

# Research on an Adversarial Attack Noise Enhancement Data Method for the SAR Ships Target Detection

Wei Gao<sup>1,2</sup>, Yunqing Liu<sup>1,\*</sup>, Yi Zeng<sup>1</sup>, Qi Li<sup>1</sup>

<sup>1</sup>Changchun University of Science and Technology, 7089 Weixing Road, Changchun City, Jilin Province, China;

<sup>2</sup>ChangChun University ChangChun, 6543 Weixing Road, Changchun City, Jilin Province, China

## Abstract

The target of maritime ships is the key goal of maritime monitoring and war blows. In recent years, research using SAR images for marine target detection and surveillance has attracted great attention in the field of marine remote sensing, becoming one of the most important marine applications for SAR data. Whether it can quickly and accurately identify the tactical intention of the marine battlefield ship's goals, and provide support for the decision-making of the commander, which is greatly related to the success or failure of the battle. If the automatic monitoring system is maliciously attacked by monitoring data, it will cause significant monitoring errors in the SAR image ship detection model. Reducing the impact of interference data on the model is high in generalization of models and can process unknown data. However, the existing offensive data does not take into account the improvement of the general performance of the model itself. Therefore, we introduced a method that combines adversarial attack methods and data enhancement, so that the model has the ability to improve generalization while retaining anti-offensive capabilities.

## Keywords

SAR image; ship target detection; noise; data enhancement; adversarial attack;

## 1. Introduction

Synthetic Pore Radar (SAR) is a two-dimensional imaging radar, which is very suitable for data sources for ship detection. The remote sensing monitoring system introduces deep learning models to improve the accuracy of detection, but data doping data seriously affects the performance of the model. The confrontation data that contains extra information with low disturbance rates may not affect humans. When deep learning network encounters malicious offensive data containing noise, the detection accuracy of the model will decrease, which will have a catastrophic impact on the detection system.

In order to make the model perform well when entering the examples and normal data, many researchers have proposed many training methods. Many researchers have studied how to design offensive data for offensive models. Szegedy et al. [1] found that "confrontation data" was formed by adding extra information in the image, resulting in a serious and wrong image classification result for interference's machine learning models. Such an example is generated by algorithms, the purpose is to deceive the machine learning model.

The working principle of FGSM [2] is to calculate the gradient of the loss function relative to the input, and to generate a sm-all disturbance by multiplying the selected small frequency  $\epsilon$  by the gradient symbol vector:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y)) \quad (1)$$

$\nabla_x \mathcal{L}(x, y)$  is the first guide of the loss function to the input z. In deep neural networks, this can be

---

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control  
EMAIL: e-mail: 2017200067@mails.cust.edu.cn (Wei Gao), \*e-mail: mzlyq@cust.edu.cn (Yunqing Liu), e-mail: 2019200080@mails.cust.edu.cn (Yi Zeng), e-mail: 2020200107@mails.cust.edu.c (Qi Li)



© 2022 Copyright for this paper by its authors.  
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).  
CEUR Workshop Proceedings (CEUR-WS.org)

calculated through back propagation algorithms [3]. Under the constraints of the model, the gradient vector symbol is to maximize the input disturbance amplitude, which also enlarges the confrontation changes in the model.

Other researchers put forward FGSM -based use iteration methods in subsequent studies: BIM[4][5], DDN[6], MI-FGSM[7]. The core of all these methods is to increase the "blunt feeling" of the model and reduce the sensitivity to disturbance. However, the calculation time and the number of iterations are linear. It takes a long time to produce a strong interference to accumulate to make it. And many researchers are for the purpose of attack, and they have not considered whether such adversarial attack can help the model improve the generalization of the model.

This study aims at the target of SAR images, and uses data to enhance means to inject data containing disturbing information into the model to improve the "blunt feeling" of the model. In addition, because of the basic and explanatory noise, the impact of exploration of noise on the model, in the research, finding a method of improving model stability and generalization.

## 2. Materials and Methods

For the target detection data of SAR images, you want to implement the noise enhanced data training model. Four important contents are: noise screening, noise attack method, Sensitive directional estimation and noise offensive enhancement. The overall generation process is as follows:

### 2.1. Noise screening

Artificial judgment SAR image ship's goal is more of the form and scattering. Some noise that cannot be detected by the naked eye cannot affect the effect of manual reading. However, this is not the case for deep learning models. In actual design, researchers hope that the accuracy of deep learning model recognition is high, but it also hopes that it can stabilize and generally not to be affected by noise. During the training, special enhanced data is needed, and the training data of the mixed standard is injected into the model to improve the blunt feeling of the model. Before that, choose a suitable noise [8]. The noise must meet several conditions: try to be closer to actual noise or simulate actual noise by simplicity, and the average noise is zero to ensure that the data after stacking noise is the same as the average value of the original data; after interference data, people cannot distinguish it.

