Journal of *Marine Science and Engineering*

MDPI

Article

Study on Small Samples Active Sonar Target Recognition Based on Deep Learning

Yule Chen , Hong Liang * and Shuo Pang

School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: lianghong@nwpu.edu.cn

Abstract: Underwater target classification methods based on deep learning suffer from obvious model overfitting and low recognition accuracy in the case of small samples and complex underwater environments. This paper proposes a novel classification network (EfficientNet-S) based on EfficientNet-V2S. After optimization with model scaling, EfficientNet-S significantly improves the recognition accuracy of the test set. As deep learning models typically require very large datasets to train millions of model parameter, the number of underwater target echo samples is far more insufficient. We propose a deep convolutional generative adversarial network (SGAN) based on the idea of group padding and even-size convolution kernel for high-quality data augmentation. The results of anechoic pool experiments show that our algorithm effectively suppresses the overfitting phenomenon, achieves the best recognition accuracy of 92.5%, and accurately classifies underwater targets based on active echo datasets with small samples.

Keywords: underwater target classification; deep learning; small samples; EfficientNet; generative adversarial network; active sonar

Citation: Chen, Y.; Liang, H.; Pang, S. Study on Small Samples Active Sonar Target Recognition Based on Deep Learning. *J. Mar. Sci. Eng.* **2022**, *10*, 1144. https://doi.org/10.3390/jmse10081144

Academic Editor: Hugo Guterman

Received: 4 July 2022
Accepted: 14 August 2022
Published: 19 August 2022

1. Introduction

Many classification methods have been proposed to deal with the underwater target classification task. Most of them rely on manual feature extraction and preset classifiers [1,2] . However, recent years have witnessed that recognition methods based on deep learning (DL) [3] can automatically extract features. DL can not only fit target mappings directly from the original signal to learn multi-level class features, but avoid feature loss in the manual extraction process and effectively improve generalization capability of the algorithm as well. Many researchers have already extracted the time-frequency features of the echoes for underwater target recognition based on DL which have achieved effective improvement compared to traditional methods [4–7]. However, most of the above research works are on the account of passive sonar or synthetic aperture sonar (SAS). By contrast, the studies based on active sonar are relatively lacking. In addition, due to the complex underwater environment and the rapid development of mechanical noise reduction technology, it is increasingly difficult to identify underwater targets [8]. On account of the many difficulties existing in underwater target recognition based on passive sonar in practice, the research for active sonar target classification is of great significance for the underwater acoustic area.

A signal traveling in real underwater environments is interfered by random noise, water scattering, and sonar reflection properties of the target material. All of those will result in strong reverberation and noise effects on the target's echo. How to extract effective class features in real underwater scenes is the focus of research. Bu M et al. applied several pre-trained deep convolutional neural networks (DCNN) to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and experimentally demonstrated that their DCNNs outperform classical machine learning methods in active sonar target recognition [9]. Seungwoo Lee used power-normalized cepstral coefficients (PNCC) for feature extraction; this method classifies real underwater target echoes and clutter with convolutional neural

networks [10]. Karl Thomas Hjelmervik et al. found several good hyperparameter values for DL classifiers in active sonar target classification using Bayesian optimization [11].

Since DL is a typically data-driven technology, according to the previous experiments, we must use a large number of data to extract valid category features. However, it is hard to obtain such a large number of practical underwater target echoes. This shortage of samples is called the few shot learning (FSL) situation [12], which has become a universal problem that hinders the further improvement of DL. To address the problem of underwater echo recognition under small sample conditions, Henrik Berg [13] applied replicating some training examples into DL to improve the recognition accuracy of active sonar echoes, but the simple replicating did not effectively solve the overfitting problem caused by FSL. Haiwang Wang et al. [14] proposed a new convolutional neural network (CNN), IAFNet, to improve the recognition accuracy of hydroacoustic communication signals in an underwater impulsive noise environment, but the IAFNet performed poorly at low signal–noise ratio (SNR). A. Testolin [15] achieved high recognition accuracy on an active sonar echo dataset of fish in a high SNR environment by using CNN and long short-term memory (LSTM). The research exceeded the recognition accuracy of traditional machine learning methods, even the authors did not conduct extended research on the proposed models.

To solve the problems in the task of classifying underwater targets based on active sonar with small samples, we propose an algorithm that derives from deep generative adversarial networks and convolutional neural networks. The main contributions are as follows:

(1) We obtain Mel-spectrograms of the target's echoes after Mel spectrum feature extraction. A novel network structure named EfficientNet-S is proposed to achieve higher recognition accuracy with small samples.
(2) We propose a novel generative adversarial network model (SGAN) model by combining group padding and even-sized convolution kernel. The simulation experiments show that the proposed method can efficiently generate high-quality time–frequency images.
(3) We validate our algorithm on the anechoic pool experimental dataset. The result shows that it effectively solves the problem of insufficient accuracy faced with underwater target classification recognition in the case of small samples.

2. Methods

In this paper, the original echoes are processed through Mel feature extraction. Then the Mel-spectrograms with accurate labels are delivered into the generating adversarial network (SGAN). Expanded data and raw data constitute the enhanced dataset. Finally, the CNN model (EfficientNet-S) can effectively classify the underwater targets with the enhanced dataset. Figure 1 shows the structure of the classification system.

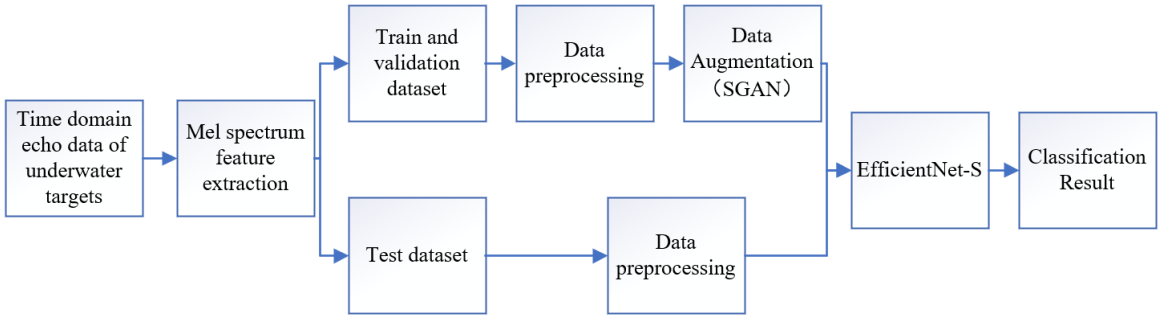

Figure 1. The structure of the proposed classification algorithm.

2.1. Setting up Dataset Based on Active Sonar Echoes

In order to simulate the real underwater target classification scene, it is necessary to build a simulation dataset that matches the real echo signals. The LFM signal is chosen as the transmitting signal, and the propagation medium is assumed uniform and homogeneous. The receiving array is a 6×6 equally spaced planar array with 0.026 m spacing. We change the conditions of reverberation and noise background to find the performance of this method at different SNRs. The SNRs of the signal are, respectively, 5 dB, 0 dB, -5 dB, and -10 dB. The left of Figure 2 shows the echoes of four types of targets at -10 dB. There are strong noise and reverberation interference with the signal. Therefore, it is hard to distinguish directly the four types of echoes only in the time domain. The spectrum is commonly chosen as one of effective feature for underwater acoustic target classification. Mel spectrum [16] has a visual representation of the spectral characteristics. It is basically the same as the human ear's speech recognition. Actually, in the past, the traditional target recognition relied on the sonar of the soldier's hearing.

