Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056075

ARTICLE

Adversarial Defense Technology for Small Infrared Targets

Tongan Yu[1], Yali Xue[1,*], Yiming He[1], Shan Cui[2] and Jun Hong[2]

[1]College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, 210000, China

[2]Shanghai Electro-Mechanical Engineering Institute, Shanghai, 201109, China

*Corresponding Author: Yali Xue. Email: xueyali@nuaa.edu.cn

Received: 13 July 2024 Accepted: 03 September 2024 Published: 15 October 2024

ABSTRACT

With the rapid development of deep learning-based detection algorithms, deep learning is widely used in the field of infrared small target detection. However, well-designed adversarial samples can fool human visual perception, directly causing a serious decline in the detection quality of the recognition model. In this paper, an adversarial defense technology for small infrared targets is proposed to improve model robustness. The adversarial samples with strong migration can not only improve the generalization of defense technology, but also save the training cost. Therefore, this study adopts the concept of maximizing multidimensional feature distortion, applying noise to clean samples to serve as subsequent training samples. On this basis, this study proposes an inverse perturbation elimination method based on Generative Adversarial Networks (GAN) to realize the adversarial defense, and design the generator and discriminator for infrared small targets, aiming to make both of them compete with each other to continuously improve the performance of the model, find out the commonalities and differences between the adversarial samples and the original samples. Through experimental verification, our defense algorithm is not only able to cope with multiple attacks but also performs well on different recognition models compared to commonly used defense algorithms, making it a plug-and-play efficient adversarial defense technique.

KEYWORDS

Adversarial defense; adversarial robustness; small infrared targets; transferable perturbation; GAN

1 Introduction

In recent years, deep learning has seen rapid advancements across various domains, including computer vision, natural language processing, image generation, and autonomous driving. However, Szegedy et al. [1] identified vulnerabilities in neural network-based classification tasks, highlighting the susceptibility of deep neural networks to adversarial samples, which can significantly degrade performance. This vulnerability is particularly concerning in applications such as infrared small target detection, which is critical in many fields and is often challenged by adversarial attacks. To enhance the security of infrared recognition models, it is imperative to incorporate counter-defense technologies.

Among the array of available defense mechanisms, the primary strategies include adversarial training [2], improvements in model architecture [3], and detection of adversarial samples [4]. Nevertheless, the effectiveness of these techniques is generally confined to specific attack algorithms, and

their defensive performance heavily relies on the depth of understanding of the recognition model. An approach that transcends the dependency on specific recognition models is adversarial disturbance elimination via GAN. GAN leverage their potent generative capabilities to process diverse input samples effectively. For instance, Defense-GAN was initially introduced by Samangouei et al. [5], and more recently, Collaborative Defense-GAN was proposed by Laykaviriyakul et al. [6].

1.1 Purpose

Due to the minimal proportion of target pixels and the subtle distinction between the background and foreground, most attack algorithms can efficiently compromise the recognition model. Consequently, it is imperative to address the following challenges: (1) The design of a defense architecture tailored for infrared small target detection, leveraging existing mature defense technologies; (2) The enhancement of the defense model's generalizability to effectively counter adversarial samples generated by diverse attack algorithms.

In this study, we draw inspiration from the concept of dynamic noise injection proposed by Liang et al. [7], and introduce a transfer perturbation algorithm centered around multidimensional feature distortion. This algorithm aims to supply high-quality training samples for GAN, building upon the work of Naseer et al. [8]. Furthermore, we develop a high-frequency feature extraction module, incorporating a variable radius Fourier transform, channel and pixel-level attention mechanism, and a reconstruction loss function. This module enables the generator to fundamentally restore the perturbation information present in adversarial samples targeting infrared small targets.

1.2 Contributions

(1) We introduce an adversarial defense technique tailored for infrared small target recognition models. This approach employs GAN to train generators endowed with superior denoising capabilities.

(2) Leveraging the concept of multidimensional feature distortion, we generate high-quality, cross-task adversarial samples for GAN training. This strategy significantly reduces training costs and enhances the generality of the generator.

(3) We develop a generator that focuses on high-frequency features, paired with a discriminator that utilizes a channel and pixel-level attention mechanism. Additionally, we implement a loss function specifically designed to guide the GAN's updates.

The efficacy of our method is demonstrated through comparative analysis with other leading defense algorithms, across a variety of attack scenarios.

2 Related Works

2.1 Adversarial Attack Methods

Adversarial samples are specifically crafted input samples designed to deceive recognition models. This section primarily discusses various methodologies employed to generate such adversarial samples.

Initial attack strategies primarily exploited the nonlinearity of neural networks to generate adversarial samples. For instance, the L-BFGS method [1] generates adversarial samples by calculating the second derivative of the gradient and the objective function. With the advent of cutting-edge countermeasures, specialized attack algorithms for different vulnerabilities are emerging. As a seminal attack algorithm, the Fast Gradient Sign Method (FGSM) [9] has been extensively studied and refined by numerous researchers. DeepFool [10] estimates the minimal distance required to alter a sample's classification by linear approximation, employing FGSM's underlying principles to generate

adversarial samples. Madry et al. [11] introduced the Projected Gradient Descent (PGD) algorithm, which iterates over multiple gradients to identify optimal adversarial samples. Building on FGSM, the Basic Iterative Method (BIM) [12] employs multiple iterations to better simulate the effects of physical attacks. Dong et al. [13] enhanced BIM's generalization by integrating momentum into the process. Concurrently with these FGSM-based approaches, several other notable attack algorithms have been developed. Papernot [14] devised a first-order attack algorithm aimed specifically at defeating defensive distillation strategies [15]. The Carlini-Wagner (CW) attack [16] optimizes an objective function tailored to the target model, minimizing the perturbation introduced to the original image to generate adversarial samples. Since the inception of adversarial sample research, numerous innovative algorithms have been proposed. For example, Zhang et al. [17] proposed a neuron-based feature level attack that manipulates specific neurons within the model to induce misclassification; Wang et al. [18] introduced the Boundary Attack, which incrementally moves an input sample towards the model's decision boundary, causing misclassification with minimal perturbations near the boundary.

