Training with Noise Adversarial Network: A Generalization Method for Object Detection on Sonar Image

Qixiang Ma[1] Longyu Jiang[1,2,*] Wenxue Yu[1] Rui Jin[1] Zhixiang Wu[1] Fangjin Xu[1]
[1]Laboratory of Image Science and Technology, Southeast University, Nanjing 210096, China
[2]Acoustic Science and Technology Laboratory, Harbin Engineering University, Harbin 150001, China
{qixiangma, JLY, ywx.list, ruijin, zhixiangwu, xfj_1109}@seu.edu.cn

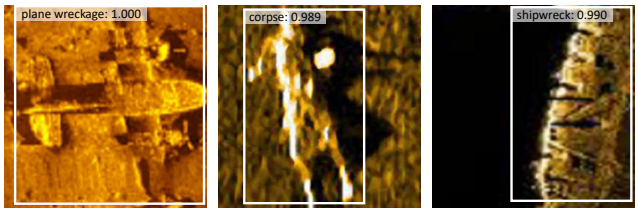
Abstract

Object detection tasks for sonar image confront two major challenges, scarcity of dataset and perturbation of noise, which cause overfitting to models. The state-of-the-art object detection designed for optical images cannot address the issues because of the inherent differentiation between the optical image and sonar image. To tackle this problem, in this paper, we propose an adversarial training method to generalize the detector by introducing perturbation with specific noise property of sonar images during training stage. We design a sideway network which we name Noise Adversarial Network (NAN). The NAN is embedded into the state-of-the-art detector to generate adversarial examples which serve as assistant decision-making items to predict both class and bounding box, aiming to improve the generalization and noise robustness of the detector. To provide prior knowledge of noise perturbation to NAN, we also design a Noise Block (NB) for introducing noise in the upstream layers, which further improves noise robustness. Following the Faster R-CNN framework, the results of our experiments indicate a 8.9% mAP boost on our sonar image dataset. The detector equipped with NAN and NB also outperforms the baseline on noised test sets. Furthermore, it gains a 2.4% mAP boost on the optical image dataset PASCAL VOC 2007.

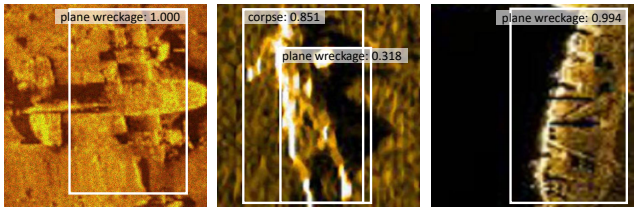
1. Introduction

Sonar (sound navigation and ranging) image, generated by imaging sonar system leveraging acoustic echo, depicts the underwater environment [2]. Object detection tasks for sonar images aim at locating and recognizing the object instances in sonar images. It serves as a prerequisite for an extensive set of downstream underwater vision tasks, such as sea mines detection [9] and fish tracking [6]. Classical approaches for object detection in sonar images contain various representations of template matching (TM) [29, 21], utilizing engineered features to implement both classification and localization. Other detection tasks in sonar images are based on paired highlight-shadow regions [10] and step-by-step feature extraction [42]. However, the above-mentioned methods have exceeding dependency of hand-engineered features which cannot interpret inherent representation of sonar images, inevitably inducing ineffective learning of sonar image data.

Thanks to Convolutional Neural Networks (CNNs), the Deep Learning techniques achieve quality results in optical image datasets such as ImageNet [8] and COCO [19]. The CNN-based object detectors designed for optical images can be directly used on sonar image detection tasks



(a) Original sonar images



(b) Sonar images with speckle noise (μ=0, σ^2=0.1)

Figure 1. Deterioration of sonar images with speckle noise. We use Faster R-CNN [32] with ResNet-101 [19] to train and test on our sonar image dataset. (a) The detector performs well when test on original sonar images. However, (b) the accuracy severely degrades when the sonar images are attacked by speckle noise with zero mean and variance $\sigma^2 = 0.1$. The three images in (b) illustrate wrong bounding box regression, false alarm and misclassification, respectively.

*Corresponding author.

because of the similarity between the features of object instances in optical images and sonar images [24, 38, 23]. For example, both a plane wreckage in a sonar image and an airplane in an optical image show the analogous inswept fuselage.

However, directly applying the state-of-the-art architectures such as Faster R-CNN [32] or SSD [26] to sonar image induces two major deficiencies. One is that the quantity of available sonar images is scarce. Different from the abundant optical image datasets [8, 19, 11], high-quality sonar images are hard to be gathered due to the high cost of exploration, let along an integrated public sonar image dataset with diversity. The other is that the perturbation of noise is diversified. The sonar imaging is usually accompanied by various noises of different types and magnitudes, which is caused by intricate underwater environment and imperfectly compensated vehicle motion [43]. As Figures 1 shows, these noises make sonar images degraded and plagues existing detection algorithms.

These two deficiencies inevitably lead to overfitting. Specifically, the finite samples which are used to train detectors make the parameters only fit the target values over the sample space. Meanwhile, training with existing noised sonar images cannot definitely match the test data which are subsequently gathered, because the test data may be generated from different sea areas with different underwater vehicles, inducing variety of noises. Furthermore, with the perturbation of noise in sonar images, the CNN-based detectors cannot reflect robustness in such detection tasks.

