applied sciences

MDPI

Article

Underwater Object Classification Method Based on Depthwise Separable Convolution Feature Fusion in Sonar Images

Wenjing Gong [1,2,3], Jie Tian [1,3,*] and Jiyuan Liu [1,3,*]

[1] Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; gongwenjing@mail.ioa.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Key Laboratory of Science and Technology on Advanced Underwater Acoustic Signal Processing, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: tianjie@mail.ioa.ac.cn (J.T.); ljy@mail.ioa.ac.cn (J.L.)

Abstract: In order to improve the accuracy of underwater object classification, according to the characteristics of sonar images, a classification method based on depthwise separable convolution feature fusion is proposed. Firstly, Markov segmentation is used to segment the highlight and shadow regions of the object to avoid the loss of information caused by simultaneous segmentation. Secondly, depthwise separable convolution is used to learn the deep information of images for feature extraction, which produces less network computation. Thirdly, features of highlight and shadow regions are fused by the parallel network structure, and pyramid pooling is added to extract the multi-scale information. Finally, the full connection layers are used to achieve object classification through the Softmax function. Experiments are conducted on simulated and real data. Results show that the method proposed in this paper achieve superior performance compared with other models, and it also has certain flexibility.

Keywords: sonar image; Markov segmentation; depthwise separable convolution; highlight; shadow; underwater object classification

Citation: Gong, W.; Tian, J.; Liu, J. Underwater Object Classification Method Based on Depthwise Separable Convolution Feature Fusion in Sonar Images. *Appl. Sci.* **2022**, *12*, 3268. https://doi.org/10.3390/app12073268

Academic Editor: Alexander Sutin

Received: 21 February 2022
Accepted: 22 March 2022
Published: 23 March 2022

1. Introduction

Underwater automatic target recognition (ATR) technology is widely used in military and civil fields, and object classification in sonar images is an important research field [1–3]. The highlighted echo and black shadow in sonar images are the important attributes of the object, both of which have rich features and are the focus in the field of object classification. In ref. [4], by minimizing the Euclidean distance between the points on the shadow contour and the hyper ellipse, the shadow is fitted as a hyper ellipse. The hyper ellipse parameters are then used to achieve the classification of the object in side-scan sonar images. In ref. [5], the mean clustering method is used to separate the shadow region in the image and fuse their features into the classification process as an auxiliary feature set. Different from refs. [4,5], the highlight regions in sonar images are segmented by level set in [6], then the invariant moment features are extracted and SVM is used to realize the recognition of ships.

The above research indicates the significance of highlight and shadow regions in sonar images for object classification. However, these studies achieve object classification by using only the highlight or shadow region. Due to the large interference of underwater noise, small objects are often submerged in the noise. Conversely, in the process of sonar image acquisition, there may be no shadow in the image, or objects of different shapes may produce the same shadow at some special angles. Therefore, it has certain limitations to realize object classification only relies on the highlight region or shadow region in sonar images.

On this basis, the method based on feature description [7] segments sonar image into the highlight region and shadow region, and then finds the optimal classification feature set based on their features. In ref. [8], the image is divided into two parts by

fuzzy morphology, and the object classification is achieved through the extracted shape features combined with Markov Chain Monte Carlo theory (MCMC). In addition, several other studies have used the idea of segmenting the sonar image into different parts for classification or detection [9–11], but there are some deficiencies in feature extraction. In recent years, a convolutional neural network has a mature application in the field of image processing [12,13], such as image denoising, image restoration, and image separation. It also has been used in the classification of objects in sonar images. For example, a convolutional neural network is used [14] to classify moment features extracted from a shadow region which are obtained by graph cut algorithm. The method of transfer learning [15] is introduced to realize classification, and the requirement of the amount of data is reduced while training the network. Therefore, the use of depthwise convolution may be to improve the classification performance by joint using highlight and shadow features of the underwater object in sonar images.

In this paper, Markov random field theory is used to segment highlight and shadow regions of the object in sonar images respectively. The features of the two regions are extracted through the depthwise separable convolution. Then, the extracted feature maps are fused at the last convolution layer. The fusion features are used to achieve the object classification through full connection layers. The contribution of the proposed method is stated below. The object information is preserved to a large extent by segmenting the object image separately. The feature extraction capability and lightweight advantage of depthwise separable convolution are utilized to improve the classification capability of the network. Pyramid pooling is added in front of the full connection layer to extract the multi-scale information. This method can also be used to achieve classification by a single region when the highlight is blocked or the shadow is missing.

The remainder of this paper is organized as follows. In Section 2, we introduce the structure of the underwater object classification method proposed in this paper. The methods and material including the image classification method, improved depthwise separable convolution, single feature extraction network, and classification network are presented in Section 3. Data set description and experimental results for the classification tasks are demonstrated in Section 4. The conclusions of this work are highlighted in Section 5.