2.2 Adversarial Defense Methods

In order to deal with the influence of adversarial samples on recognition models, many scholars have focused their research on adversarial defense. Defense technology can be divided into the following two categories: model-based defense technology and data-based defense technology.

Model-based defense mechanisms aim to desensitize the model to adversarial perturbations. Adversarial training, a prominent defense algorithm, was introduced alongside the FGSM and PGD algorithms, offering defense strategies against these specific types of attacks. Shafahi et al. [19] introduced the Free-AT method to diminish computational overhead by reusing gradient information. Zhang et al. [20] developed TRADES, which dissects robustness error into natural error and boundary error, elucidating the trade-off between accuracy and robustness in classification problems, and introduced novel loss functions. Techniques such as ANP [21] and mixup [22] defend by enriching the diversity of training data. Furthermore, model-based defense algorithms enhance the model architecture to bolster defense; for instance, defensive distillation reduces gradient output through secondary training, thereby curtailing the generation of adversarial samples. SLL [23] employs decision boundary smoothing to counter model-specific attack algorithms.

Conversely, data-based defense technologies focus on diminishing disruptive information within data. Currently, two principal defense technologies dominate this realm. The first involves the detection of adversarial samples, with the BUE algorithm [4] distinguishing and amplifying feature discrepancies for adversarial sample detection. The second approach seeks to minimize the disturbance introduced by adversarial samples. Liao et al. [24] proposed a feature-level method for noise removal; Dubey et al. [25] reverted perturbations back to the nonadversarial manifold using the K-nearest neighbor method. Image reconstruction, a rapidly evolving field, has also yielded notable advancements in defense technologies. Through the reconstruction of benign samples, reconstruction algorithms supply the model with undisturbed samples, facilitating effective defense. GAN play a pivotal role in this context: initially, APE-GAN by Jin et al. [26] and Defense-GAN by Samangouei et al. [5] demonstrated defense capabilities against most attacks, albeit not concurrently. More recently, Zhao et al. [27] refined the GAN framework and introduced a twin synthesizer structure to eliminate disturbances. Laykaviriyakul et al. [6] drew inspiration from the cyclic GAN structure, applying it to adversarial sample elimination for efficient disturbance mitigation.

3 Methodology

3.1 General Frame

In this paper, we propose an adversarial defense technique for small infrared targets, which is based on the GAN framework. The overall framework is illustrated in Fig. 1.

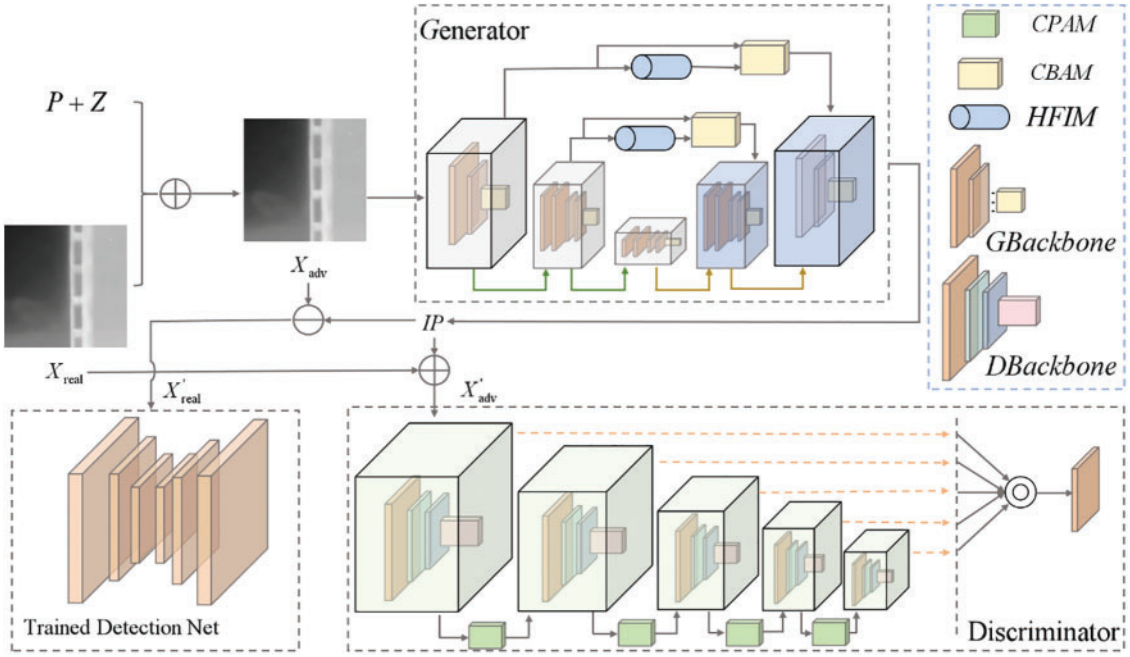

Figure 1: Adversarial defensive defense-based training framework

Our method comprises four main components: 1) Initially, transferable perturbations (P) are generated through the maximization of multidimensional feature distortion, forming adversarial samples (X_{adv}) together with real samples (X_{real}) as the output of the generator. The input samples will simultaneously have a perturbation budget between 0 and 1 with noise z making it dynamic; 2) Generate inverse perturbation (IP) more accurately using a generator capable of extracting high-frequency information combined with Convolutional Block Attention Module (CBAM) [28]; 3) Fake adversarial samples (X'_{adv}), composed of inverse perturbations and real samples, are discriminated by the discriminator to determine whether they belong to the same domain as the real adversarial samples. The discriminator possesses pixel-level discrimination capability, which aligns closely with the characteristics of infrared small target images; 4) Fake real samples (X'_{real}), formed by combining inverse perturbations with real adversarial samples, are fed into a pre-trained recognition network. The restoration degree of fake real samples is judged by a pixel-level Intersection over Union (IOU) loss. We will describe these four components in the following sections.