Inspired by adversarial learning [15, 28], a regularization method preventing model from overfitting, we propose a generalization method for detectors to tackle the problems above: training sonar images with adversarial perturbation which is generated from its own feature space. This perturbation has the property of noise which frequently arises in sonar images. The architecture that generates adversarial perturbation is a sideway network named Noise Adversarial Network (NAN). Instead of adding noise by yielding pixels directly, the NAN creates adversarial perturbation in convolutional feature space, which reduces computational cost and guarantees the end-to-end training. We also design a auxiliary component Noise Block (NB) to further improve the noise robustness. In our experiment, we demonstrate that the detector equipped with NAN and NB not only gains substantial improvement in detection performance, but obtains noise robustness compared with its baseline Faster R-CNN[32] on our own sonar image dataset.

The main contributions of our work are (1) proposing a Noise Adversarial Network (NAN) which enhances the ability of sonar image detection tasks; (2) yielding a sonar image dataset with remarkable variety, which will be shared in the near future to facilitate the underwater researches; (3) proving that this method not only improves performance on sonar image dataset, but is applicative on optical image dataset such as PASCAL VOC 2007.

2. Related Work

2.1. CNN-based Object Detector

Recently, significant progress has been made in the research field of object detection via convolutional neural networks. With regard to two-stage detectors, R-CNN [13] is one of the first architectures which brings CNN to object detection takes. The following SPP-Net [18] and Fast R-CNN [12] also achieves prominent performance on several benchmarks (*e.g.*, PASCAL VOC [11] and MS COCO [19]). The dominant pipeline Faster R-CNN [32] revolutionizes the CNN-based object detector by combining the proposal generator and detector into a unified end-to-end version. From then on, numerous intriguing two-stage detectors has emerged [7, 25, 17], boosting the representational power and progressively improved performance.

2.2. Object Detection for Sonar Image

The detection for underwater objects in sonar images is a universal task across various research domains. It facilitates archaeology [35], guarantees the pipeline monitoring [31] and maps habitat [45]. Classical detection methods yield unreliable performance because of the overdependence on environmental conditions such as seafloor elevation [44] and illumination intensity [16]. Meanwhile, the methods based on man-made features such as highlight-shadow patterns [9] and template matching [29, 21] are also considered as brittle algorithms with low generalization.

Recently, CNN-based object detectors yield an improvement of accuracy on sonar images [39, 38, 23, 40, 24]. However, with the finite training data, neither strong generalization nor robustness to noise is proven. The direct usage of state-of-the-art object detectors on sonar images is prone to overfitting.

Meanwhile, sonar images are notorious for speckle noise perturbation [46]. To mitigate the effect of speckle noise on detectors, a common wisdom is to introduce additional noise besides the noise of the sonar image itself [49], which is helpful to enhance the image and improve noise robustness of detectors. Nevertheless, directly inducing speckle noise to sonar image with fixed parameters is restricted to a limited sample space. We conjecture that the parameters of noise model are learnable and can be self-adaptively estimated during the training stage, which is also applied in the noise model estimation of sonar imagery [48].

2.3. Adversarial Learning with Random Perturbation

Random perturbation is a current regularization method for generalization. It was proved in [1] that adding Gaus-

sian perturbation to inputs illustrates an analogous effect to regularization term in the objective function. The original adversarial learning [37] induces the similar idea, showing that training with generated adversarial examples can generalize models, which yields robustness against adversarial examples while reducing the test error.

In object detection tasks, OHEM [33] extracts realistic features from real images to make hard examples, strengthening robustness of models. A-Fast-RCNN [41] generates the hard examples of occlusion and deformation in convolutional feature space to learn invariance. Our work shows analogous idea to A-Fast-RCNN [41]. However, instead of using dropout or spatial transformation to block or deform the convolutional features, we introduce adversarial noise perturbation related to sonar image to generalize model as well as improve robustness of noise attack.

3. Training with Noise Adversarial Network

We start this section with a definition of notations. Let us assume that $x \in S$ and y $\in L$ denote an input image and its label respectively, where S=$\{x_1, x_2, ..., x_n\}$ is the set of all samples and L=$\{y_1, y_2, ..., y_n\}$ represents the set of labels, each y includes the ground-truth class y_c and bounding box location y_l. We also define that $D_c(x)$ and $D_l(x)$ respectively represent the predicted class and bounding box location, output values of an object detector.

3.1. Adversarial Learning for Object Detection

The current criterion in object detector can be written as,

$$L_D = L_{soft\max}(D_c(\chi), y_c) + [y_c \notin bg] L_{bbox}(D_l(\chi), y_l), \quad (1)$$

where χ is the representative feature extracted from its corresponding input x. The χ can also be regarded as the original example without perturbation. The first term is SoftMax loss while the second one is bounding box loss which only includes the foreground classes.

The adversarial learning based approach aims to generate adversarial example to fool the detector such as A-Fast-RCNN [41]. In our approach, we directly use the adversarial examples as a kind of extra representation derived from the original examples. These adversarial examples serve as assistant decision-making items to predict both class and bounding box, which is same mechanism as original ones,

$$L_P = L_{soft\max}(D_c(\chi + r_{adv}), y_c) + [y_c \notin bg] L_{bbox}(D_l(\chi + r_{adv}), y_l), \quad (2)$$

where r_{adv} denotes the perturbation noise that is generated by our Noise Adversarial Network (NAN). The item $\chi + r_{adv}$ is the adversarial example corresponding to χ. We also measure the divergence between the distributions of original and adversarial examples to approximate a true distribution which is robust to the noise attack r_{adv}. The loss function can be written as,

$$L_{adv}(\chi, \theta) := D_{KL}[q(y|\chi), p(y|\chi + r_{adv}, \theta)], \quad (3)$$

where $q(y|\chi)$ represents the distribution of original examples which is supposed to be approximated to a true distribution by $p(y|\chi + r_{adv}, \theta)$, the distribution of adversarial examples. θ represents the vectors of the parameters of NAN. $D_{\mathrm{KL}}[q, p]$ measures the KL divergence between two distributions q and p.

