Table 10. We observe that the performance of TAA decreases when the temperature is set to a small value. A lower temperature causes the model to focus more on hard negative samples [67,68]. As noted in [68], a smaller temperature equals a smaller number of negative samples. Consequently, using a lower temperature is expected to negatively impact performance when a large number of negative samples is needed, as it reduces the diversity of negative samples. In our work, we set the temperature to 0.1.

Table 10. Effect of the temperature of InfoNCE. We report the results with different InfoNCE temperatures T in the black box setting.

T	AlexNet	VGG16	ResNet18	EfficientNet	MobileNet	ConvNeXt
0.01	16.6	17.7	14.7	12.4	13.7	15.0
0.05	16.7	19.3	16.0	13.4	14.2	16.3
0.1	**17.9**	**20.0**	**17.3**	**14.3**	**15.2**	**17.5**
0.5	16.8	19.9	17.2	14.2	14.5	17.3

Size of perturbation budgets. The perturbation budget is crucial in determining the maximum allowable magnitude. It limits how much the input can be modified while keeping the perturbation imperceptible. Following [48], we experiment with four different perturbation budgets, 4/255, 8/255, 12/255, and 16/255, with the results presented in Table 11. As expected, larger perturbation budgets lead to better performance, but this comes at the cost of the increased visibility of perturbations. It is worth noting that many previous works [14,41] commonly use a perturbation budget of 16/255 as it strikes a balance between effectiveness and subtlety. For this reason, we adopt the same budget in our work to ensure comparability and maintain imperceptibility while maximizing attack performance.

Step size. Our CL-TAA method builds on I-FGSM to update the perturbation, leveraging the benefits of a multi-step attack. The step size plays a critical role in this process, as it significantly impacts the effectiveness of the attack. A larger step size can lead to faster convergence, but it is also hard to find the optimal point. On the other hand, a smaller step size allows for finer adjustments, but it requires more iterations and increases the risk of getting stuck in a local optimum. Therefore, a proper step size is essential for balancing

attack efficiency and accuracy. The results of different step sizes are shown in Table 12, where the step size of 2/255 performs best. By default, we set the step size to 2/255.

Table 11. Effect of perturbation budget. We report the results with different values of the maximum perturbation ϵ under the black box setup.

ϵ	AlexNet	VGG16	ResNet18	EfficientNet	MobileNet	ConvNeXt
4/255	11.8	12.3	11.0	11.8	12.2	11.9
8/255	13.9	16.1	13.0	13.1	12.8	13.6
12/255	16.7	17.3	15.1	14.1	15.0	15.6
16/255	**17.9**	**20.0**	**17.3**	**14.3**	**15.2**	**17.5**

Table 12. Effect of step size. We report the results with different step sizes α under the black box setting.

α	AlexNet	VGG16	ResNet18	EfficientNet	MobileNet	ConvNeXt
1/255	16.0	18.2	16.3	14.1	15.0	17.0
2/255	**17.9**	**20.0**	**17.3**	**14.3**	**15.2**	**17.5**
4/255	17.2	19.4	16.5	13.2	13.8	17.0
6/255	17.0	19.0	15.6	12.6	13.0	16.5

Number of iterations. We investigate the impact of the number of training iterations on the performance of adversarial examples. For this, we experiment with various training iterations and present the results of adversarial examples against AlexNet, VGG16, ResNet, EfficientNet, MobileNet, and ConvNeXt, depicted in Figure 7. It helps us understand how training duration influences the effectiveness of adversarial attacks across different model architectures. We can observe a clear improvement in performance as the number of iterations increases from 1 to 20. However, beyond this point, the performance of transferable TAA saturates or slightly decreases with additional iterations. Since the number of iterations of 20 yields the best performance, we adopt this setting in our experiments.