SAR image noise uses its probability distribution function and probability density distribution function, which comes from the quality of the environmental conditions and sensing components itself from the image acquisition. The main factor of the image noise during transmission is that the transmission channel used is contaminated by noise. It mainly includes the following: salt and salt noise, random noise and Gaussian noise. Considering the complexity and controllable interpretation analysis of salt and pepper noise and random noise, choose Gaussian noise as the basic data to enhance noise.

### 2.2. Noise attack method

The variables used below are: input sample  $X$ , training model  $f$ , and the confrontation sample generated by the gradient rising:

$$\hat{X} = \sigma + X \quad (2)$$

$$\|\sigma\| < \epsilon \quad (3)$$

Where  $\sigma$  is disturbance. The confrontation data of the disturbance information is added to the  $f(\hat{X}) \neq Z$  in the model. Compared with the model without confrontation data  $f(X) = Z$ , the model performance is attacked. The performance of Model  $f$  depends on the target output of the opponent's target.

### 2.3. Sensitive directional estimation

The sensitive direction is estimated to be estimated to form the direction of strong offensive in the offensive sample, which is most likely that the direction of the final result determination of the target category changes. Model  $f$  evaluates the sensitivity of changes in the characteristics of each input sample. We need to explore the sensitivity of the model to noise, and we need to join the disturbances that the human eye cannot detect. In order to achieve this effect experiment, the disturbance added to the original sample should be as small as possible. Assuming the number of use norm  $\|\cdot\|$  Describe the differences between the points in the input domain, the confrontation sample in the model  $f$  can be formally turned into the following optimization problem:

$$X' = X + \operatorname{argmin}\{\|\varphi\|: f(\varphi + X) \neq f(X)\} \quad (4)$$

For the research goals of this article, when the formula (4)  $f(\sigma + X) \neq f(X)$ , two cases are charged. One is the goal classification error (omissions or error). The other is that although the detection box detects the target, the detection box deviations are huge. The input component value of the data  $X$  is added with the sensitivity value of these changes after the evaluation model  $f$ , which is a common method for changing sensitivity sensitivity.

### 2.4. Noise offensive enhancement (NOE)

The designer not only has a high accuracy of deep learning model recognition, but also hopes that it can be stable and generalized without being affected by noise. Then in the training, special enhanced data is needed, and the training data of the mixed standard is injected into the model to reduce the sensitivity of the model to noise. The enhancement method is mainly through reinforcement of noise. The noise in the actual environment of SAR images mainly includes salt and pepper noise, random noise and Gaussian noise. Considering the complexity and controllable interpretation analysis of salt and pepper noise and random noise, choose Gaussian noise as the basic data to enhance noise. The design process is as follows:

$$\eta \sim N(\mu, \delta) \quad (5)$$

$$\|x\|_2 = \sqrt{(x - \bar{x})^2} = \sqrt{\delta j} \quad (6)$$

$$E[\|\hat{X} - x\|] = 0 \quad (7)$$

$$\sum \eta_i = \sum \delta j \quad (8)$$

Formula (5) is noise expression. During the experiment, multiple noise was sampled from the Gaussian distribution, and the condition was  $\mu = 0$ . As shown in (6), the reason for the sampled noise is to ensure that the noise sample after the noise is added is the same as the average of the original sample. Then adjust the variance performance of the control sample by controlling the  $\delta$  in the noise distribution parameter. At the same time as the offensive, the upper formula (3) constraints in the definition of the problem must be met, so as the formula (7), the noise of the design must meet the constraints. The  $\delta$  value in the formula (6) will change with the number of iterations. Each noise meets the formulas (6) and (8). After ensuring optimization, multiple noise superposition still meets the constraints.

$$\|x + \sum_{j=1}^J \eta - x\|_p < \epsilon \quad (9)$$

$$\|\sum_{j=1}^J \eta\|_p < \epsilon \quad (10)$$

Follow the design to satisfy the formula (7) selection multiple groups to meet the noise of the upper formula (9) (10) distribution. When  $p = 2$ , it is equivalent to  $\|\sum_{j=1}^J \eta\|_2 = \sqrt{\sum \delta j} < \epsilon$  which is  $\delta j < \epsilon^2$ . According to the adjustment of the adjustment square  $\delta$  value according to the distracting  $\epsilon$  disturbance limit.

$$\delta j = \frac{\epsilon^2}{2^j} \quad (11)$$

$$\sum_{j=1}^J \delta j = \frac{\epsilon^2(1-\frac{1}{2^J})}{1-\frac{1}{2}} < \epsilon^2 \quad (12)$$