In our training approach, we use the abovementioned three loss functions assembly to fine-tune the model to be robust to various noise attack as well as gain generalization. The loss of the model is the sum of the three loss.

$$L_{total} = L_D + L_P + L_{adv} \quad (4)$$

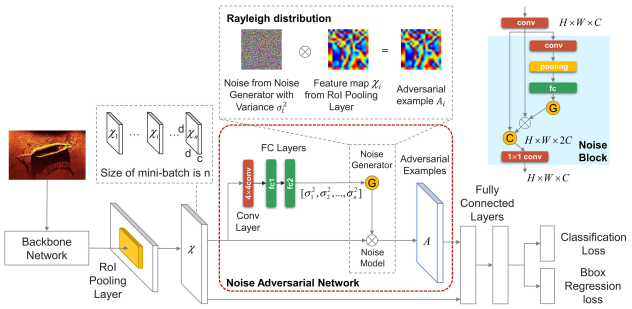


Figure 2. The Faster R-CNN equipped with NAN (red box) and NB (upper right). The NAN uses a convolutional layer with kernel size $k = 4$ following two fully connected layers to predict variance σ_i^2 for each feature map χ_i from 1 to n in the mini-batch. The variances are used to generate noises whose amplitudes follow Rayleigh distribution. Then these noises are introduced into the original feature maps to generate adversarial examples. NB acts as a auxiliary component, only inducing speckle noise in the upstream to provide prior knowledge for NAN without changing the shape of features.

3.2. Noise Adversarial Networks Design

Our Noise Adversarial Network (NAN) is prone to be integrated in various state-of-the-art object detection frameworks. For the purpose of easy comprehension and implementation, we choose Faster R-CNN [32] as our baseline model. It also serves as a prerequisite detection framework referenced by several subsequent detection tasks [22, 36, 4].

We embed NAN to the object detector to gain generalization as well as achieve noise robustness without computationally expensive cost. Therefore, the design of NAN is supposed to follow the two principles. One is that the NAN should generate noise perturbation of specific distribution from convolutional feature space instead of adding random pixel values on the input image. The other is that NAN

should be easily embedded into a detector to guarantee end-to-end training.

Properties of Noise in Sonar Image. The echo signals received by sonar are mainly affected by three kinds of noises, namely environmental noise, reverberation and self-noise while reverberation is the dominating one, especially in shallow water areas [30]. It is defined as the scattered sound energy of water or heterogeneous bodies in water boundary received by hydrophone [5]. The intensity of reverberation varies with the distance of the scatterer and the intensity of the transmitted signal. We only consider reverberation as the source of perturbation noise in our model. According to Middleton's seabed reverberation model [27], the reverberation at moment t can be defined as the sum of the real and imaginary parts,

$$X(t) = \mathrm{Re}(t) + j\,\mathrm{Im}(t). \tag{5}$$

We assume that $V_n(t)$ and $\varphi_n(t)$ respectively represents the instantaneous amplitude and the instantaneous phase of the n-th scatterer. The real and imaginary parts can be formulated as,

$$\mathrm{Re}(t) = \sum_{n=1}^{N} V_n(t)\cos\varphi_n(t), \tag{6}$$

$$\mathrm{Im}(t) = \sum_{n=1}^{N} V_n(t)\sin\varphi_n(t), \tag{7}$$

according to Central Limit Theorem(CLT) [27], $\mathrm{Re}(t)$ and $\mathrm{Im}(t)$ will converge in Gaussian distribution if N is large enough. Meanwhile, for the random scatters, $V_n(t)$ and $\varphi_n(t)$ are random variables independent of each other. $\varphi_n(t)$ can be regarded as Uniform distribution from 0 to 2π [47]. Therefore, the mean of $\mathrm{Re}(t)$ and $\mathrm{Im}(t)$ can be written down as,

$$\begin{aligned}\langle \mathrm{Re}(t)\rangle &= \sum_{n=1}^{N}\langle V_n(t)\cos\varphi_n(t)\rangle \\ &= \sum_{n=1}^{N}\left[\langle V_n(t)\rangle \cdot \tfrac{1}{2\pi}\int_0^{2\pi}\cos\varphi_n(t)d(\varphi_n(t))\right] = 0\end{aligned}, \tag{8}$$

$$\begin{aligned}\langle \mathrm{Im}(t)\rangle &= \sum_{n=1}^{N}\langle V_n(t)\sin\varphi_n(t)\rangle \\ &= \sum_{n=1}^{N}\left[\langle V_n(t)\rangle \cdot \tfrac{1}{2\pi}\int_0^{2\pi}\sin\varphi_n(t)d(\varphi_n(t))\right] = 0\end{aligned}, \tag{9}$$

where $\langle A\rangle$ is mean of A. Therefore, $\langle \mathrm{Re}(t)\,\mathrm{Im}(t)\rangle = 0$ and the two parts have equivalent variance σ^2. Since both $\mathrm{Re}(t)$ and $\mathrm{Im}(t)$ follow the Gaussian distribution and they are independent of each other, with the equivalent variance σ^2 and zero mean, the amplitude of reverberation $|X(t)| = \sqrt{\mathrm{Re}^2 + \mathrm{Im}^2}$ follows Rayleigh distribution.

Based on the above analysis, because of the widespread scatterers, a single measurement for each pixel is in random variability. The probability density function of its amplitude follows Rayleigh distribution with zero mean and variance σ^2. If the parameter σ^2 can be predicted, the noise with its amplitude following Rayleigh distribution can be generated.