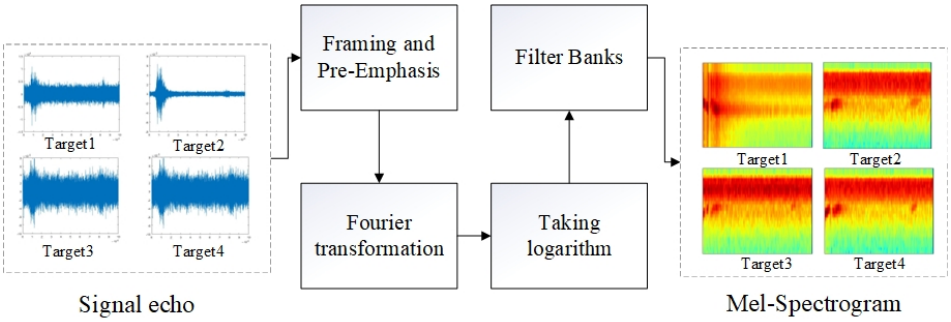

Figure 2. Mel-spectrogram extraction flow chart.

The flow block diagram of Mel feature extraction for the four types of simulated target echoes is shown in Figure 2. Firstly, the purpose of framing and pre-emphasis is to compensate for the loss of high-frequency components and boost the high-frequency components; then the FFT transformation is performed on each frame signal; finally the Mel-Spectrogram is obtained after calculating Mel frequency and energy spectrum. The signal is filtered through a set of filters, and the Mel spectrum is obtained by calculating the logarithmic energy of each filter. The transformation equation from the frequency domain to the Mel domain is (1).

$$\mathrm{Mel}(f) = 2595 * \lg(1 + \frac{f}{700}) \tag{1}$$

The Mel-Spectrogram is a two-dimensional visual plane diagram that can reflect the change of the Mel-Spectrogram with time. The horizontal ordinate of the Mel-Spectrogram represents the time, and the vertical ordinate represents the frequency. The gray value of each pixel represents the signal energy density of the corresponding frequency at a certain time.The parameters for Mel spectrum analysis in this paper are set as follows: sampling frequency, Fs = 50 kHz; number of the fast Fourier transform points, NFFT = 512; and number of Mel filters, 26.Rotate counterclockwise each type of target through 360 degrees to obtain 4×360 sets of echoes. According to the actual irradiation angle of the sonar, we select the echoes of specific angles for Mel spectrum transformation. The obtained 544 Mel-Spectrogram are randomly divided into three subsets: the training set has 380 images (about 70%), the validation set has 84 images (about 15%), and the rest of the images are used as the test set. The model was run on a desktop workstation with AMD Ryzen 9 3990X CPU, 3070 GPU, 64G of RAM, and Windows 10 as the operating system.

2.2. The Underwater Acoustic Target Recognition Method Based on Efficientnet

In this section, a baseline network model consists of the construction experience in optical image tasks, then the optimal composite scaling coefficients are obtained by model scaling. Thus, we propose our classification network model (Efficientnet-S) based on the Mel-spectrogram dataset with the optimal coefficients.

2.2.1. Baseline Model

CNN is a kind of feed-forward neural network inspired by biological neural networks. Unlike traditional neural networks, CNN consists of convolutional layers, pooling layers, nonlinear activation functions, etc. These underlying layer structures in any combination can be superimposed to obtain deep networks, and deep CNNs can extract higher-order features ideal for solving image recognition problems. EfficientNets are a network series published by Google in May 2019 [17] that combine neural architecture search (NAS) with model scaling to jointly optimize the training speed and efficiency. However, the architecture of EfficientNet is too large compared with the dataset constructed in this paper. DCNN may be inadequately trained in case of FSL, and undergo a larger risk of overfitting. Inspired by the fuzed-MBconv module of EfficientNet-V2 [18], we construct the base convolution block following the below steps: (1) using a residual structure to add a constant mapping as a branch next to the regular main road in order to prevent the network from degenerating; (2) replacing the standard 3×3 convolution kernel with a combination of several sizes of convolution kernels; (3) replacing the conventional Relu activation function with a hard swish activation function to reduce the number of parameters and computational consumption of the model; (4) regularizing with BN after each convolution operation; (5) adaptively weighting the features in the channel dimension by using a self-attentive mechanism to improve the feature representation; and (6) using a dropout strategy to alleviate model overfitting. The structure of the base convolutional block is shown in Figure 3. Based on the idea of inverted bottleneck [19], we have developed the baseline model by stacking the base convolution block. A baseline is used as a benchmark to compare with the optimization model. Its structure is shown in Table 1.

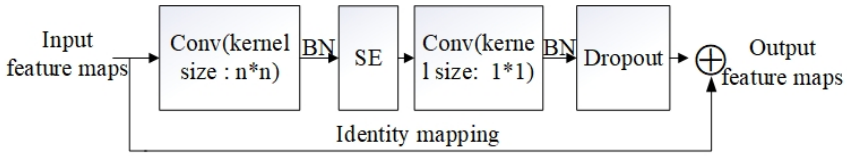

Figure 3. Base convolution block.

Table 1. Baseline model and Efficientnet-S structure table.

Stage and Operator		Channels		Repetitions	
Baseline model (B) and Efficientnet-S (S)		**B**	**S**	**B**	**S**
0	Conv3×3	24	28	1	1
1	Base-Conv, k7×7	24	28	1	1
2	Base -Conv, k5×5	96	116	1	1
3	Base -Conv, k3×3	384	460	2	3
4	Base -Conv, k3×3, SE0.25	96	116	2	3
5	Base -Conv, k3×3, SE0.25	128	154	3	4
6	Base -Conv, k3×3, SE0.25	256	308	4	6
7	Conv1×1 + Pooling + FC	4	4	1	1

2.2.2. The Optimization Model (Efficientnet-S)

Further research found that increasing the width, height, channels or layers can improve the performance of DCNN. As the idea of model scaling gradually spreads to the design of new network structures, research works have shown that (1) the width–height scaling makes the number of parameters rise exponentially; (2) channel scaling tends to

lose deep-level features; and (3) depth scaling tends to produce gradient disappearance or explosion. In a word, the expansion of a single dimension tends to encounter bottleneck. By conducting extensive NAS work, Ref. [17] systematically investigates model scaling. The authors empirically quantify the relationship among all the three dimensions of width, depth and resolution. According to formulation (2), we use five sets of coefficients for a uniform extension of the three dimensions to maximize the recognition accuracy, where N is the network model; X is the input; i is the network stage; $\hat{f}_i$ is the structure of a stage; $\hat{L}_i$, $\hat{W}_i$, $\hat{H}_i$ and $\hat{C}_i$ are the number of layers, width, height, and channels of a stage of the baseline model; and d, $w$and $r$are a set of coefficients for scaling network depth, channels, and resolution.

$$\begin{aligned} &\max_{d,w,r} Accuracy(\mathrm{N}(d,w,r)) \\ &\mathrm{N}(d,w,r) = \bigodot_{i=1\ldots s} \hat{f}_i^{d\cdot\hat{L}_i}\left(X_{(r\cdot\hat{H}i, r\cdot\hat{W}_i, \cdot\, w\cdot\hat{C}_i)}\right) \end{aligned} \tag{2}$$