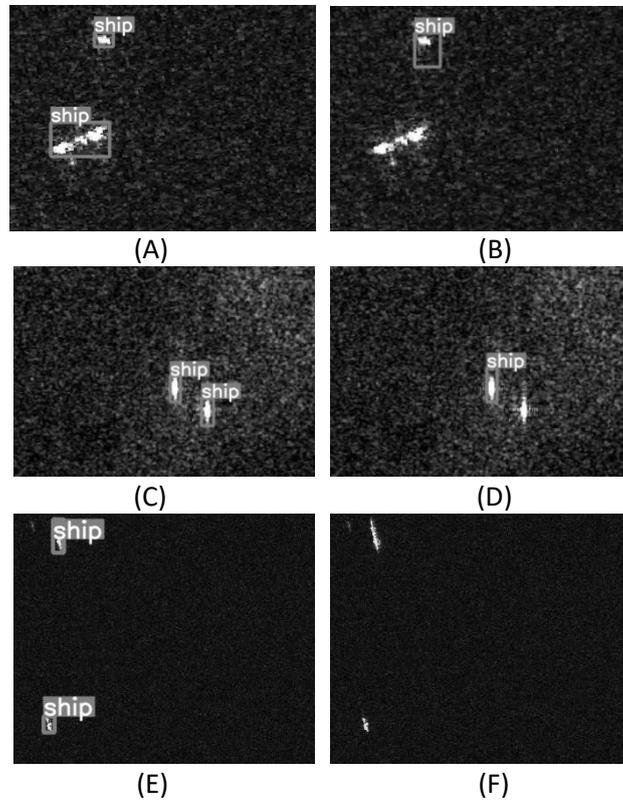
Let's set the value of  $\delta j$  as shown in the formula (11). It can be seen from the formula (12) that after any number of iterations, the constraints are still met. At the beginning, the weight will be higher, and it will gradually decrease as iterations. Select the highest noise  $\eta_t$  that allows the loss function  $L(f(x + \Delta x), y)$ . Its mathematical expression is:

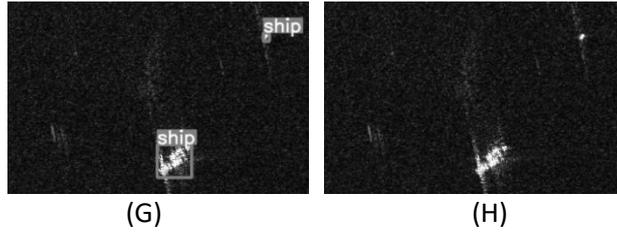
$$\eta_t = \arg \max_{\eta \in H} L(f(I_{t-1} + \eta), y^t) \quad (13)$$

In the formula (13), the maximum disturbance noise selection of the value of L. Through the above steps, the noise model of the effective offensive SAR ship detection model will be obtained.

### 3. Results & Discussion

Data for the SAR ship dataset enhances the use of widely used Yolov3 as the detection model in the offensive experiment. The abundant data of offensive scenarios SSDD [9] data set is used as an experimental set. Although multiple data sets are disclosed for the target of SAR image ships, the application of the data set is very different. It is difficult to have similar third-party datasets that can be used as verification sets of this experiment. In addition to the conventional detection model accuracy testing experiment, in order to fully detect the enhancement of the attack effect on the model, generalization experiments on the verification set are essential. In order to solve the above problems, the data is divided into 3 parts in the experiment: the first part of the  $\alpha\%$  data set is used as a training set, and 20% of the second part of the data is used as a test set, and the third part  $(100 - 20 - \alpha)\%$  data instead of third-party data Use  $\alpha \in (20,60)$  as a verification set.





**Figure 1** The comparison of the detection effect before and after the offensive (A, C, E, G is the detection effect of the previous model, B, D, F, H is the model detection effect after the offense)

Through cross -verification, multiple experiments are used to obtain persuasive data. It is known that the reduction of training data will have the accuracy of the model, but the data change trend of data on the test set and verification set is a more interested information.

As shown in Figure 1, the experimental data shows the effect of increasing the effect of the model detection before and after the noise increases. Compared with the first two groups (A) (B) and (C) (D) data in Figure 1, the target offset and missed inspection occurred after the attack. The latter two groups (E) (F) and (G) (H) data clearly show that the target of the ship that can be detected can be detected, and the loss target and the bounding box change after increasing the noise can be detected.