Details in Noise Adversarial Network. To ensure that the noise can be generated in a simple but efficient way, we propose a lightweight sideway network which includes a convolutional layer followed by two cascaded fully connected layers. The convolutional layer has a large $k \times k$ kernel to aggregate representative features from RoI pooling layer. Then the down-sampled feature maps are fed into the following fully connected layers to predict the output. This process is unsupervised because no label of noise is used. The only purpose of NAN is to generate noise whose amplitude follows Rayleigh distribution. Specifically, for the two parameters of noise mean and variance, it only predict the variance σ^2 of $\mathrm{Re}(t)$ and $\mathrm{Im}(t)$ because we had proved that both real and imaginary parts had constant zero mean.

In Faster R-CNN, length-fixed convolutional features are generated from RoI pooling layer. These features represent the high dimensional semantic information of foreground object proposals. Motivated by A-Fast-RCNN [41], we utilize the region-based features from RoI pooling layer as the input of our Noise Adversarial Network to predict σ^2. One reason is that the structure of NAN is constant and length-fixed features are prone to be manipulated. The other is that embedding NAN following RoI pooling layer will not ruin the integrality of the upstream parts such as backbone or RPN networks.

Since the noise generation in NAN is unsupervised, intuitively, it is necessary for NAN to obtain the prior knowledge of speckle noise. We empirically embed a component named Noise Block (NB) in the upstream network (*e.g.* following the first convolutional layer in ResNet-101) to introduce noise in the features with low semantics but high resolution. The NB is to provide prior knowledge to automate the choice of noise magnitude of NAN in a reasonable range.

As Figure 2 shows, given a mini-batch of length-fixed features generated from RoI pooling layer, each feature χ_i in the mini-batch has a size of $c \times d \times d$, where c is the number of channels and d denotes the spatial length (*e.g.*, in ResNet-101 [19], c=1024, d=7 and the size of mini-batch n=128). The feature χ_i is fed into the NAN to predict the variance ${\sigma_i}^2$ of noise. The noise generated by NAN with its amplitude follows Rayleigh distribution can be formulated as,

$$noise_i = \sqrt{\left(noise_i^{\mathrm{Re}}\right)^2 + \left(noise_i^{\mathrm{Im}}\right)^2}, \tag{10}$$

where $noise_i^{\mathrm{Re}}$ and $noise_i^{\mathrm{Im}}$ denote the noise of real part and imaginary part respectively in Eq.(5). Both of them follow Gaussian distribution with zero mean. They have

equivalent value of variance ${\sigma_i}^2$. The distribution can be written down as,

$$noise_i^{\mathrm{Re}} \sim N(0, NAN(\chi_i)), \tag{11}$$

$$noise_i^{\mathrm{Im}} \sim N(0, NAN(\chi_i)), \tag{12}$$

where $NAN(\cdot)$ denotes the output of Noise Adversarial Network. $NAN(\chi_i)$ represents the variance ${\sigma_i}^2$ predicted from the $i\text{-}th$ feature χ_i of the mini-batch.

The upper right of Figure 2 showcases the NB, which follows the same principle of generation and addition of speckle noise as NAN. To guarantee the unified input of fc layer, we leverage RoI Align Pooling [17] to crop the feature map into a fixed size. Then the noised feature and the original one are concatenated on the channel dimension, fed into a 1×1 convolutional layer. It keeps the shape of feature map remaining the same.

Noise Model. Once the variance ${\sigma_i}^2$ is predicted, $noise_i$ which follows Rayleigh distribution can be generated by the noise generator in a random way. The $noise_i$ has the identical shape to the feature χ_i and is added to the original feature to generate the adversarial example by a specific noise model. We follow the idea in [3] that noise model is supposed to cope with mutable signal dependence caused by various imaging conditions,

$$A_i = \chi_i + (\chi_i)^{\gamma} noise_i \tag{13}$$

where A_i represents the $i\text{-}th$ generated adversarial example of the mini-batch, γ is a non-negative exponential parameter which is related to the dependence on the feature χ_i.

Training Approach. We design this adversarial learning method with a stage-wise training approach. It consists of two stages. In the first stage, we initialize the standard Faster R-CNN by pre-training from ImageNet [8] and train this model without NAN, aiming to make the model have an approximate distribution fitting of the object instances in dataset. Then in the second stage, we jointly train the pre-trained Faster R-CNN model and our NAN during each iteration. The NAN model is initialized with Xavier [14]. In a single forward propagation, the NAN generates noise with Rayleigh distribution and adds it to the original feature abiding by the formulation of Eq.(13), yielding the adversarial examples to introduce perturbation for the detector. The process of generation of adversarial examples only occurs in the training stage. We remove the NAN in the test stage, only leveraging the detector with trained weights, which ensures that no extra time cost will be induced. We treat NB in the same way with NAN, analyzing its auxiliary function in Section 4.4.

4. Experiments

We conduct our experiments on our own dataset and a benchmark. On our dataset, we first compare our method with the baseline and another two adversarial learning based object detectors. Then we analyze the noise robustness and generalization of NAN and NB. Furthermore, we apply this method on PASCAL VOC 2007 to verify its correctness.

4.1. Collected Dataset

The purpose of our adversarial training method is to generalize the standard object detectors to avoid overfitting while obtain robustness to noise attack in sonar images. To this end, we built up a sonar image dataset that satisfies the following three requirements. First, the dataset contains remarkable variety in terms of object size, illumination, position and noise distribution. Secondly, it is significant that the dataset do not illustrate systematic bias such as a preference of images that contains centered objects with ideal illumination and orientation. Thirdly, the annotations of each image need to be consistent, precise and exhaustive.