3.2 Transferable Perturbation

Currently, prevalent attack methodologies share a common characteristic: the adversarial samples they generate induce significant distortions at the feature layer of the target network, thereby hindering accurate identification by the target network. In Fig. 2, we used the Mean Squared Error (MSE) between the adversarial sample and the real sample on this layer as a measure of feature distortion, along with the IOU value of as an indicator of adversarial sample quality. From the trend of the curves

in the figure, it can be seen that there is a positive relationship between the degree of feature distortion and the strength of the attack on the antagonistic samples. Therefore, with the goal of maximizing the degree of feature distortion, we propose a method for generating transferable perturbations based on multidimensional feature distortion (MFDP).

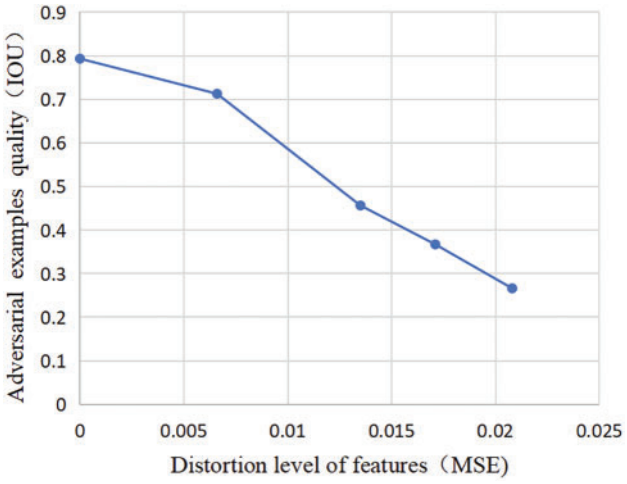

Figure 2: Diagaph of adversarial sample quality and degree of feature distortion

This method produces high-quality adversarial samples that are capable of spanning across tasks and networks, aimed at enhancing the generalizability of our defense algorithm.The self-supervised perturbation generation approach proposed by Naseer et al. [8] focuses predominantly on feature distortion analysis from the network's deep features, yet it overlooks the variance in distortion across multidimensional features. By simultaneously perturbing multiple feature dimensions, adversarial samples can be more covertly integrated into the data, rendering the model's detection and countermeasures more challenging.

Moreover, considering the small pixel proportion of small infrared targets, adversarial samples generated by mainstream attack algorithms often contain isolated noise points that resemble target features. Therefore, we aim for the adversarial samples to retain partial small target features of the deep features. From the output of the real samples on the deep features, areas likely to be targets are selected. Randomly, 1–3 areas outside of this region in the deep feature output image of the sample are chosen. The generation of adversarial samples is then guided by the objective of minimizing the degree of feature distortion between these areas. Adversarial samples are subjected to random perturbations with pixel values ranging between (0,1) as input samples for GAN training. This approach is adopted because, although perturbations generated through multidimensional feature distortion possess strong transferability, the degree of distortion produced on the target network's feature layer by different attack methods can vary.

The generation method of transfer disturbance is shown in Fig. 3. In our defined loss function, α, β, and γ are hyperparameters, set to 0.2, 0.6, and 0.2, respectively. The loss function of the four feature layers can be explained by the following formula:

$$L_{layeri} = d\,(Layeri\,(X_{adv}), Layeri\,(X_{real})) \tag{1}$$

This loss function represents the difference between the output of X_{adv} and X_{real} in the i th feature layer, i takes 1–4; the fourth layer is the difference between parts of the output of the feature layer.

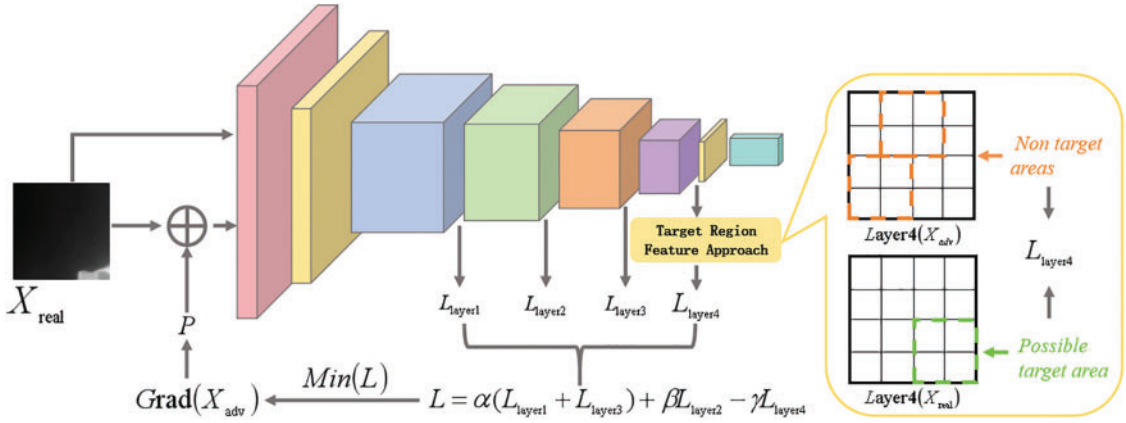

Figure 3: Tranfer disturbance based on multidimensional feature distortion

3.3 Reverse Disturbance

3.3.1 Generator

To extract features from high-quality adversarial samples and restore the perturbations within them, we have designed a generator structure based on high-frequency features, incorporating an encoder-decoder architecture. In infrared images, high-frequency features often contain more noise information; hence, we designed the high-frequency feature extraction module (HFIM) structure, as shown in Fig. 4. The parameter Radius is involved in GAN training, allowing for the dynamic adjustment of Radius to extract high-quality high-frequency feature information. This generator structure will also serve as a denoising framework within the recognition network, filtering out perturbations from the input samples.