The training setting of the used network is a batch size of 16; the optimizer selects the stochastic gradient descent (SGD) with the strategy of decaying cosine learning rate = 0.01. The training process is carried out for 50 epochs based on our dataset. Figure 4 shows the relationship between scaling coefficients and model parameter quantity based on the Mel-spectrogram dataset. Figure 5 shows the result of scaling the model under different SNRs. The results show that when the composite expansion parameters are set to (d, w, r) = (1.4, 1.2, 1.5), the model achieves the highest test set accuracy under different SNRs. We propose the optimal classification network model (Efficientnet-S) through this set of composite expansion parameters based on the baseline model. The structure of Efficientnet-S is shown in Table 1. Notably, EfficientNet-S model reaches above 95% accuracy with 50 epochs under SNR = 10 dB (Table 2), which achieves better accuracy than baseline model under all SNRs. It is noticeable that, as is shown in Figure 4, the model parameters will multiply as the scaling coefficients increase, but the recognition accuracy does not increase when the scaling coefficients continue to increase on the basis of Efficientnet-S (Figure 5). We speculate that this might be due to the FSL situation and high inter-category similarity of our dataset. In the circumstance of an insufficient dataset, the results of feature extraction and feature selection will differ greatly from the essential features of the target, leading to problems, such as low recognition accuracy and large bias of model parameter estimation. How to achieve the recognition and classification of underwater targets with small samples is the current problem that needs to be solved.

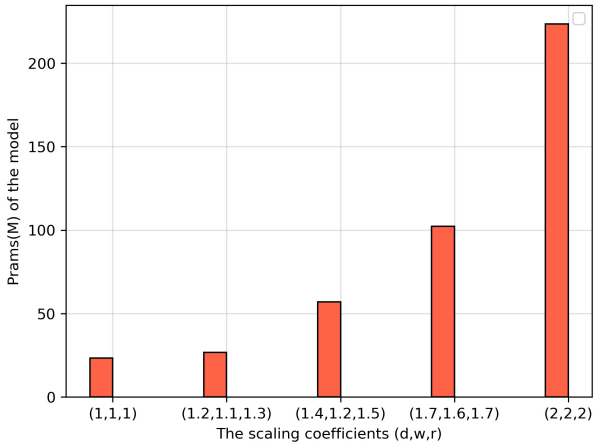

Figure 4. Relationship between parameters and scaling coefficients.

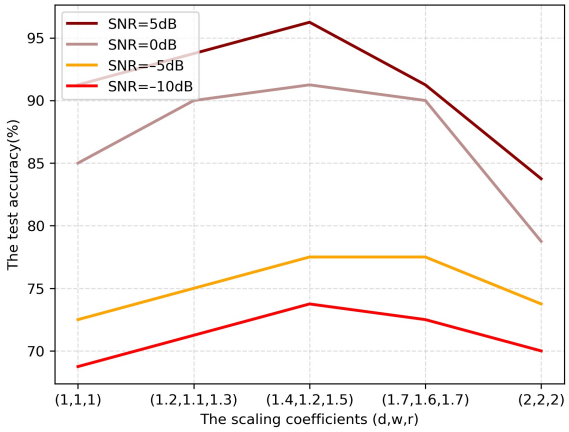

Figure 5. Scaling the model under different SNRs.

Table 2. Experimental results based on the baseline and Efficientnet-S under different SNRs.

Models	SNRs (dB)	Test Set Accuracy (%)
Baseline	5	91.25
Efficientnet-S	5	96.25
Baseline	0	85.00
Efficientnet-S	0	91.25
Baseline	−5	72.50
Efficientnet-S	−5	77.50
Baseline	−10	68.75
Efficientnet-S	−10	73.75

2.3. A Novel Generating Adversarial Networks

In the model scaling part, we speculate that the lack of samples in the Mel-spectrogram dataset may be the reason why the accuracy diminishes for bigger models. In this section, we propose a novel deep convolution generative adversarial network (SGAN) suitable for spectrograms to expand the Mel-spectrogram dataset. The simulation results show that the overfitting problem in DCNN under FSL is solved.

2.3.1. Principle of Generating Adversarial Network

Generative adversarial network (GAN) [20] has been widely used in recent years to expand the dataset for neural network training. GAN is an unsupervised model that consists of a generator model (G) and a discriminative model (D). The former can capture details of the distribution of data features, while the latter can estimate the sample's category. The task of the G is to maximize the probability of the "error in judgment" of D, while the task of the D is to accurately discriminate between the real samples and the samples generated by D.

Define the generator's distribution of the data as P_g and the prior variable of the input noise as $P_z(z)$, using $G(z;\theta_g)$ to represent the mapping of the data space, where G is a multilayer perception with the parameter θ_g. Then define $D(x;\theta_d)$ also as a multilayer perceptron to output a separate label scalar. $D(x)$ represents the probability that x comes from the real data distribution $P_{data(x)}$ instead of P_g. By training D to maximize the probability of discriminating the correct label and G to minimize $\log(1 - D(G(z)))$, the training process $V(G;D)$ of D and G is a two-person game problem with a minimization of the value function shown in Equations (3) and (4).

$$\min_G \max_D V(D,G) = E_x P_{data(x)}[\log D(x)] + E_{zP_z(z)}[\log(1 - D(z))] \tag{3}$$

$$D(x) = \frac{P_{data}(x)}{P_{data}(x) + P_g(x)} \tag{4}$$

By means of training D and G until D cannot discriminate between the data generated by G and the real data, there is $P_g(x) = P_{data}(x)$, at which point $D(x) = 0.5$, i.e., and the training process has reached its optimum.

2.3.2. Sgan Architecture

Deep convolutional generative adversarial networks (DCGAN) is a type of GAN proposed by Radford and Metz in 2015 [21]. The most significant feature of DCGAN is that it uses convolutional neural networks as the architecture to build the G and D, which solves the problems of training instability and mode collapse. In this section, we propose a new deep convolutional generative adversarial network to extend our Mel-Spectrogram dataset.

As shown in Figure 6, the generator of SGAN firstly generates 100-dimensional random noise. The data are enlarged through the fully connected layer, then reconstructed to transform the one-dimensional data into a three-dimensional matrix of 12 × 12 × 1024, doubling the feature map by the transposed convolution (TransConv) of five deconvolution layers. The final output is a feature map of 384 × 384 × 3. The discriminator performs the feature extraction by convolution operation (Conv) on the 384 × 384 × 3 size feature map outputted by the generator. Finally, the recognition result is outputted by the fully connected layer as shown in Figure 7. A batch normalization layer (BN) and an activation function (LeakyReLU) are introduced into both the generator and the discriminator. They are used to normalize the feature values, accelerate the convergence and improve the learning ability. In addition, a dropout layer is added to the discriminator to suppress the overfitting phenomenon and improve the generalization ability of the model. The structure of SGAN is shown in Figure 8.

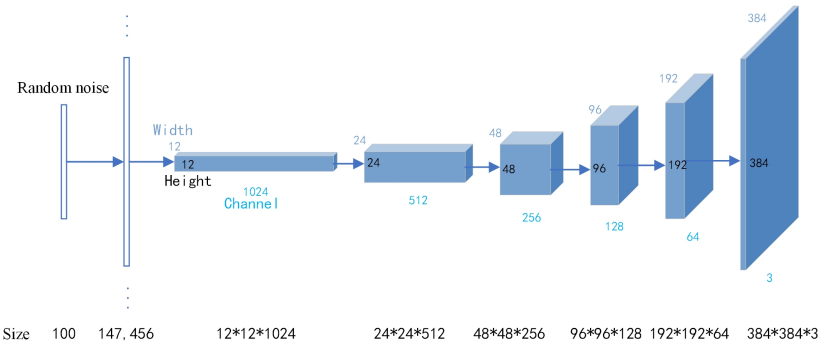

Figure 6. Generator structure.