We gathered a set of 216 sonar images that were captured by Side Scan Sonar (SSS) and Synthetic Aperture Sonar (SAS) from public photo-sharing websites. These sonar images are taken without the purpose of computer vision tasks such as object detection, which guarantee that these images are not 'biased'. The set of the images contains an extensive range of observed conditions such as lighting as well as ambient noise and the images are not excessively partial to a particular object. These sonar images include three categories: corpse, shipwreck and plane wreckage. We split them into two subsets A (168 images) and B (48 images), then obtained 25 samples from each original image by random cropping for data augmentation. The cropping size was 80% of the original size on each side. We got augmented subsets A+ (4200 samples) and B+ (1200 samples), then resized the 5400 samples to 600×600 pixels and marked each object of the samples with an annotation following the guideline of PASCAL VOC [11]. We use subset A+ and B+ for training and testing, respectively. There are no intersections between the train and the test set.

4.2. Implementation details and optimization

We train the detector with NAN following the aforementioned two-stage approach in Section 3.2. We apply stochastic gradient descent (SGD) for the two-stage training with 10K and 80K iterations, respectively. Both stage starts a learning rate of 0.001, decreasing to 0.0001 after 0.8K and 60K, respectively. The entire model is trained with a momentum of 0.9 and a weight decay of 0.0001 on four NVIDIA GeForce GTX TITAN X GPUs with 12 GB memory. The detector leverages RoIAlign Pooling [17] to generate the length-fixed RoI features. No extra data augmentation was used for the generated samples in Section 4.1 except standard horizontal image flipping. Each mini-batch involves only one image per GPU with 128 (in ResNet-101) or 256 (in VGG16) RoIs per image.

Method	Arch	mAP	cp	shw	plw
FCNN [12]	VGG	65.6	39.5	86.1	71.1
FRCNN [32]	VGG	68.8	48.8	86.3	71.3
FRCNN+	VGG	72.8	63.4	**88.5**	66.6
Ours	VGG	**80.4**	**81.3**	87.7	**72.3**
OHEM [33]	VGG	63.0	33.3	85.4	70.4
A-FCNN [41]	VGG	61.1	19.1	87.3	76.9
FRCNN [32]	ResNet	79.8	72.8	90.6	76.0
FRCNN+	ResNet	81.7	84.1	90.5	70.5
Ours	ResNet	**88.3**	**96.7**	**90.8**	**77.4**

Table 1. Detection average precision (%). FRCNN+ refers to FRCNN [32] with our training schedule. Cp, shw and plw refer to the three categories, corpse, shipwreck and plane wreckage, respectively.

Method	mAP	cp	shw	plw
FRCNN+NA	83.5	81.1	90.4	79.1
Ours(FRCNN+NB)	85.8	92.5	90.6	74.4
Ours(FRCNN+NAN)	88.3	96.7	**90.8**	77.4
Ours(FRCNN+NAN+NB)	**90.6**	**97.7**	90.0	**84.0**
FRCNN+NAN+NB+NA	85.7	89.8	89.8	77.4

Table 2. Detection average precision (%) of ablation study with NB (with ResNet-101). NA represents Noise Augmentation, which means directly adding speckle noise (zero mean, variance varying from 0 to 1) to sonar images.

Additionally, for the purpose of aggregating representative features as well as generating the adversarial examples in a simple but efficient way, we set the kernel size $k = 4$ of convolutional layer in NAN with zero padding and one stride. (Figure 2). We set the exponential parameter γ=1 in Section 4.3.

4.3. Results on Our Dataset

We report the results on our own dataset for training Faster R-CNN with NAN. Table 1 showcases that the main results of NAN with the backbones VGG16 [34] and ResNet-101 [19] are 80.4% mAP and 88.3% mAP, respectively. They obtain competitively higher performance compared with the results of their corresponding standard pipelines 68.8% mAP and 79.8% mAP. In addition, they also outperform the baseline detector trained with our training schedule (the stage-wise training approach proposed in Section 3.2), which have a result of 72.8% mAP and 81.7% mAP, respectively. Furthermore, we also compare our method with another two widely used adversarial training strategies on our dataset. The OHEM [33] yields a result of 63.0% mAP while A-Fast-RCNN [41] gives 61.1% mAP, both of them are lower than our results.

4.4. Ablation Study with Noise Block

Since NAN generates the noise in an unsupervised way, we design a Noise Block (NB) to yield noise in the shallow layers (in this case it follows the first convolutional layer in ResNet-101 and VGG16). We hypothesis the NB can offer prior knowledge of noise generation to NAN because the features from shallow layers have low semantics but high resolution and geometric details, which is reasonable to estimate the speckle noise in the space domain.

Table 2 verify that the cooperation of NAN and NB improve the performance of detector. When the detector is equipped with both NAN and NB, the mAP is further improved to 90.6%. Additionally, the result of plane wreckage was largely boosted, which explains that the NB implicitly enhances the ability of NAN. To validate the effect of NB, we record the variances predicted by NAN during 50k training iterations with and without NB. Figure 3 (a) and (b) shows the number and distribution. It is obvious that in Resnet-101, the combination of NAN and NB (w/ NB) can generate the variances in a concentrated range. In VGG16 however, the variances generated in this way (w/ NB) follow a long-tail distribution. This indicates that with the effect of NB, NAN can predict the speckle noise perturbation in a different domain, which varies according to structure of backbones.