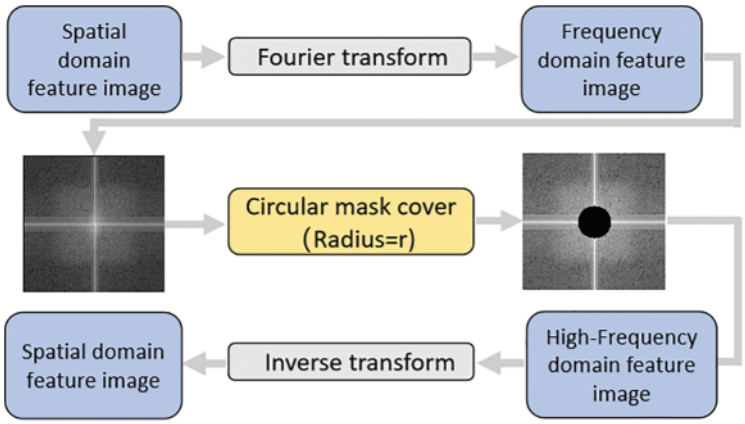

Figure 4: HFIM based on variable fourier radius

The process of the generator: Adversarial samples first enter the encoding structure, where features across different dimensions are extracted. Features of different dimensions are transformed into high-frequency features after HFIM. After this, an attention mechanism is utilized to feed the feature fusion map of different dimensions into the decoder. Finally, the decoder then gradually restores the perturbations based on the feature fusion information, outputting an inverse perturbation image.

3.3.2 Discriminator

The discriminator, tasked with discerning whether the inverse perturbations generated by the generator share the same content as the original perturbations, can guide the generation of inverse perturbations. We believe that directly discriminating between inverse and original perturbations is challenging; adversarial samples, however, contain not only the information of the original samples but also perturbation information, providing the discriminator with more comprehensive insights. Therefore, discriminating the authenticity of adversarial samples is more appropriate. We design the discriminator as a fully convolutional network, utilizing channel and pixel-level attention, as shown in Fig. 5, to discern pixel differences, ultimately resulting in the production of a scoring map for the adversarial samples. Each pixel's confidence score ranges between 0 and 1, with scores closer to 1 indicating authenticity and vice versa. The advantage of this design lies in the pixel-level detection required for small infrared targets; the smaller the pixel difference in restored samples, the higher the defense efficiency.

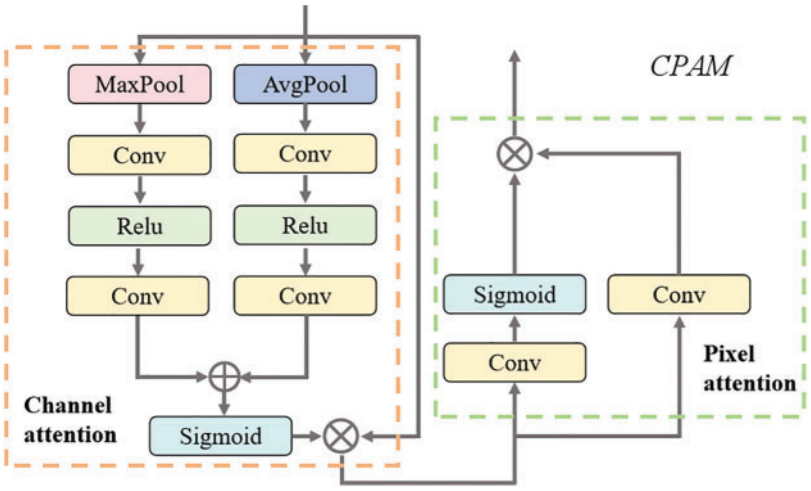

Figure 5: The channel pixel attention mechanism (CPAM)

3.3.3 Loss Function

The loss function serves as a crucial guide for the generator to produce correct inverse perturbations and for the discriminator to accurately judge adversarial samples. Its design significantly determines the quality of the generated samples, and a well-designed loss function can guide the generator and discriminator towards effective adversarial training. Therefore, we provide a detailed introduction to the loss functions we employed.

Generator loss function (L_G): The loss function of the generator is as follows, comprising four components:

$$L_G = L_{xreal} + L_{xadv} + L_{GD} + L_F \tag{2}$$

$$L_{xreal} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\left\|xreal_i - xreal_i'\right\|_2^2 & xreal_i > k\\ \left\|xreal_i - xreal_i'\right\|_1 & xreal_i < k\end{cases} \tag{3}$$

$$L_{xadv} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}\left\|xadv_i - xadv_i'\right\|_2^2 & xadv_i > k\\ \left\|xadv_i - xadv_i'\right\|_1 & xadv_i < k\end{cases} \tag{4}$$

$$L_{GD} = \left\| D\left(xadv'\right) - 1 \right\|_1 \tag{5}$$

$$L_F = 1 - \frac{1}{N} \sum\nolimits_N \frac{\sum_{pixels} F\left(xreal'\right) mask}{\sum_{pixels} F\left(xreal'\right) + mask - F\left(xreal'\right) mask} \tag{6}$$

The loss function Lxreal calculates the difference between xreal and x′real by dividing xreal and x'real into n regions. If the pixels in the first region exhibit significant differences, we will impose a heavier penalty. Lxadv is implemented in a similar manner. The advantages of this approach are as follows: compared to the direct application of MSE or MAE losses, our loss function can more accurately guide the generator towards the correct update direction and expedite the training process. For instance, in the case of local outliers, using MSE alone would amplify the overall loss, leading to substantial changes in model weight updates; relying solely on MAE could result in the loss overlooking local outliers, and the model weights might continue updating in the direction of the last update. By employing L_{GD} as the loss function for the discriminator's feedback, we aim for x′adv to deceive the discriminator, such that the closer the output scoring matrix is to the identity matrix, the better. LF calculates the average intersection ratio between the output of the recognition network F and the labeled image to guide the generator training.