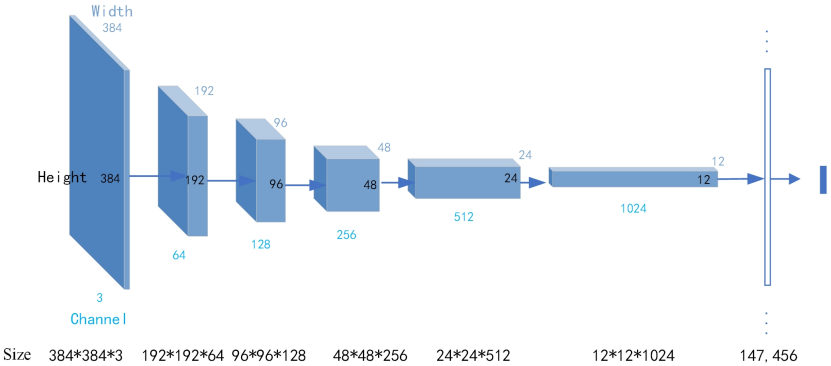

Figure 7. Discriminator structure.

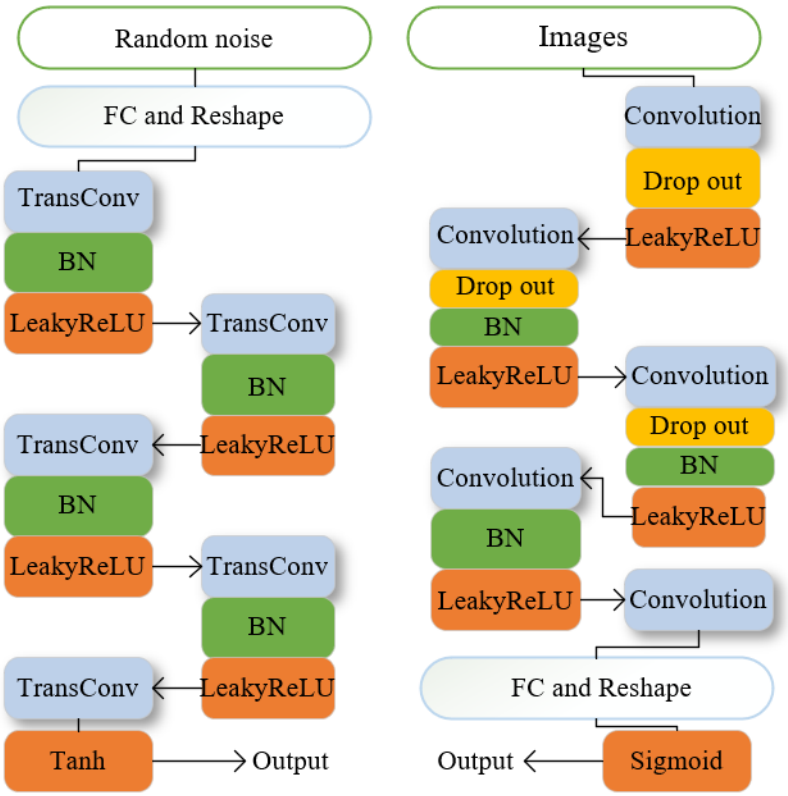

Figure 8. SGAN structure.

Compared with the conventional optical images, Mel-Spectrograms have more dense texture features and less distinct contour features, which requires less perceptual field in the deconvolution and convolution process. In consideration of this requirement, we propose to use 2×2 (C2) even-sized convolution kernels instead of the normal conventional 3×3 (C3) convolution kernels to achieve the purpose of parameter reduction. It is known from the properties of convolution that even-sized convolution kernels cause asymmetric receptive fields (RFs) that produce pixel shifts in the resulting feature maps, and such position shifts accumulate during multiple convolution superpositions and severely erode spatial information. Inspired by Ref. [22], we propose a group-symmetric padding (GP) method to divide the 1024-channel feature map into a uniform number of 256 groups in terms of channel order, and symmetrically pad the 4-channel feature map of each group. The specific steps are shown in Figure 9. The feature map size after padding is (channel, width, height) = (1024, 13, 13). Then the length and width of the matrix are expanded. Meanwhile, the dimension of the matrix is reduced by a deconvolution operation with an even-sized kernel (size = 2 × 2, stride = 2). This whole process is called (C2-GP).

To evaluate the generative effect of SGAN with various convolutional kernel sizes, we use Mel-Spectrograms of four types of targets as input in turn. 5 spectrograms of the same targets are put into SGAN each time. We record the output images after every 100 iterations of the generator G. Figure 10 shows the generation results of SGAN-C2GP with 5 Mel-Spectrograms of target2 as input. The images generated after 1000, 2000, and 3000 iterations are compared with the original input images. It can be seen that as the iterations proceed, the number of noise points in generated images is significantly reduced. Both the contours and the detail textures are closer to the original images, which visually illustrates the effectiveness of the SGAN algorithm.

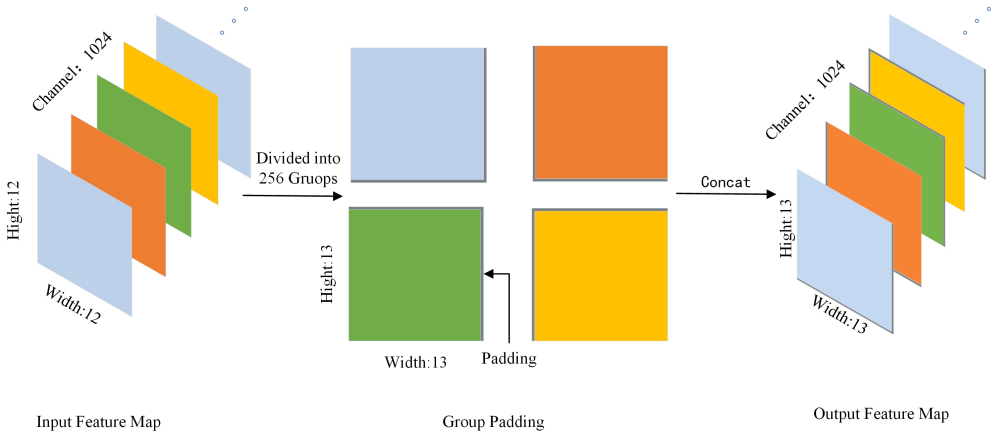

Figure 9. Group padding diagram.

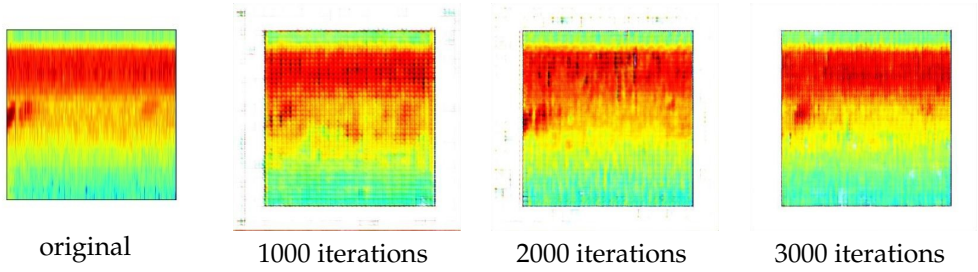

Figure 10. Generated samples under different iterations.