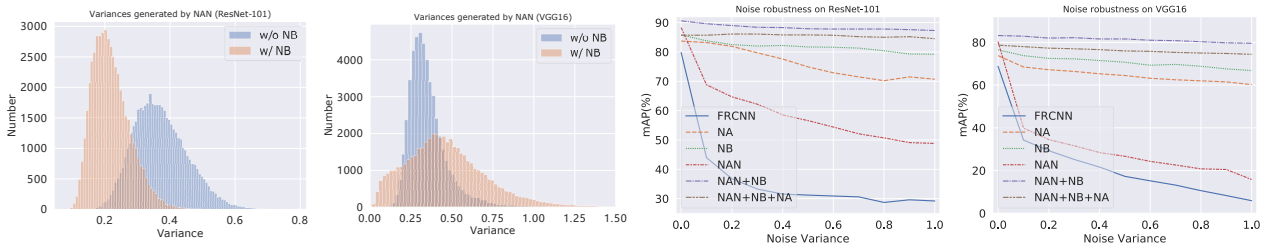
Utilizing NB alone still generates a competitive of 85.8% mAP. We also compare this mechanism with Noise Augmentation (NA), a common protocols to avoid overfitting, which introduces speckle noise manually to the sonar images before training. The noise we add to image has zero mean and the variance randomly varies from 0 to 1. This method yields a mAP of 83.5%, falling short of NAN and NB, which is mainly caused by the restricted noise model. An intriguing fact is that when we combine the three methods (NAN+NB+NA), the mAP is reduced (85.7% in Table 2), we conjecture it is the excessive addition of speckle noise that plagues the detector.

4.5. Analysis of Noise Robustness

The original dataset implicitly includes noise which is generated by the intricate underwater environment. However, the noise cannot be analyzed quantitatively. Therefore, we manually introduce noises to the test set to analyze the noise robustness of NAN and NB. We add speckle noise with zero mean and variance σ^2 varying from 0 to 1 to simulate the varying underwater reverberation, the speckle noise model can be expressed as,

$$I_s = I_o + n * I_o, \tag{14}$$

where I_s denotes the image with speckle noise, I_o represents the original image while n is speckle noise with mean 0 and variance σ^2. The model is equivalent to Eq.(13) in the case that γ=1.



(a) (b) (c) (d)

Figure 3. Result for further investigation. (a), (b) Histogram of variances generated by NAN with and without NB on ResNet-101 and VGG16, respectively. (c), (d) Noise Robustness with various intensities of noise attack (μ=0, σ^2 from 0 to 1). (c) is on ResNet-101. (d) is on VGG16

Figure 3 (c) and (d) illustrate the noise robustness of various mechanisms on both ResNet-101 and VGG16, respectively. The original Faster R-CNN shows highly vulnerable to speckle noise. With the variance σ^2 of noise attack raising to 1, the performance exponentially drops to 29.2% and 5.9% on the two backbones respectively (blue line). Equipped with NAN, the detector shows an improvement to 48.8% and 15.8% (red dotted line), but still falling short of a robust model. It is obvious that NAN substantially boosts the performance on original test set ($\sigma^2 = 0$) but falls to tackle the problem in noise cases. The coordination of NAN and NB overcomes the noise attack with all intensities (purple dotted line, specifically in the case of $\sigma^2 = 1$ yielding 87.3% and 79.6% on the two backbones, respectively), which demonstrates that with the assist of NB, the NAN achieves high noise robustness. It is also showed that to some extent noise attack can be mitigated with noise augmentation (yellow dotted line). However, when combining our NAN and NB with this mechanism, the mAP drops with a certain degree (brown dotted line), which can be explained as the reduction caused by excessive addition of noise.

Error analyses. We use the analysis method from [20] to understand type of false positive suppressed by NAN and NB. The detection results are classified into four groups: 1) Correct detection **(Cor)**: correct classification with IoU > 0.5. 2) Localization error **(Loc)**: correct classification with misaligned bounding box ($0.1 <$ IoU < 0.5). 3) Similar **(Sim)**: confusion with wrong class with IoU > 0.1. 4) Background **(BG)**: confusion with background (IoU < 0.1).

Figure 4 showcases the top-ranked detections with noise attack ($\sigma^2 = 0.5$). It is obvious that the coordination of NAN and NB extensively eliminates the false positive of Loc and Sim, reducing them from 44% (FRCNN) to 8% (NAN+NB). Equipped with NAN and NB, the detector shows stronger localization (Loc) and classification (Sim) ability. It is worth noting that there are no BG error in top-ranked detections, which means the detector has less confusion with background in the case of noise attack.

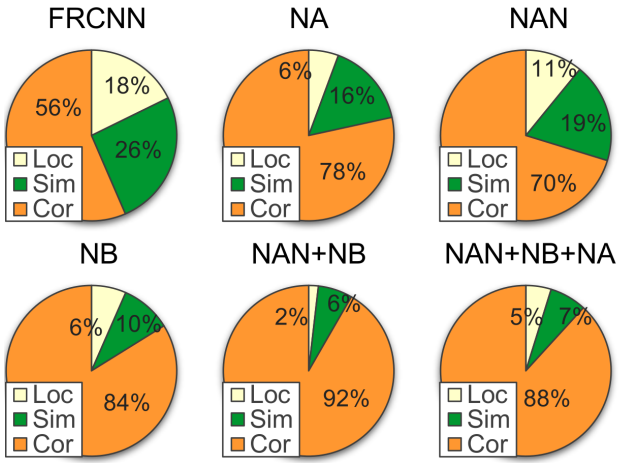


Figure 4. Error analyses on different models. Test set is attacked by speckle noise (μ=0, σ^2=0.5). Distribution of top-ranked detections include Cor (correct), Loc (misaligned localization) and Sim (confusion with a wrong class). similar category)

4.6. Analysis of Generalization with Various Noises

To investigate the generalization of NAN and NB, we manually enlarge variation between train and test sets. We construct four duplicate datasets from original sonar datasets, introducing each of them a specific kind of noise: speckle (**Spe**), gaussian (**Gau**), poisson (**Poi**) or salt-and-pepper (**S&P**). In datasets of Spe and Gau, the mean and variance of both speckle and gaussian noise are 0 and randomly varying from 0 to 1, respectively. In S&P, the proportion of salt-and-pepper noise replacing the image pixels is 0.1. The detectors are trained with each train sets and tested on all the test sets, respectively.