Discriminant loss function (L_D): Because we want the discriminant to distinguish between x′adv and x′adv, that is, D(xadv) is true and D(x'adv) is false. The formula is shown below:

$$L_D = L_{D1} + L_{D2} \tag{7}$$

$$L_{D1} = \left\| D\left(xadv\right) - 1 \right\|_1 \tag{8}$$

$$L_{D2} = \left\| D\left(xadv'\right) - 0 \right\|_1 \tag{9}$$

4 Experiments

In this section, we will validate the proposed strategy. Considering that in the infrared small target samples, the adversarial samples generated by small disturbances have no significant attack effect on the recognition model. The differences in visual perception of the adversarial samples generated by too large disturbances will be particularly obvious, which does not have much research value.In this, we choose 8/255 as the perturbation budget in the experiment, such a perturbation budget not only has a good attack effect, but also has only a slight deviation in visual perception.

4.1 Dataset

We used the dataset from literature [29]. There were 1367 images, single target images accounted for 65.97%, dual target images for 30.71%, and multi-target images for 3.32%, and all the targets met the definition of small targets. We divided the data set into training set, validation set, and test set with a ratio of 6:2:2. When training the model, the samples in the training set are randomly flipped and the brightness enhancement with 50% probability to achieve the purpose of data enhancement.

4.2 Training Environment

We verified the proposed algorithm based on the PyTorch1.13.1 platform, with a GPU of NVIDIA GeForce RTX 3070 and an OS system as Windows 11, 8 GB RAM.

4.3 MFDP Experiment

To validate the efficacy of transferable perturbations generated through multidimensional feature distortion, we utilized ResNet18 as the source network to guide the generation of perturbations at various depths. These perturbations were then applied to prevalent deep learning recognition models (including ALC-Net [30], UIU-Net [31], DNA-Net [32], CNU-Net). Under the same perturbation budget conditions, using the mean Intersection over Union (mIOU) as the metric, it can be deduced from Table 1 that the impact of feature distortion at different depths on the perturbation is substantial.

Table 1: Attack effectiveness of perturbations generated by feature distortion at different depths

Loss	ALC-Net	UIU-Net	DNA-Net	CNU-Net
Origin	**0.7469**	**0.5109**	**0.6450**	**0.7861**
LB1	0.2911	0.0718	0.1966	0.1558
LB2	0.5171	0.0834	0.1876	0.2428
LB3	0.4515	0.1144	0.2306	0.2618
LB4	0.6505	0.2473	0.4546	0.4589
Ours	**0.4881**	**0.0847**	**0.2050**	**0.2447**

We can analyze through Fig. 6 combined with the data in Table 1 to get that the attack effect of LB1 and LB3 is relatively excellent, but the cost is that they will be relatively obvious from the visual perception; while LB2 can reduce the visual difference caused by the noise on the basis of guaranteeing the attack effect, so we assigned more weights to LB2; although LB4 is the least obvious from the visual perception, but similarly it is also the least effective in attacking, so we do not expect LB4 to play a positive role in the generation of perturbations, instead we expect the perturbations to have a reduced feature difference on B4. Our method generates adversarial samples in graphs with only a smaller visual difference than LB4 and guarantees a certain attack strength.

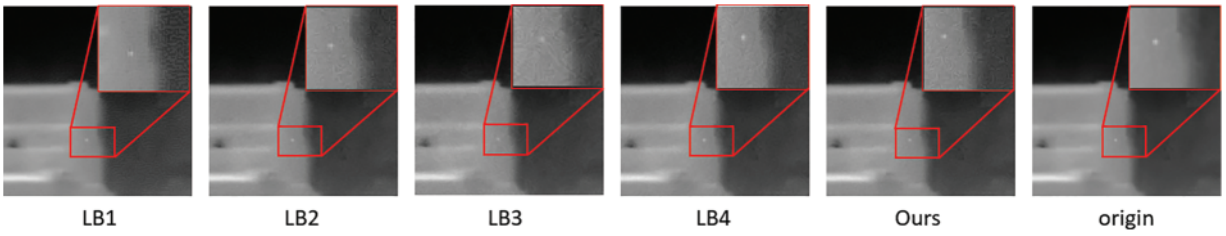

Figure 6: Comparative images of adversarial samples

4.4 Denfense Experiments

In this section, we will validate the effectiveness of our defense strategy. MIOU, precision, and recall will serve as the evaluation metrics. These three indicators are instrumental in assessing the detection quality of adversarial samples post-defense when fed into the recognition model. To verify the generalizability of our defense strategy, we not only utilized adversarial samples generated by the MFDP method but also selected adversarial samples produced by mainstream white-box and black-box attack methods: PGD [11], FGSM [9], BIM [12], ZOO [33], Boundary [34] and CW [16]. Both PGD and BIM are both done in 10 iterations. Both ZOO and Boundary are both done in 60 iterations. The

maximum number of iterations for CW is 100. The process of generating adversarial samples does not take the means associated with data augmentation to ensure the rigor of the experiment.