In addition to the subjective analysis, we also judge the quality of the generated images by calculating the peak signal-to-noise ratio ($PSNR$). The formula for calculating the $PSNR$ of two images I and K [23] is

$$PSNR = 10\log_{10}(\frac{MAX_I^2}{MSE}) \tag{5}$$

where MAX_I is the pixel maximum of picture I, the size of I and K are both $m \times n$. The mean-square error (MSE) is defined as

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{6}$$

where μ_I and μ_K are the means of I and K, σ_I and σ_K are the variances of I and K, σ_{IK} is the covariance of I and K.

Structural similarity ($SSIM$) [23] is probably currently the most popular evaluation metric for image similarity, which is commonly used in the analysis of image generation by GAN. $SSIM$ measures the brightness, contrast and structure of two images. Its calculation formula is as below.

$$SSIM(I,K) = \frac{(2\mu_I\mu_K + c_1)(2\sigma_I\sigma_K + c_2)(\sigma_{IK} + c_3)}{(\mu_I^2 + \mu_K^2 + c_1)(\sigma_I^2 + \sigma_K^2 + c_2)(\sigma_I\sigma_K + c_3)} \tag{7}$$

where the parameters setting are

$$c_1 = (0.01 \times 255)^2; c_2 = (0.03 \times 255)^2; c_3 = 0.5c_2 \tag{8}$$

From Equations (5) and (7), it can be seen that the larger the value of PSNR and the closer the value of $SSIM$ to 1, the higher the similarity of the two images, which means that the generated images are closer to the real images. We conducted ablation experiments under different SNRs (5 dB, 0 dB, −5 dB, −10 dB) to intuitively observe the impact of different convolutional kernel sizes and padding methods on the model performance, where $PSNR$ and $SSIM$ are obtained by averaging the generated images after 1000, 2000, and 3000 iterations of the generator. Time cost is the average training elapsed time for 3000 iterations. Figure 11 shows the result of scaling the model under differenet SNRs with SGAN(C2-GP). In conclusion, the simulation results at low SNR can illustrate the better performance of the model.The results of Figures 12 and 13 are obtained under SNR = −10 dB.

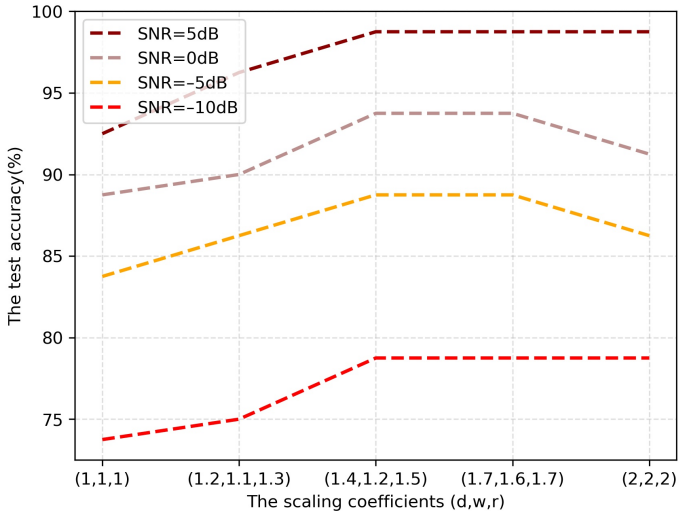

Figure 11. Scaling the model under differenet SNRs with SGAN.

From the results in Figure 12, it can be seen that the even-sized convolutional kernels (C2, C4) achieve poorer performance than the odd convolutional kernel (C3). Both PSNR and SSIM of C2, C4 are significantly lower than C3. Notably, after introduced group padding (GP), the results of GAN with C2-GP and C4-GP both have a significant improvement. In particular,the PSNR value of C2-GP reaches 20.73, while the value of SSIM is 0.607. Those results verify that the effectiveness of the group padding method proposed in this paper. The computational complexity and the number of parameters are reduced due to the reduction of the convolutional kernel size. It can be seen that C2-GP takes less time than C3, and C4-GP takes less time than C5 for the same 3000 iterations. We obtain several generated images by SGAN with different padding methods and convolution kernels for 3000 iterations. The high-quality images are selected according to the criteria of SSIM, PSNR and intuitive sense. Finally, the number of training and validation datasets for each kind of targets is expanded from 95 to 300, while the test set remain constant. The results of the simulation experiments based on the expanded dataset are shown in Figure 13. The accuracy on the expanded dataset is larger than the accuracy on original dataset. To be precise, the baseline model achieves a 2.5% accuracy improvement while Efficientnet-S achieves a 3.8% improvement on the SGAN-C2GP expanded dataset. The results are consistent with the excellent performance of SGAN-C2GP shown in Figure 12. Moreover, comparing Figure 5 with Figure 11, it not only illustrates that the dataset expansion partly solves the decreasing accuracy due to scaling, but reflects the effectiveness of the SGAN as well.

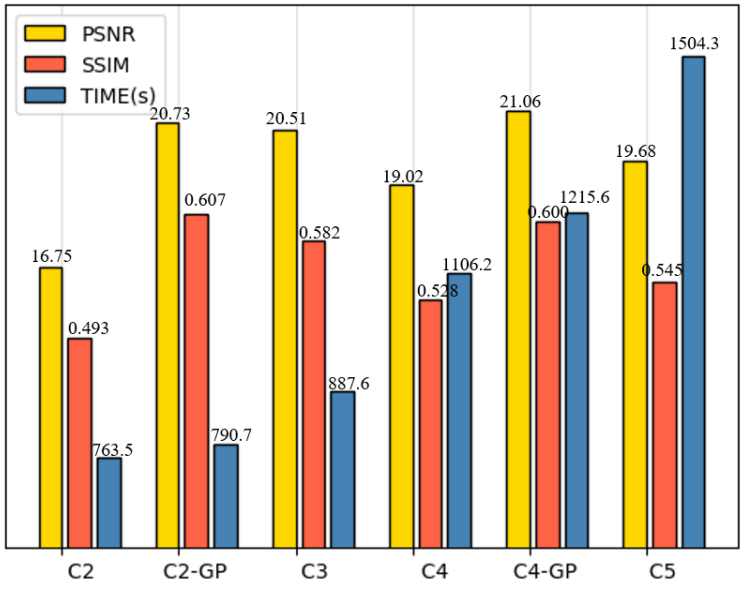

Figure 12. The results of changing the convolution kernel and padding method under SNR = −10 dB.

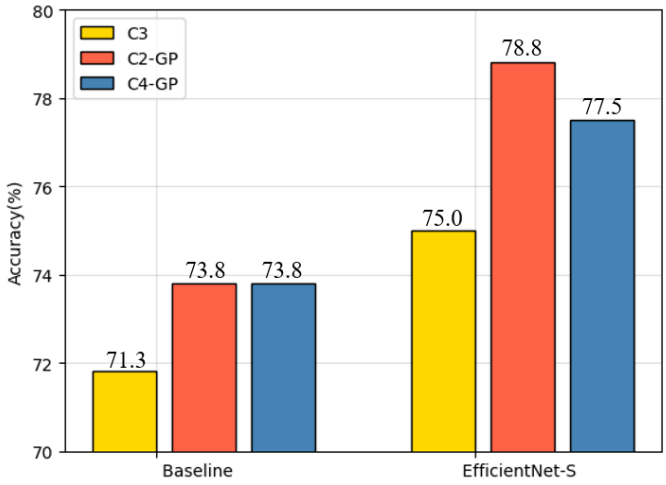

Figure 13. The results of different models after SGAN expansion under SNR = −10 dB.