Figure 5 shows the heat map of results on both baseline (left) and detector equipped with NAN and NB (right), both of them are trained and tested on different kinds of noise. It is remarkable that (1) detector with NAN and NB outperforms its counterpart base network in almost all the cases, with only one exception that training with S&P and

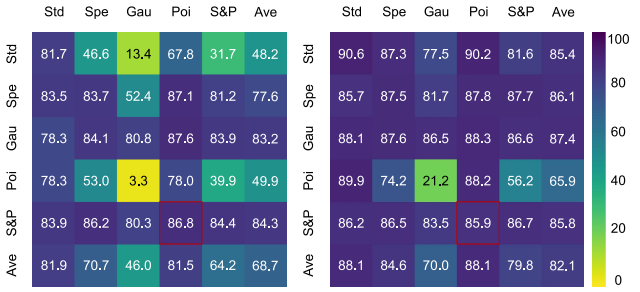


Figure 5. Heat map of detection results (%) on both baseline (left) and detector equipped with NAN and NB (right). The column labels represent the train sets with its related type of noise while the row labels mark the test sets. The value in each cell is the result generated by its corresponding train and test set, *e.g*. the result in red box is yielded by training on salt-and-pepper noised train set while testing on poisson noised test set. We also illustrate the case on original dataset (Std) and average value (Ave) of its related column or row.

test on Poi test set (red box); (2) 14 over 36 cases of baseline show mAP lower than 70%, which illustrates its poor generalization, since it over fits the finite training samples while yields low mAP on test sets with other kinds of noise; (3) comparing the first row with the rest of rows of baseline heat map, it is obvious that noise augmentation indeed improves the results of detection except for Poi; (4) training with Poi dataset dramatically degrades the performance of detectors on each test set. It is mainly because that as a signal-dependent noise, poisson noise is largely related to sonar image itself, inherently corrupting train sets.

In most cases, the train and test sets are attacked by different types of noise, the results prove that the NAN and NB allow for greater variation between train and test set, the ability of generalization of NAN and NB is explicitly measured.

4.7. Results on PASCAL VOC 2007

To verify the universality of NAN and NB, we apply them on the optical image benchmark PASCAL VOC 2007. We train the models on VOC 2007 and VOC 2012 `trainval` sets ('07+12') and test on VOC 2007 `test` sets. In addition, we introduce two types of noise to the test set. Besides the speckle noise we leveraging in Section 4.5, we utilize Gaussian noise which is ubiquitous in optical images to simulate the degraded images that are captured in the abnormal illumination or transmission conditions. To cope with the Gaussian noise, we also set a compensatory experiment in which NAN generates perturbation noise following Gaussian distribution, using additive noise model.

As Table 3 shows, in both ResNet-101 and VGG16 architectures, NAN and NB with Gaussian perturbation noise outperforms another two methods in original and gaussian

Method	ResNet			VGG16		
test set	O	Spe	Gau	O	Spe	Gau
FRCNN+	76.6	58.7	56.8	73.5	47.3	47.2
Rayleigh	77.8	67.3	66.7	74.5	**65.2**	64.1
Gaussian	**79.0**	**72.2**	**74.2**	**75.7**	64.4	**68.5**

Table 3. PASCAL VOC 2007 test mAP(%). FRCNN+ refers to FRCNN [32] with our two-stage training schedule. Rayleigh denotes training model with NAN and NB which generates noise following Rayleigh distribution ($\gamma = 1$, Eq.(13)). Gaussian refers to training NAN and NB which generates noise following Gaussian distribution with an additive noise model. 'O', 'Spe' and 'Gau' represent original test set, speckle and Gaussian noised (μ=0, σ^2=0.1) test set, respectively.

noised test set. In ResNet-101, the detector equipped with NAN achieves 79.0% mAP, 2.4% higher than baseline. The model equipped with NAN and NB reflects its robustness on both speckle and Gaussian noise attack. NAN and NB with perturbation noise following Rayleigh distribution does not achieve high performance on original test set as Gaussian model but still competitive. A speculative explanation is that without a coherent imaging system under normal conditions, the imaging mechanism of optical images does not involve speckle noise, which is different from sonar images.

5. Conclusion

We present an adversarial strategy to generate adversarial examples with peculiar noise property of sonar image by a sideway Noise Adversarial Network with its auxiliary part Noise Block. Extensive experiments on our sonar image dataset demonstrate that detector equipped with NAN and NB gains a substantial improvement as well as noise robustness compared with its baseline and other strategies. It is demonstrated that this mechanism allows for greater variation between train and test sonar datasets. Furthermore, the universality is proved on the optical image dataset. In the future, we plan to explore other inherent properties such as illumination and shadow to improve the performance of the detectors for sonar images.

6. Acknowledgements

The Project Supported by the National Natural Science Foundation of China (No: 61871124 and 61876037), The national defense Pre-Research foundation of China, by the fund of Science and Technology on Sonar Laboratory (No: 6142109KF201806), by the Stable Supporting Fund of Acoustic Science and Technology Laboratory (No: JCKYS2019604SSJSSO12).

References

[1] C. M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[2] P. Cervenka and C. De Moustier. Sidescan sonar image processing techniques. *IEEE journal of oceanic engineering*, 18(2):108–122, 1993.