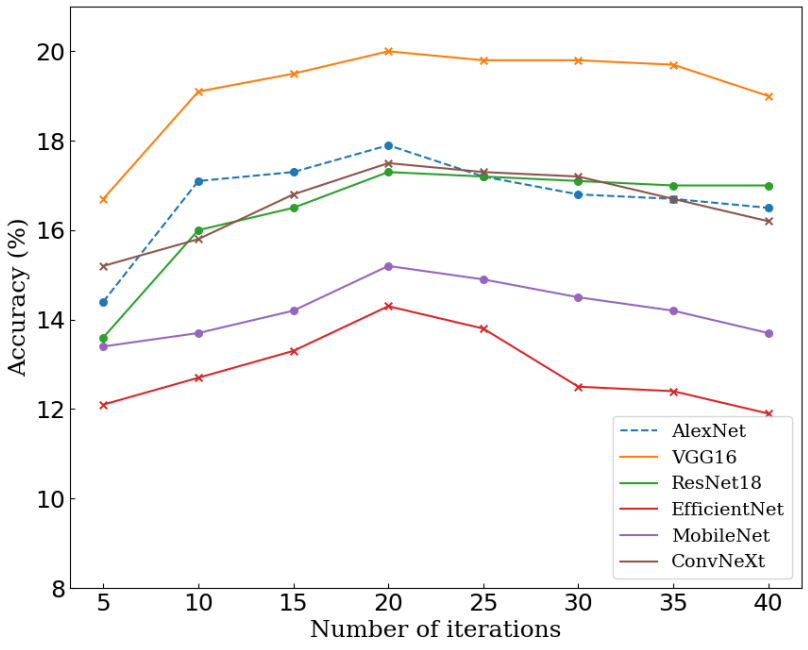


Figure 7. Effect of the number of iterations. We report the results of different iterations under the black box setting by using AMS-CNN as the source model.

5. Conclusions

This work offers the first comprehensive study on TAA against SAR target recognition models in a black box setting, where the attacker has no access to these models. Given the limited generalization of adversarial examples generated by the classical I-FGSM method, we propose leveraging CL to enhance their generalization and improve their effectiveness. Unlike the classical CL used for pre-training, we apply it during the training stage by leveraging the logits of different samples. Specifically, we use the logit of the anchor sample to attract the positive sample while repelling the negative sample, significantly improving the transferability of adversarial examples. Extensive experiments have confirmed the high transferability and effectiveness of the proposed CL-TAA for TAA. The success of our attack method highlights the importance of developing adversarially robust SAR target recognition models.

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

Funding: This research was partly supported by Start-up Funding for Newly Introduced Talents in Shenzhen (CA11409031), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2022-II220078, Explainable Logical Reasoning for Medical Knowledge Generation) and under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2023-00259004) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Data Availability Statement: The research data is available at https://www.sdms.afrl.af.mil/index.php?collection=mstar&page=targets (accessed on 1 October 2024).

Acknowledgments: We thank the above grants for providing help in APC. We also thank the reviewers and editors for their constructive comments in improving the quality of our work.

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

Abbreviations

The following abbreviations are used in this manuscript:

SAR	Synthetic aperture radar
TAA	Targeted adversarial attack
CL	Contrastive learning
CL-TAA	Contrastive learning-based targeted adversarial attack
DCNN	deep convolutional neural network
non-TAA	Non-targeted adversarial attack
FGSM	Fast Gradient Sign Method
PGD	Projected Gradient Descent
I-FGSM	Iterative-Fast Gradient Sign Method
MI-FGSM	Momentum Iterative–Fast Gradient Sign Method
TI-FGSM	Translation Iterative–Fast Gradient Sign Method
DI-FGSM	Diverse-Input-Fast Gradient Sign Method
MoCo	Momentum Contrast
LRN	Local response normalization
VGG	Visual Geometry Group
DNNs	Deep Neural Networks
MSTAR	Moving and Stationary Target Acquisition and 335 Recognition
AFRL	Air Force Research Laboratory
DARPA	Defence Advanced Research Projects Agency