3. Experiment

To test the performance of the proposed algorithm, an experimental dataset was constructed based on the active sonar echoes obtained from the anechoic pool; the model generalizability analysis was performed by comparing the results of the baseline model and EfficientNet-S. We also analyzed the result before and after introducing the generative adversarial strategy into the experimental dataset.

3.1. Experimental Dataset

The relevant experiments were all conducted in an anechoic pool. Both the target and the transducer are 2.5 m away from the water surface with 2.5 m separation distance between each other. The time-domain waveforms and Mel-Spectrograms of the four types of targets are shown in Figure 14. The experimental system is shown in Figure 15.

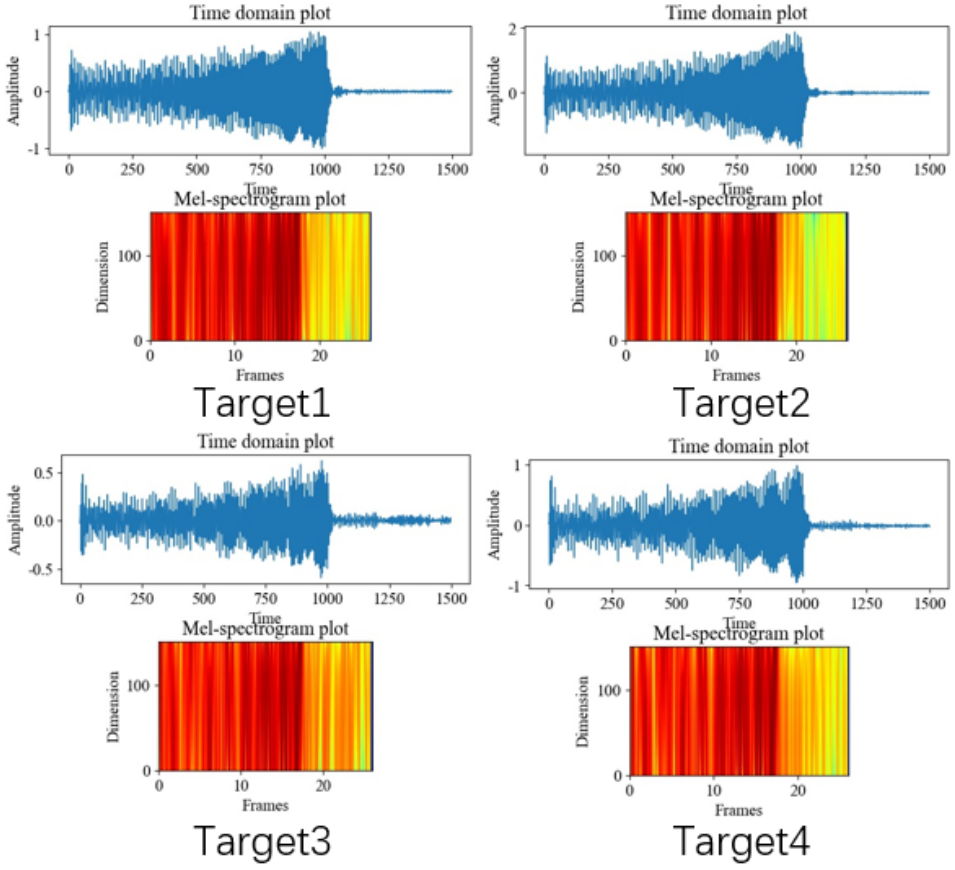

Figure 14. Schematic diagram of the underwater targets' echoes and Mel-Spectrograms.

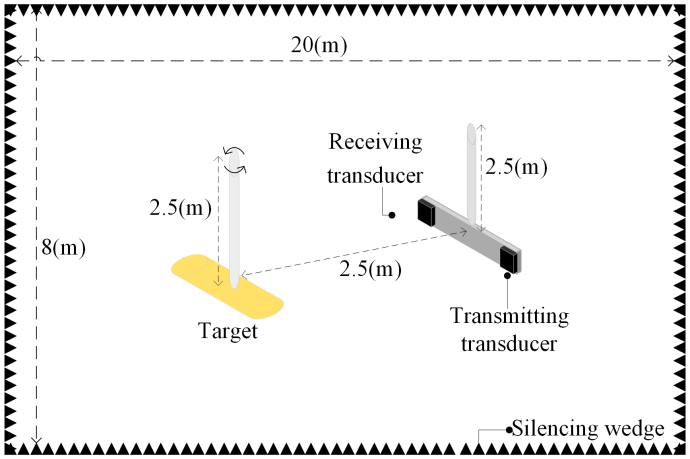

Figure 15. Block diagram of the pool experiment system.

We put four types of targets along the central axis, rotated counterclockwise in 1 degree step, and received echo data at each rotation angle by a data acquisition system. A total of 4×360 sets of echo data is collected, then divided into four groups with 136 sets of each type of targets. These data groups were randomly selected to obtain the time–frequency images using the Mel spectrum feature extraction method proposed in Section 2.1, which will construct the experimental Mel-spectrogram dataset for the following steps. The obtained 544 images were randomly divided into three subsets, where there were 380 images (about 70%) in the training set, 82 images (about 15%) in the validation set, and the remaining images that were used as the test set. The number of each type of targets in the dataset is shown in Table 3.

Table 3. Experimental data description.

Categories	Train and val	Train and val (Expanded Number)	Test	Total	Total (Expanded Number)
Target 1	116	300	20	136	320
Target 2	116	300	20	136	320
Target 3	116	300	20	136	320
Target 4	116	300	20	136	320

3.2. Model Evaluation Metrics

In target recognition, the prediction results can yield four potential predictions: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). If the predicted label value is consistent with the true label value, the result is marked as TP. Otherwise, it is marked as FP, and if there is no predict label matching the true label, it is marked as FN. TP represents the number of correctly identified targets, FP is the number of incorrectly identified targets, and FN is the number of targets that are not detected. The performance of the model can usually be evaluated by accuracy, which is calculated by Equation (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Furthermore, we use recall rate and F1Score value as performance indicators to describe the classifier. Each performance indicator is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1Score = \frac{2 * Precison * Recall}{Precision + Recall} \tag{12}$$

In addition, we also recorded the training time to measure the efficiency of a model. The training time represents the total time from the start of the model iteration to the end of training.

3.3. Result Analysis

Based on the same training environment and training strategy as in the previous section, the classification results on the experimental dataset shown in Table 4. Analysis of the results in Table 4 shows that the scaling model (EfficientNet-S) achieves the best 90% test set recognition accuracy (Experiment 2 in Table 5) with the compound scaling coefficients (d, w, r) = (1.4, 1.2, 1.5). The results illustrate that EfficientNet-S has an 11.2% improvement compared to the baseline model without scaling (Experiment 1 in Table 5). The increase of time consumption due to the increased parameters is also within acceptable limits. Instead of an increase in recognition accuracy, there is a decrease in recognition accuracy with the continued scaling of the baseline model. The computational time consumed by the model training increases exponentially with the larger scaling coefficients, which is consistent with the conclusion obtained in the previous simulation experiments.