Initially, we validate the modules designed for small infrared targets within the GAN to ensure all designs are effective. We conducted ablation experiments by removing HFIM and CPAM. Scheme A and Scheme B represent the scenarios without HFIM and without CPAM, respectively, while Scheme C represents the scenario where both HFIM and CPAM are removed. "P" represents Precision, and "R" represents Recall. The results are shown in Table 2. It is observed that our defense strategy maintains a relatively high level of effectiveness against various types of attacks. Without HFIM, a decrease in precision is not significantly noticeable, with occasional increases observed; however, a more pronounced decrease in recall is evident. This outcome is attributed to HFIM making the generator more sensitive to high-frequency feature points, thereby restoring target features more clearly and naturally resulting in a higher recall rate. If CPAM is excluded, there is a noticeable decline in discriminator performance, which consequently degrades the generator's performance and slightly diminishes the defense effectiveness.

Table 2: Results of ablation experiments

Attack	Metrics	A (without HFIM)	B (without CPAM)	C (without HFIM and CPAM)	Ours
MFDP	P	85.81	86.07	83.15	**86.09**
	R	70.21	71.21	70.65	**72.07**
	mIOU	63.68	64.71	62.49	**65.35**
PGD	P	85.95	**87.23**	87.04	86.36
	R	72.82	72.89	72.35	**74.43**
	mIOU	67.07	67.78	67.53	**68.78**
FGSM	P	87.42	**87.65**	87.43	86.54
	R	72.03	**73.38**	72.25	72.74
	mIOU	66.91	**67.98**	67.10	67.28
BIM	P	86.64	87.22	86.93	**87.23**
	R	70.19	71.2	70.98	**72.05**
	mIOU	64.21	65.25	64.85	**66.12**
CW	P	80.02	81.06	78.87	**81.12**
	R	55.53	54.32	51.62	**58.04**
	mIOU	49.63	47.66	44.39	**52.03**
ZOO	P	82.56	82.98	**84.23**	83.48
	R	69.25	69.01	70.08	**70.84**
	mIOU	62.19	62.18	62.89	**63.07**
Boundary	P	82.77	82.60	82.73	**82.78**
	R	68.32	67.91	67.66	**68.44**
	mIOU	60.88	60.15	59.97	**61.38**

We compare our method with several mainstream defense strategies (The methods used for comparison were all modified based on the original network architecture, tailored to our dataset

and model structure), and verify the detection accuracy of the identification model under different attack algorithms. In Table 3, we can observe that the methods listed are capable of defending against various types of attacks to a certain extent. This is because both Defense-GAN and Collaborative Defense-GAN are trained with adversarial samples generated by MFDP, which corroborates our proposed method for generating transferable perturbations can train cross-task defense algorithms. Although Learn2Perturb is trained using original samples, it essentially applies perturbations to different modules of the model, achieving high-quality adversarial defense. Comparative experiments reveal that although our method may slightly underperform on certain metrics, it overall surpasses other algorithms in defense performance against various attack algorithms. Especially in the face of black-box attacks, our method has more advantages and is superior to other methods in various indicators. And in terms of mIOU, our method is always optimal, which shows that compared with other methods, our method maximally restores the feature information of the sample.

Table 3: Comparison of experimental results

Attack	Metrics	No-Defense	D-GAN [5]	L2P [35]	CD-GAN [6]	Ours
MFDP	P	80.26	80.93	85.58	82.02	**86.09**
	R	28.74	61.50	71.89	67.95	**72.07**
	mIOU	24.47	53.62	65.14	59.83	**65.35**
PGD	P	78.16	**87.12**	83.32	85.49	86.36
	R	9.31	72.28	68.71	70.89	**74.43**
	mIOU	9.35	67.54	61.95	63.86	**68.78**
FGSM	P	81.67	86.07	83.33	85.65	**86.54**
	R	26.01	**72. 76**	69.06	70.77	72.74
	mIOU	25.28	67.02	62.41	63.76	**67.28**
BIM	P	74.39	86.64	84.18	82.80	**87.23**
	R	30.53	71.66	70.73	67.84	**72.05**
	mIOU	28.08	65.48	63.27	60.08	**66.12**
CW	P	7.51	44.02	**81.27**	54.55	81.12
	R	2.51	**58.05**	57.34	50.65	58.04
	mIOU	1.98	33.33	51.12	35.83	**52.03**
ZOO	P	68.31	80.36	83.26	81.21	**83.48**
	R	33.26	64.66	70.69	64.92	**70.84**
	mIOU	26.68	54.37	62.58	56.87	**63.07**
Boundary	P	74.35	80.03	81.89	81.34	**82.78**
	R	26.99	65.06	66.18	65.12	**68.44**
	mIOU	21.04	54.72	58.77	57.64	**61.38**

The aforementioned experiments were conducted based on the same recognition model. Next, we integrate our generator into the recognition model used in the MFDP experiment, and again add a recognition model MDvsFA [36] under a different framework to evaluate the generalization of our defense strategy. The generator acts during the data loading process, so it is guaranteed to remain constant under the condition of combining arbitrary recognition models. Keeping the attack

methods consistent, we compare the recognition accuracy under attack, with and without defense, across various recognition models. In Table 4, we observe that, regardless of the recognition model, our defense strategy achieves a success rate of over 75% (we define the defense success rate simply as the ratio of post-defense recognition accuracy to the original recognition accuracy). Furthermore, we also consider whether embedding an additional denoising module into the model would affect the detection of clean samples. From the table, it is evident that the worst-case scenario shows only a 7% decrease in recognition accuracy, while the rest are controlled within a 5% decrease. Therefore, it can be concluded that our defense strategy exhibits excellent generalizability. MIOU will be used as a measure for the experiment.