References

1. Chen, S.; Wang, H. SAR target recognition based on deep learning. In Proceedings of the 2014 International Conference on Data Science and Advanced Analytics (DSAA), Shanghai, China, 30 October–1 November 2014; pp. 541–547.
2. Gao, F.; Yue, Z.; Wang, J.; Sun, J.; Yang, E.; Zhou, H. A novel active semisupervised convolutional neural network algorithm for SAR image recognition. *Comput. Intell. Neurosci.* **2017**, *2017*, 3105053. [CrossRef]
3. Anagnostopoulos, G.C. SVM-based target recognition from synthetic aperture radar images using target region outline descriptors. *Nonlinear Anal. Theory Methods Appl.* **2009**, *71*, e2934–e2939. [CrossRef]
4. Cha, M.; Majumdar, A.; Kung, H.T.; Barber, J. Improving SAR automatic target recognition using simulated images under deep residual refinements. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2606–2610.
5. Zheng, S.; Hao, X.; Zhang, C.; Zhou, W.; Duan, L. Towards Lightweight Deep Classification for Low-Resolution Synthetic Aperture Radar (SAR) Images: An Empirical Study. *Remote Sens.* **2023**, *15*, 3312. [CrossRef]
6. Yi, G.; Hao, X.; Yan, X.; Dai, J.; Liu, Y.; Han, Y. Automatic modulation recognition of radiation source signals based on two-dimensional data matrix and improved residual neural network. *Def. Technol.* **2024**, *33*, 364–373. [CrossRef]
7. Yi, G.; Hao, X.; Yan, X.; Wang, J.; Dai, J. Automatic Modulation Recognition for Radio Frequency Proximity Sensor Signals Based on Masked Autoencoders and Transfer Learning. *IEEE Trans. Aerosp. Electron. Syst.* **2024**, 60, 8700–8712. [CrossRef]
8. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 1–9.
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
10. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
12. Szegedy, C. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
13. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9185–9193.
14. Dong, Y.; Pang, T.; Su, H.; Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4312–4321.
15. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
16. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
17. Biggio, B.; Fumera, G.; Roli, F. Security evaluation of pattern classifiers under attack. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 984–996. [CrossRef]
18. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv* **2017**, arXiv:1712.04248.
19. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box adversarial attacks with limited queries and information. In Proceedings of the International Conference on Machine Learning, Beijing, China, 14–16 November 2018; pp. 2137–2146.
20. Ilyas, A.; Engstrom, L.; Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv* **2018**, arXiv:1807.07978.
21. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrücken, Germany, 21–24 March 2016; pp. 372–387.
22. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2574–2582.
23. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
24. Mądry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *Stat* **2017**, *1050* .
25. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1765–1773.
26. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. Adversarial attacks and defences: A survey. *arXiv* **2018**, arXiv:1810.00069. [CrossRef]