[3] D. Chen, X. Chu, F. Ma, and X. Teng. A variational approach for adaptive underwater sonar image denoising. In *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, pages 1177–1181. IEEE, 2017.

[4] X. Chen and A. Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017.

[5] N. P. Chotiros. Non-rayleigh distributions in underwater acoustic reverberation in a patchy environment. *IEEE Journal of Oceanic Engineering*, 35(2):236–241, 2010.

[6] M. E. Clarke, N. Tolimieri, and H. Singh. Using the seabed auv to assess populations of groundfish in untrawlable areas. In *The future of fisheries science in North America*, pages 357–372. Springer, 2009.

[7] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

[9] G. J. Dobeck, J. C. Hyland, et al. Automated detection and classification of sea mines in sonar imagery. In *Detection and Remediation Technologies for Mines and Minelike Targets II*, volume 3079, pages 90–111. International Society for Optics and Photonics, 1997.

[10] E. Dura, Y. Zhang, X. Liao, G. J. Dobeck, and L. Carin. Active learning for detection of mine-like objects in side-scan sonar imagery. *IEEE Journal of Oceanic Engineering*, 30(2):360–371, 2005.

[11] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[12] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[14] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[16] J. Groen, E. Coiras, and D. Williams. Detection rate statistics in synthetic aperture sonar images. In *Proc. Intl. Conf. & Exh. Underwater Acoustic Measurements*, pages 367–374, 2009.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012.

[21] N. Hurtós, N. Palomeras, S. Nagappa, and J. Salvi. Automatic detection of underwater chain links using a forward-looking sonar. In *2013 MTS/IEEE OCEANS-Bergen*, pages 1–7. IEEE, 2013.

[22] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 650–657. IEEE, 2017.

[23] J. Kim, H. Cho, J. Pyo, B. Kim, and S.-C. Yu. The convolution neural network based agent vehicle detection using forward-looking sonar image. In *OCEANS 2016 MTS/IEEE Monterey*, pages 1–5. IEEE, 2016.

[24] J. Kim and S.-C. Yu. Convolutional neural network-based real-time rov detection using forward-looking sonar image. In *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*, pages 396–400. IEEE, 2016.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[27] D. Middleton. A statistical theory of reverberation and similar first-order scattered fields–i: Waveforms and the general process. *IEEE Transactions on Information Theory*, 13(3):372–392, 1967.

[28] T. Miyato, S.-i. Maeda, S. Ishii, and M. Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[29] V. Myers and J. Fawcett. A template matching procedure for automatic target recognition in synthetic aperture sonar imagery. *IEEE Signal Processing Letters*, 17(7):683–686, 2010.

[30] T. A. Palka and D. Tufts. Reverberation characterization and suppression by means of principal components. In *IEEE Oceanic Engineering Society. OCEANS'98. Conference Proceedings (Cat. No. 98CH36259)*, volume 3, pages 1501–1506. IEEE, 1998.

[31] Y. Petillot, S. Reed, and J. Bell. Real time auv pipeline detection and tracking using side scan sonar and multi-beam echo-sounder. In *OCEANS'02 MTS/IEEE*, volume 1, pages 217–222. IEEE, 2002.

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[33] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] H. Singh, J. Adams, D. Mindell, and B. Foley. Imaging underwater for archaeology. *Journal of Field Archaeology*, 27(3):319–328, 2000.

[36] X. Sun, P. Wu, and S. C. Hoi. Face detection using deep learning: An improved faster rcnn approach. *Neurocomputing*, 299:42–50, 2018.

[37] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[38] M. Valdenegro-Toro. End-to-end object detection and recognition in forward-looking sonar images with convolutional neural networks. In *2016 IEEE/OES Autonomous Underwater Vehicles (AUV)*, pages 144–150. IEEE, 2016.

[39] M. Valdenegro-Toro. Objectness scoring and detection proposals in forward-looking sonar images with convolutional neural networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 209–219. Springer, 2016.

[40] M. Valdenegro-Toro. Best practices in convolutional networks for forward-looking sonar image recognition. In *OCEANS 2017-Aberdeen*, pages 1–9. IEEE, 2017.

[41] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615, 2017.

[42] D. P. Williams. Fast target detection in synthetic aperture sonar imagery: A new algorithm and large-scale performance analysis. *IEEE Journal of Oceanic Engineering*, 40(1):71–92, 2015.

[43] D. P. Williams. The mondrian detection algorithm for sonar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):1091–1102, 2018.

[44] D. P. Williams and E. Fakiris. Exploiting environmental information for improved underwater target classification in sonar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(10):6284–6297, 2014.

[45] S. B. Williams, O. Pizarro, M. Jakuba, and N. Barrett. Auv benthic habitat mapping in south eastern tasmania. In *Field and Service Robotics*, pages 275–284. Springer, 2010.

[46] T. Xinyu, Z. Xuewu, X. Xiaolong, S. Jinbao, and X. Yan. Methods for underwater sonar image processing in objection detection. In *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*, pages 941–944. IEEE, 2017.

[47] N. Yahya, N. S. Kamel, and A. S. Malik. Subspace-based technique for speckle noise reduction in sar images. *IEEE Transactions On Geoscience and remote sensing*, 52(10):6257–6271, 2014.

[48] K. Yao, M. Mignotte, C. Collet, P. Galerne, and G. Burel. Unsupervised segmentation using a self-organizing map and a noise model estimation in sonar imagery. *Pattern Recognition*, 33(9):1575–1584, 2000.

[49] Q. Ye, H. Huang, and C. Zhang. Image enhancement using stochastic resonance [sonar image processing applications]. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 1, pages 263–266. IEEE, 2004.