Test experiment data is divided into two parts: one is the comparison test of the noise enhancement of the offensive effect, and the other is the generalization effect comparison experiment. The former uses cross-validation training methods, 80% data is used as a training set, and 20% is used as a test set. The generalized experiment is the data allocation method of  $\alpha\%$ , 20% and  $(100-20-\alpha)\%$  of the validation set mentioned earlier. In the experiment, in order to ensure the balanced training data and verification data,  $\alpha = 40$ . Also obtain average data through multiple measurement.

### 3.1. Noise enhanced offensive effect comparison test:

**Table 1** Different offensive methods comparison

Attack	Precision	Recall	SuccessRate	mAP
Original	97.02	97.62	0	97.88
Random Noise	93.12	93.36	6.2	93.58
FGSM	18.65	19.12	79.23	19.21
NOE	<b>19.04</b>	<b>19.57</b>	<b>79.46</b>	<b>19.32</b>

As shown in Table 1, we can see the impact effect of enhanced data NOE containing noise on the model. Our NOE offensive success is much higher than Random Noise, but it is lower than FGSM. FGSM is infinitely iterated, the offensive effect is better than our NOE method.

**Table 2** Different offensive methods Defense effect comparison

Attack	Precision	Recall	mAP
FGSM	95.28	95.72	95.94
NOE	<b>96.01</b>	<b>96.57</b>	<b>96.68</b>

As shown in Table 2, the experimental process is to mix the noise enhancement data and the pure training data of the original data set as a training input into the model for defense training, and then the test is then for testing. It can be seen that the defense performance of the model is improved.

### 3.2. Generalization comparison test

In order to prove the generalization of the model on third -party data, the relevant generalized experiments are performed in accordance with the previous design. Related data is as follows:

**Table 3** Test the detection effect of test sets during no defense training

Attack	Precision	Recall	mAP
FGSM	12.12	12.23	12.36
NOE	<b>13.04</b>	<b>13.09</b>	<b>12.58</b>

As shown in Table 3, the model tests the test accuracy indicator obtained on the 20% test set in the  $\alpha\%$  dataset as a training set. Similar to the effect presented by Table 1, the accuracy and recall rate of NOE are slightly lower than FGSM. Because the training data becomes less, the various data indicators are low, but it does not affect the confrontation effect and the generalization of the model of our observation model.

**Table 4** Verification data test results during no defense training

Attack	Precision	Recall	mAP
Original	78.62	78.82	79.02

As shown in Table 4, the detection effect of the test model on the  $(100-20-\alpha)\%$  verification set. There is no verification set data for offensive training in the experiment. This set of data is used for comparison with the subsequent defense detection data.

**Table 5** Defense training verification data test results

Attack	Precision	Recall	mAP
FGSM	78.63	78.83	79.03
NOE	<b>82.68</b>	<b>82.76</b>	<b>82.74</b>

As shown in Table 5, after defense training, the data performance effect on the verification set is verified. Compared with the data of Tables 4 and Table 5, we can clearly see that our NOE method has offensive effects, which is better than random noise, but worse than FGSM. However, the indicators in Table 5 are far more than FGSM, which proves that NOE's defense ability and generalization are better than FGSM, and have certain anti-interference capabilities.

## 4. Conclusions

For the characteristics of the target data of the ship in SAR images, the design can enhance the noise enhanced data method of the model "blunt feeling": the data that superimposes the noise. Compared to superimposed random noise, our data is more obvious on the training effect of the model. Compared with strong fighting offensive form FGSM, our offensive effect is slightly inferior and does not reach a particularly high offensive success rate. However, after we were screened to strengthen the data of the data mixed with the original training and training, the defense capacity of the model has been greatly improved, and the generalization has become stronger. Such a model can be closer to the actual detection environment and can give the ship target test results more stable.

## 5. Acknowledgments

This work was supported by the Science and Technology Department Project of Jilin Province (under grant no. 20200404210YY) and the Science and Technology Department Project of Jilin Province (under grant no. 20210203039SF)

## 6. References

- [1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
- [2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial

- examples. arXiv preprint arXiv:1412.6572.
- [3] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533-536.
  - [4] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
  - [5] Kurakin, A., Goodfellow, I. J., & Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security* (pp. 99-112). Chapman and Hall/CRC.
  - [6] Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., & Granger, E. (2019). Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4322-4330).
  - [7] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185-9193).
  - [8] Ulaby, F. T., Moore, R. K., & Fung, A. K. (1982). *Microwave remote sensing: Active and passive. Volume 2-Radar remote sensing and surface scattering and emission theory*.
  - [9] Zhang, T., Zhang, X., Li, J., Xu, X., Wang, B., Zhan, X., ... & Wei, S. (2021). Sar ship detection dataset (ssdd): Official release and comprehensive data analysis. *Remote Sensing*, 13(18), 3690.