27. Croce, F.; Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the International Conference on Machine Learning, Shenzhen, China, 15–17 February 2020; pp. 2206–2216.
28. Zhang, S.; Zuo, D.; Yang, Y.; Zhang, X. A transferable adversarial belief attack with salient region perturbation restriction. *IEEE Trans. Multimed.* **2022**, *25*, 4296–4306. [CrossRef]
29. Zhang, C.; Benz, P.; Imtiaz, T.; Kweon, I.S. Understanding adversarial examples from the mutual influence of images and perturbations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 14521–14530.
30. Zhang, C.; Benz, P.; Karjauv, A.; Cho, J.W.; Zhang, K.; Kweon, I.S. Investigating top-k white-box and transferable black box attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15085–15094.
31. Wang, X.; Chen, H.; Sun, P.; Li, J.; Zhang, A.; Liu, W.; Jiang, N. AdvST: Generating Unrestricted Adversarial Images via Style Transfer. *IEEE Trans. Multimed.* **2023**, *26*, 4846–4858. [CrossRef]
32. Wang, J.; Zhao, J.; Yin, Q.; Luo, X.; Zheng, Y.; Shi, Y.Q.; Jha, S.K. SmsNet: A new deep convolutional neural network model for adversarial example detection. *IEEE Trans. Multimed.* **2021**, *24*, 230–244. [CrossRef]
33. Cheng, Y.; Guo, Q.; Juefei-Xu, F.; Lin, S.W.; Feng, W.; Lin, W.; Liu, Y. Pasadena: Perceptually aware and stealthy adversarial denoise attack. *IEEE Trans. Multimed.* **2021**, *24*, 3807–3822. [CrossRef]
34. Rathore, P.; Basak, A.; Nistala, S.H.; Runkana, V. Untargeted, targeted and universal adversarial attacks and defenses on time series. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
35. 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]
36. 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]
37. Du, C.; Zhang, L. Adversarial attack for SAR target recognition based on UNet-generative adversarial network. *Remote Sens.* **2021**, *13*, 4358. [CrossRef]
38. 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. [CrossRef]
39. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial machine learning at scale. *arXiv* **2016**, arXiv:1611.01236.
40. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
41. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2730–2739.
42. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 776–794.
43. Dosovitskiy, A.; Springenberg, J.T.; Riedmiller, M.; Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. [CrossRef]
44. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3733–3742.
45. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 9729–9738.
46. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Shenzhen, China, 15–17 February 2020; pp. 1597–1607.
47. Zhuang, C.; Zhai, A.L.; Yamins, D. Local aggregation for unsupervised learning of visual embeddings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6002–6012.
48. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
49. Polyak, B.T. Some methods of speeding up the convergence of iteration methods. *Ussr Comput. Math. Math. Phys.* **1964**, *4*, 1–17. [CrossRef]
50. Duch, W.; Korczak, J. Optimization and global minimization methods suitable for neural networks. *Neural Comput. Surv.* **1998**, *2*, 163–212.
51. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; pp. 1139–1147.
52. Lin, J.; Song, C.; He, K.; Wang, L.; Hopcroft, J.E. Nesterov accelerated gradient and scale invariance for adversarial attacks. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 26–30 April 2020.

53. Gao, L.; Zhang, Q.; Song, J.; Liu, X.; Shen, H.T. Patch-wise attack for fooling deep neural network. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 307–322.
54. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 630–645.
55. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. Mitigating adversarial effects through randomization. *arXiv* **2017**, arXiv:1711.01991.
56. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
57. Kechagias-Stamatis, O.; Aouf, N.; Belloni, C. SAR automatic target recognition based on convolutional neural networks. In Proceedings of the International Conference on Radar Systems (Radar 2017), Belfast, UK, 23–26 October 2017; IET: Stevenage, UK, 2017; pp. 1–4.
58. Shao, J.; Qu, C.; Li, J. A performance analysis of convolutional neural network models in SAR target recognition. In Proceedings of the 2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), Beijing, China, 13–14 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
59. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6609–6618.
60. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
61. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. *arXiv* **2021**, arXiv:2103.03101.
62. Zheng, S.; Zhang, C.; Hao, X. Black-box targeted adversarial attack on segment anything (sam). *arXiv* **2023**, arXiv:2310.10010. [CrossRef]
63. Keydel, E.R.; Lee, S.W.; Moore, J.T. MSTAR extended operating conditions: A tutorial. *Algorithms Synth. Aperture Radar Imag. III* **1996**, *2757*, 228–242.
64. 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, Orlando, FL, USA, 13–17 April 1998; SPIE: San Francisco, CA, USA, 1998; Volume 3370, pp. 566–573.
65. 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 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
66. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 3 November 2017; pp. 15–26.
67. Wang, F.; Liu, H. Understanding the Behaviour of Contrastive Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021.
68. Zhang, C.; Zhang, K.; Pham, T.X.; Yoo, C.; Kweon, I.S. Dual Temperature Helps Contrastive Learning Without Many Negative Samples: Towards Understanding and Simplifying MoCo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.

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.