Table 4: Results of generalization validation

Attack	Whether to defend	ALC-Net	UIU-Net	DNA-Net	MDvsFA	CNU-Net
No Attack	No defense	74.69	51.09	64.50	63.83	78.61
	Defense	**73.20**	**48.87**	**57.39**	**61.29**	**73.94**
MFDP	No defense	48.81	8.470	20.50	16.38	24.47
	Defense	**66.97**	**43.39**	**49.04**	**56.14**	**65.35**
PGD	No defense	13.37	22.51	14.79	9.09	9.350
	Defense	**69.70**	**43.01**	**51.51**	**49.66**	**68.78**
FGSM	No defense	28.47	23.12	21.74	19.23	25.28
	Defense	**63.60**	**40.96**	**59.04**	**56.81**	**67.28**
BIM	No defense	26.68	9.360	21.65	13.29	28.08
	Defense	**69.61**	**43.97**	**49.10**	**55.37**	**66.12**
CW	No defense	21.08	6.510	15.04	7.53	1.980
	Defense	**51.05**	**38.82**	**54.68**	**46.87**	**52.03**
ZOO	No defense	30.71	25.01	25.67	20.79	26.68
	Defense	**68.25**	**47.36**	**52.99**	**59.46**	**63.07**
Boundary	No defense	21.09	35.14	25.06	24.72	21.04
	Defense	**62.17**	**45.91**	**53.46**	**53.24**	**61.38**

In Fig. 7, we have selected a set of examples showing a clean sample, an adversarial sample generated after MFDP, and a sample after denoising with a trained generator. Not only that, the figure also shows the feature images of each of the three in the recognition process and the corresponding recognition results. From the comparison of the feature images, it can be seen that the target features of the adversarial sample have been almost obscured and new approximate target features have been generated; while the denoised sample and the clean sample have similar feature images, and the target features can be distinguished more clearly. The recognition results are also similarly consistent with the judgment of the feature image.

To further visualize the results, we calculate the difference between adversarial samples, denoised samples and clean samples, using MSE as the metric; The IOU values between the recognition results of the three samples and the label images are also calculated. In Table 5, it can be found that there are some differences between adversarial samples and clean samples. However, the denoised sample has little difference with the clean sample, and it may also be obvious from the recognition results that the noisy sample has been able to be recognized more accurately.

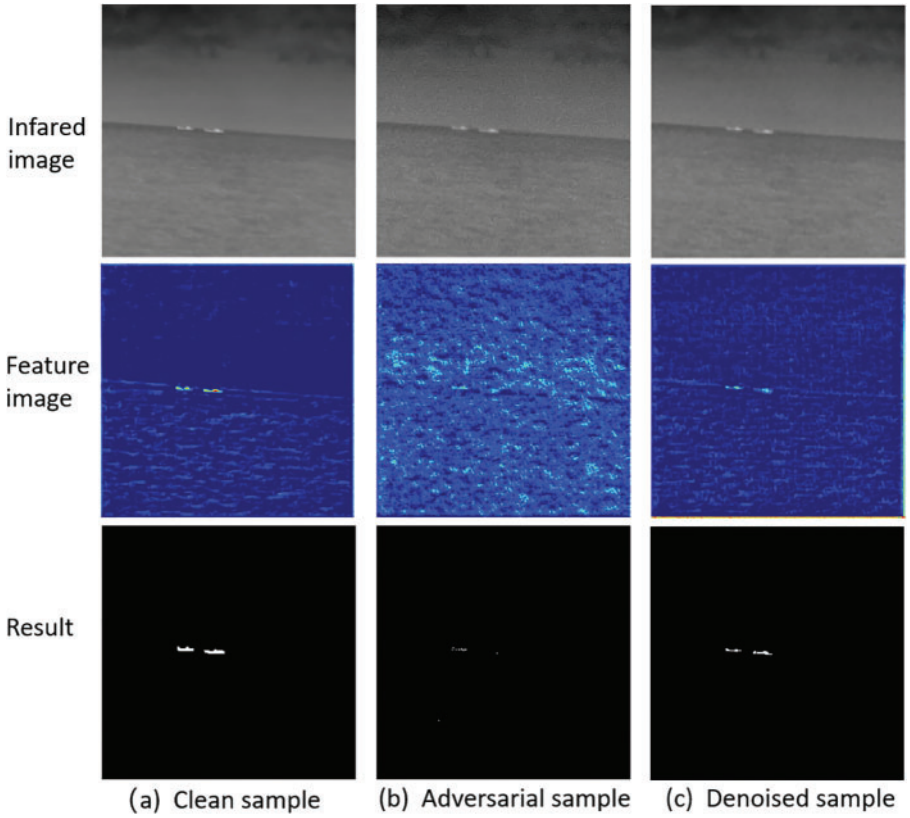

Figure 7: Schematic diagram of the defense effect

Table 5: Example sample comparison results

Sample	MSE (other sample, clean sample)	IOU (result, mask)
Clean sample	0	69.56
Adversarial sample	19.82	9.48
Denoised sample	3.71	58.43

5 Conclusion

In this study, we propose an adversarial defense strategy for small infrared targets. It generates migration perturbations using common features of multiple attacks, so our training process does not require a large number of adversarial samples; we also use GAN to train generator structures with excellent performance, which can be embedded into different target detection models, and ultimately, we can achieve good defense against multiple attacks with excellent generality.

Acknowledgement: None.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 62073164; and in part by the Shanghai Aerospace Science and Technology Innovation Foundation under Grant SAST2022-013.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Tongan Yu, Yali Xue; data collection: Shan Cui; analysis and interpretation of results: Shan Cui, Jun Hong; draft manuscript preparation: Tongan Yu, Yiming He. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *2014 Int. Conf. Learn. Rep. (ICLR)*, Banff, AB, Canada, Apr. 2014.

[2] C. K. Mummadi, T. Brox, and J. H. Metzen, "Defending against universal perturbations with shared adversarial training," in *2019 Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Republic of Korea, IEEE, Nov. 2019, pp. 4928–4937.

[3] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in *2014 Int. Conf. Learn. Rep. (ICLR)*, Banff, AB, Canada, Apr. 2014.

[4] R. Feinman, R. Ryan, S. S. Curtin, and A. B. Gardner, "Detecting adversarial samples from artifacts," 2017. doi: 10.48550/arXiv.1703.00410.