Table 4. Scaling the baseline model with different coefficients.

Models	Scaling Factor d	Scaling Factor w	Scaling Factor r	Test Set Accuracy (%)	Training Time (s)
Baseline	1	1	1	78.8	223.5
Baseline	1.2	1.1	1.3	86.3	422.7
Baseline	1.4	1.2	1.5	**90.0**	546.8
Baseline	1.7	1.6	1.7	85.0	1500.4
Baseline	2	2	2	83.8	2721.6

Table 5. Experimental results of different models.

Experiment Serial Number	Network Model	Test Set Accuracy (%)	Training Time (s)
1	Baseline	78.8	223.5
2	Efficientnet-S	90.0	546.8
3	Efficientnet-S+SGAN	**92.5**	1203.6
4	IAFNet	73.8	411.2
5	Efficientnet-V2S	82.5	1108.7

Based on the SGAN proposed in Section 2.3, we expanded dataset based on experimental echoes. The number of training datasets for each type of targets was expanded from 95 to 300 based on the screening of generated images by SSIM, PSNR and intuitive sense. The number of test dataset remained constant. The same training strategy and initial parameters as in the previous section were used to train in the experimental dataset after the SGAN expansion (Experiment 3). To compare with other current models, the IAFNet [14] and EfficientNet-V2S [18] network models were also experimented on the original dataset (Experiments 4 and 5). The results of each experiment are shown in Table 5. Figures 16 and 17 show the graphs of validation accuracy and validation loss according to the different numbers of epochs.

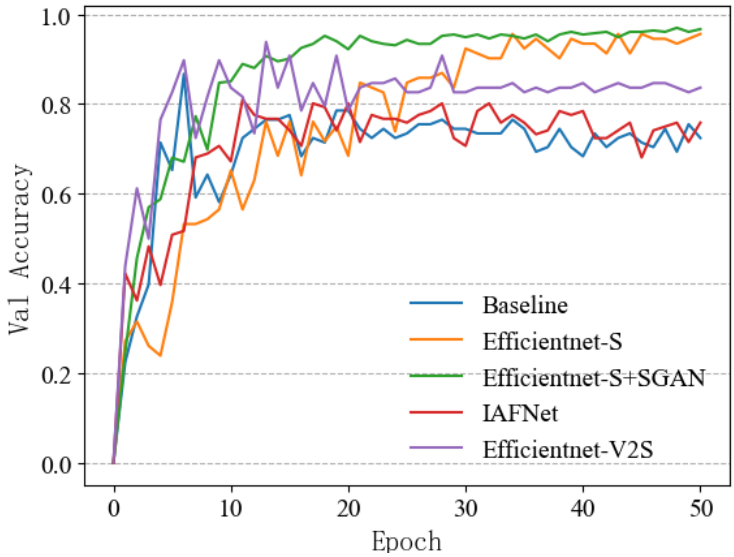

Figure 16. The validation accuracy for different models.

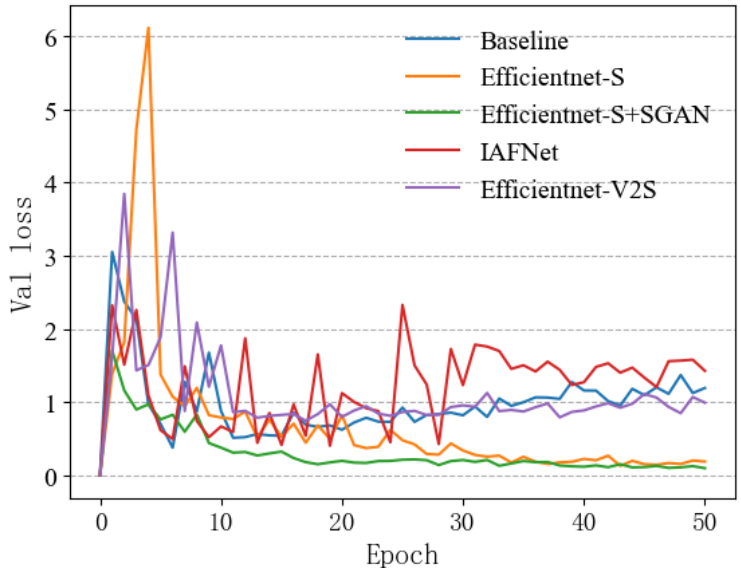

Figure 17. The validation loss for different models.

Comparing the results of Experiments 1, 2, and 3, it can be seen that the combination of Efficientnet-S and SGAN achieves 92.5% recognition accuracy in the test set. We observed that our algorithm outperformed the baseline with big margins on the experimental dataset. It also achieves an effective improvement of 2.5% compared with Efficientnet-S alone. These results suggest that the SGAN can achieve effectively regular augmentation for the Mel-Spectrograms dataset. As shown in the green lines of Figures 16 and 17, we notice that the case of the decrease in validation accuracy and the increase in validation loss no longer appear, and the training process is smoother. These results fully indicate that the overfitting problem is effectively suppressed after the data expansion by SGAN. Such a result illustrates that our algorithm can effectively solve the problem of an insufficient amount of underwater target data. Another important observation is that the continuous expansion of the data set can still reduce the overfitting to a certain extent, but it will also bring more time consumption. In contrast with Experiments 4 and 5, the result of Experiment 3 shows that the model of Efficientnet-S combined with SGAN achieved superior performance on the experimental dataset compared to the other existing algorithms. Due to the decreasing depth of the convolutional layer, the Efficientnet-S training time consumption is also significantly lower than Efficientnet-V2S. The IAFNet is a lightweight network. Although the training speed is fast, its test set accuracy is only 73.8%, which is much lower than our algorithm.

In Table 6, the average recall, precision and F1Score are respectively 92.43%, 92.50% and 0.9246. These values are all obtained by combining Efficientnet-S and SGAN. The experiment results show that the proposed algorithm has excellent recognition ability for four categories of echoes, which are consistent with the previous conclusions.

Table 6. The performance of the classifiers.

Experiment Serial Number	Network Model	Average Precision (%)	Average Recall (%)	Average F1Score
1	Baseline	78.68	78.75	0.7871
2	Efficientnet-S	90.00	90.00	0.9000
3	Eff-S+SGAN	**92.43**	**92.50**	**0.9246**
4	IAFNet	75.03	73.75	0.7438
5	Efficientnet-V2S	82.83	82.50	0.8266

4. Discussion

Due to the shortcomings of DCNN for target recognition in underwater scenes, we innovatively introduce EfficientNet-S as the basic backbone network of our algorithm to extract the target features, as well as proposing a novel generative adversarial network (SGAN) to expand the dataset. Experimental results show that the improved model proposed in this paper has excellent performance in the underwater target classification task. As is shown in Figure 18, compared with other existing networks and baselines, our algorithm effectively improves the recognition accuracy of four types of underwater targets. However, it can be seen from the confusion matrices that the recognition accuracy of our algorithm for some similar classes are relatively low, such as target 1 and target 3. These two classes are confused with each other in all of the confusion matrices. Therefore, it is worth considering how to improve the recognition accuracy of strongly confusing categories in our later work. The reverberation makes it difficult to detect targets in shallow water. The space–time reverberation modeling for active sonar array is worthy of further studies. We would discuss the performance of our algorithm when it works with the different non-Gaussian noise models.