[5] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in *2018 Int. Conf. Learn. Rep. (ICLR)*, Vancouver, BC, Canada, Apr. 2018.

[6] P. Laykaviriyakul and E. Phaisangittisagul, "Collaborative Defense-GAN for protecting adversarial attacks on classification system," *Expert. Syst. Appl.*, vol. 214, 2023, Art. no. 118957. doi: 10.1016/j.eswa.2022.118957.

[7] Q. Liang, Q. Li, and W. Z. Nie, "Learning perturbations for adversarial defense based on GAN structure," *Signal Process. Image*, vol. 103, 2022, Art. no. 116659. doi: 10.1016/j.image.2022.116659.

[8] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *2020 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, IEEE, Jun. 2020, pp. 262–271.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *2014 Int. Conf. Learn. Rep. (ICLR)*, Santiago, Chile, Apr. 2015.

[10] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *2016 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, IEEE, Jun. 2016, pp. 2574–2582.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *2018 Int. Conf. Learn. Rep. (ICLR)*, Vancouver, BC, Canada, Apr. 2018.

[12] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*, Chapman and Hall/CRC, 2018, vol. 1, pp. 99–112.

[13] Y. P. Dong *et al.*, "Boosting adversarial attacks with momentum," in *2018 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, IEEE, Jun. 2018, pp. 9185–9193.

[14] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE Eur. Symp. Secur. Priv. (EuroS&P)*, Saarbrucken, Germany, IEEE, Mar. 2016, pp. 372–387.

[15] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symp. Secur. Priv. (SP)*, San Jose, CA, USA, IEEE, May 2016, pp. 582–597.
[16] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symp. Secur. Priv. (SP)*, Paris, France, IEEE, Apr. 2017, pp. 39–57.
[17] J. P. Zhang *et al.*, "Improving adversarial transferability via neuron attribution- based attacks," in *2022 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, IEEE, Jun. 2022, pp. 14993–15002.
[18] D. Wang, J. Lin, and Y. G. Wang, "Query-efficient adversarial attack based on Latin hypercube sampling," in *2022 Proc. 2022 IEEE Int. Conf. Image Process. (ICIP)*, Qingdao, China, IEEE, May 2022, pp. 546–550.
[19] A. Shafahi *et al.*, "Adversarial training for free!" in *2019 Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019, p. 32.
[20] H. Y. Zhang, Y. D. Yu, J. T. Jiao, E. P. Xing, L. E. Ghaoui and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *2019 Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, Jun. 2019, pp. 7472–7482.
[21] A. Liu, X. Liu, C. Zhang, H. Yu, Q. Liu and J. He, "Training robust deep neural networks via adversarial noise propagation," *IEEE Trans. Image Process.*, vol. 30, pp. 5769–5781, 2021. doi: 10.1109/TIP.2021.3082317.
[22] H. Y. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *2017 Int. Conf. Learn. Rep. (ICLR)*, Toulon, France, 2017.
[23] R. Muthukumar and J. Sulam, "Adversarial robustness of sparse local lipschitz predictors," *SIAM J. Math. Data Sci.*, vol. 5, no. 4, pp. 920–948, 2023. doi: 10.48550/arXiv.2202.13216.
[24] F. Z. Liao, M. Liang, Y. P. Dong, T. Y. Pang, X. L. Hu and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *2018 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, IEEE, Jun. 2018, pp. 1778–1787.
[25] A. Dubey, L. Maaten, Z. Yalniz, Y. X. Li, and D. Mahajan, "Defense against adversarial images using web-scale nearest-neighbor search," in *2019 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, Long Beach, CA, USA, IEEE, Jun. 2019, pp. 8767–8776.
[26] G. Q. Jin, S. W. Shen, D. M. Zhang, F. Dai, and Y. D. Zhang, "Ape-gan: Adversarial perturbation elimination with gan," in *2019 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, UK, IEEE, May 2019, pp. 3842–3846.
[27] M. W. Zhao, H. M. Qusay, and A. Sanaa, "Evaluation of GAN-based model for adversarial training," *Sensors*, vol. 23, no. 5, 2023, Art. no. 2697. doi: 10.3390/s23052697.
[28] W. Sanghyun, P. Jongchan, L. Joon-Young, and S. W. In, "CBAM: Convolutional block attention module," in *2018 Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, IEEE, Sep. 2018, pp. 3–19.
[29] Y. L. Xue, T. A. Yu, S. Cui, and Z. Zhou, "Infrared small target detection based on cascaded nested U-Net," (in Chinese), *J Jilin Univ. (Eng. Technol. Ed.)*, 2023. doi: 10.13229/j.cnki.jdxbgxb.20230785.
[30] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional local contrast networks for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9813–9824, 2021. doi: 10.1109/TGRS.2020.3044958.
[31] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, no. 3, pp. 364–376, 2023. doi: 10.1109/TIP.2022.3228497.
[32] B. Li *et al.*, "Dense nested attention network for infrared small target detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1745–1758, 2023. doi: 10.1109/TIP.2022.3199107.
[33] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Dallas, TX, USA, 2017, pp. 15–26.
[34] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *2018 Int. Conf. Learn. Rep. (ICLR)*, Vancouver, BC, Canada, Apr. 2018, pp. 1–12.

[35] J. Ahmadreza, J. S. Mohammad, K. Michelle, S. Christian, and W. Alexander, "Learn2Perturb: An end-to-end feature perturbation learning to improve adversarial robustness," in *2020 Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, IEEE, Jun. 2020, pp. 1241–1250.

[36] H. Wang, L. Zhou, and L. Wang, "Miss detection *vs.* false alarm: Adversarial learning for small object segmentation in infrared images," in *2019 Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Republic of Korea, IEEE, Nov. 2019, pp. 8509–8518.