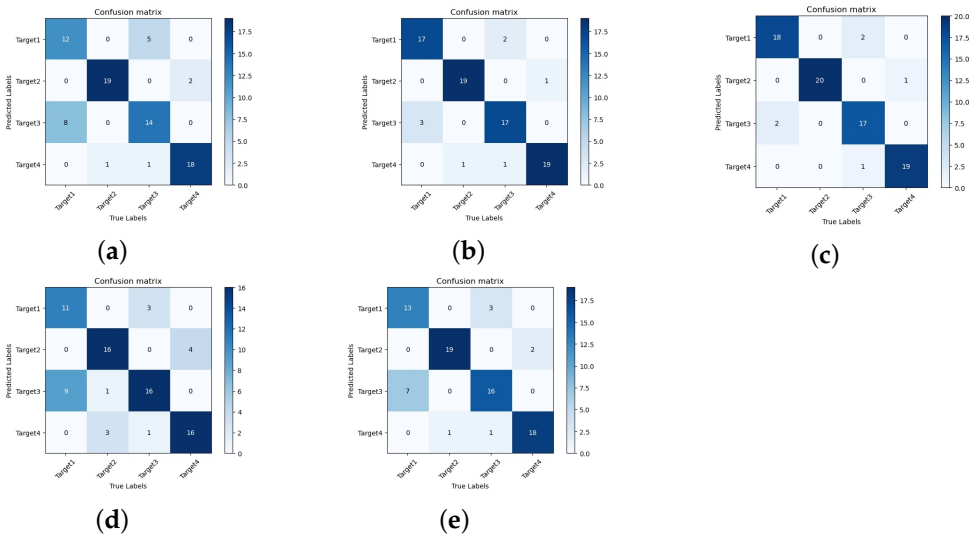

Figure 18. The confusion matrices of different models. (**a**) Baseline. (**b**) EfficientNet-S. (**c**) EfficientNet-S+SGAN. (**d**) IAFNet. (**e**) EfficientNet-v2S.

5. Conclusions

In this paper, we proposed a deep learning algorithm for the task of underwater target recognition under small samples. Mel-Spectrograms of target echoes were calculated through the Mel spectrum feature extraction method. It was regarded as the classification feature applied into the classification network. Aiming at the problem of insufficient recognition accuracy of DCNN in the process of underwater target recognition, we proposed a network model (Efficientnet-S) adjusted by compound scaling based on the baseline model. Combining the group padding and even-sized convolution kernel, the SGAN model was designed to achieve the effective augmentation of samples. The designed classification network based on the expand dataset realizes target classification with high efficiency and high performance. According to the experiments based on anechoic pool echoes, the proposed model achieves more than 90% recognition accuracy, which is a better classification performance than other current methods. The results show that our algorithm accurately recognizes real underwater targets with small samples. The advantages of the deep learning method include its stronger capability to identify underwater targets and generalization than traditional methods. In contrast, its drawbacks consist of more computational power required and expensive GPUs necessary to process a large amount of data and complex data models. For future works, researchers may opt to further improve

the DCNN models in small samples with other optimization algorithms, such as transfer learning and meta-learning. In some fields, those approaches are used to efficiently solve the problem that DL requires a huge amount of data for it to capture the key features of classification. In addition, scaling the model with more sets of parameters may also allow our model to perform better on other datasets.

Author Contributions: Data curation, Y.C.; methodology, Y.C.; project administration, H.L.; software, Y.C.; supervision, H.L.; validation, S.P.; writing—original draft, Y.C.; writing—review and editing, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under grant number 61971354.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Li, X.; Wu, Y. The classification of spherical shells with varying thickness-to-radius ratios based on the auditory perceptive features. *J. Acoust. Soc. Am.* **2019**, *145*, 1693. [CrossRef]
2. Zou, L.; Ke, T.; Zha, J. Active sonar detection using adaptive time-frequency feature. In Proceedings of the 2016 IEEE/OES China Ocean Acoustics (COA), Haerbin, China, 9–11 January 2016.
3. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
4. Yang, H.; Li, J.; Sheng, M. Underwater acoustic target multi-attribute correlation perception method based on deep learning. *Appl. Acoust.* **2022**, *190*, 108644.
5. Zhang, T.; Feng, G.; Liang, J.; An, T. Acoustic scene classification based on Mel spectrogram decomposition and model merging. *Appl. Acoust.* **2021**, *182*, 108258. [CrossRef]
6. Miao, Y.; Zakharov, Y.V.; Sun, H.; Li, J.; Wang, J. Underwater Acoustic Signal Classification Based on Sparse Time–Frequency Representation and Deep Learning. *IEEE J. Ocean. Eng.* **2021**, *46*, 952–962. [CrossRef]
7. Lakshmi, M.D.; Santhanam, S.M. Underwater Image Recognition Detector using Deep ConvNet. In Proceedings of the 2020 National Conference on Communications (NCC), Kharagpur, India, 21–23 February 2020.
8. Wei, Z.; Yang, J.; Min, S. A Method of Underwater Acoustic Signal Classification Based on Deep Neural Network. In Proceedings of the 2018 5th International Conference on Information Science and Control Engineering (ICISCE), Zhengzhou, China, 20–22 July 2018.
9. Bu, M.; Benen, S.; Kraus, D. False Alarm Reduction for Active Sonars using Deep Learning Architectures. In Proceedings of the Undersea Defence Technology (UDT), Stockholm, Sweden, 15 May 2019.
10. Lee, S.; Seo, I.; Seok, J. Active Sonar Target Classification with Power-Normalized Cepstral Coefficients and Convolutional Neural Network. *Appl. Sci.* **2020**, *10*, 8450. [CrossRef]
11. Berg, H.; Hjelmervik, K.T. Deep Learning on Active Sonar Data Using Bayesian Optimization for Hyperparameter Tuning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021.
12. Feifei, L.; Fergus, R.; Perona, P. One-shot learning of object categories. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2006**, *28*, 594–611. [CrossRef] [PubMed]
13. Berg, H.; Hjelmervik, K.T. Classification of anti-submarine warfare sonar targets using a deep neural network. In Proceedings of the MTS/IEEE Charleston OCEANS Conference, Charleston, SC, USA, 22–25 October 2018.
14. Wang, H.; Wang, B.; Li, Y. IAFNet: Few-Shot Learning for Modulation Recognition in Underwater Impulsive Noise. *IEEE Commun. Lett.* **2022**, *26*, 1047–1051. [CrossRef]
15. Testolin, A.; Kipnis, D.; Diamant, R. Detecting Submerged Objects Using Active Acoustics and Deep Neural Networks: A Test Case for Pelagic Fish. *Appl. Sci.* **2020**, *10*, 2776–2788. [CrossRef]
16. Sun, F.; Wang, M.; Xu, Q.; Xuan, X.; Zhang, X. Acoustic Scene Recognition Based on Convolutional Neural Networks. In Proceedings of the 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 19–21 July 2019.
17. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
18. Tan, M.; Le, Q.V. EfficientNetV2: Smaller models and faster training. *arXiv* **2021**, arXiv:2104.00298.
19. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–23 June 2022.
20. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Neural Information Processing Systems, Montreal, QC, Canada, 8 December 2014.

21. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the 4th International Conference on Learning Representations (ICLR 2016), San Juan, Puerto Rico, 2–4 May 2016.
22. Wu, S.; Wang, G.; Tang, P.; Chen, F.; Shi, L. Convolution with even-sized kernels and symmetric padding. In Proceedings of the NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019.
23. Alain, H.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010.