2. Structure of Underwater Object Classification Method

The classification system of underwater object joints using highlight and shadow in sonar images is shown in Figure 1. Starting from the whole process of object classification, three steps are mainly considered in model structure design, namely the image segmentation part, feature extraction part, and fusion classification part. The main content of the image segmentation part is to complete the segmentation of the input images. The feature extraction part mainly introduces the depthwise separable convolution and its improved methods, including the introduction of activation function, by aiming at the construction of a single feature extraction network. The fusion classification part mainly includes the fusion and classification process of the previous feature extraction network results. The specific implementation process of each part is as follows:

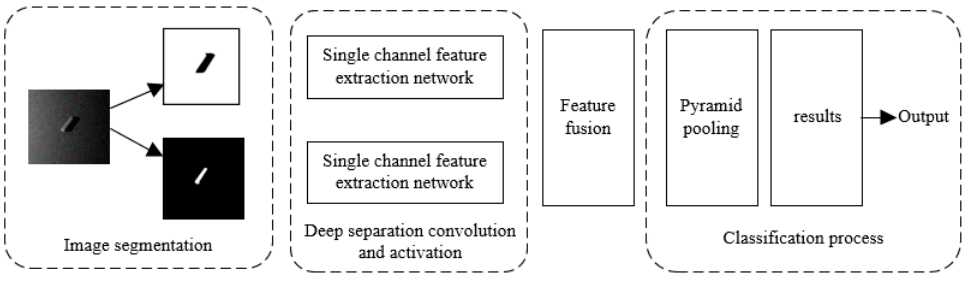

Figure 1. Joint classification system of highlight and shadow of small objects in underwater sonar images.

2.1. Image Segmentation

It is usually necessary to segment the sonar image before classifying suspicious objects in the images. The image segmentation algorithm based on Markov random field makes use of the relevant information between adjacent pixels, which can describe the local statistical characteristics of the image and segment the image with noise effectively. Markov random field theory combined with the conditional iterative algorithm is used to segment the image preliminarily. The morphological corrosion and expansion operation is applied for post-processing. After obtaining the highlight and shadow of the object in the sonar image, the binary image dataset of the object is established.

2.2. Feature Extraction

Feature extraction is mainly achieved by improved depthwise separable convolution. In this part, the defect of depthwise separable convolution is improved, and the feature extraction network is designed and optimized according to the image characteristics. The binary images of the segmented highlight and shadow segmented by method in Section 2.1 are input into the feature extraction network to obtain the depth convolution features of the object.

2.3. Fusion Classification

In the process of fusion classification, the above convolution features are fused by a connection function and predicted by the Softmax classification function. The depth convolution features extracted from the two parts are fused to obtain multi-scale features, which are input into the Pyramid pooling. Then the feature vector of the object is put into the full connection layer to obtain the classification result of the object.

3. Implementation of Proposed Method

The implementation process of the method proposed in this paper is described. Specifically, the segmentation method of sonar image is introduced in Section 3.1; then, the basic depthwise separable convolution is provided in Section 3.2, including its shortcomings and improvement measure, as well as the design of single feature extraction network; finally, the parallel classification network architecture is explained in Section 3.3, and the process of feature fusion and object classification is further presented.

3.1. Highlight and Shadow Segmentation

It is well known in the machine learning community that the interference of background information may affect the accuracy of network classification if the original sonar image is directly used as the input of the network. Conversely, the huge amount of computation brought by the large image size will increase the network training and prediction time. Based on these reasons, the original sonar image is firstly segmented before input into the network. This can make the network pay attention to the key regions in the sonar image that affect the classification decision to improve the accuracy and efficiency of object classification.

The image segmentation algorithm based on Markov random field uses Gibbs field and maximum posterior probability MAP to achieve image segmentation, which has good robustness to noise images. For each sonar image, the image segmentation algorithm of the Markov random field (MRF) [16–19] is used in this paper. Hammersley–Clifford theorem [20] proves that the joint probability of Markov random fields obeys Gibbs distribution, so it can be obtained:

$$p(x|\beta) = \frac{1}{Z(\beta)} \exp(-H(x|\beta)), \tag{1}$$

where $Z(\beta)$ is the regular constant and H is the energy function defined by potential function V_c and a nonnegative scalar constant β:

$$H(x|\beta) = \sum_c V_c(x_c|\beta), \tag{2}$$

After that, a conditional iterative model (ICM) algorithm is applied to optimize the noise and model parameters to obtain the region of highlight and shadow. Morphological expansion and corrosion operations [21] are used to reprocess the segmentation results, and then the small connected regions were removed to obtain a relatively complete binary image of highlight and shadow. The segmentation result of a measured object in a synthetic aperture sonar (SAS) image is shown in Figure 2. Figure 2a is the original sonar image, Figure 2b is the corresponding highlight region of the object, and Figure 2c is the shadow image of the object. Through the above operations, the highlight and shadow regions of the underwater object are completely segmented.

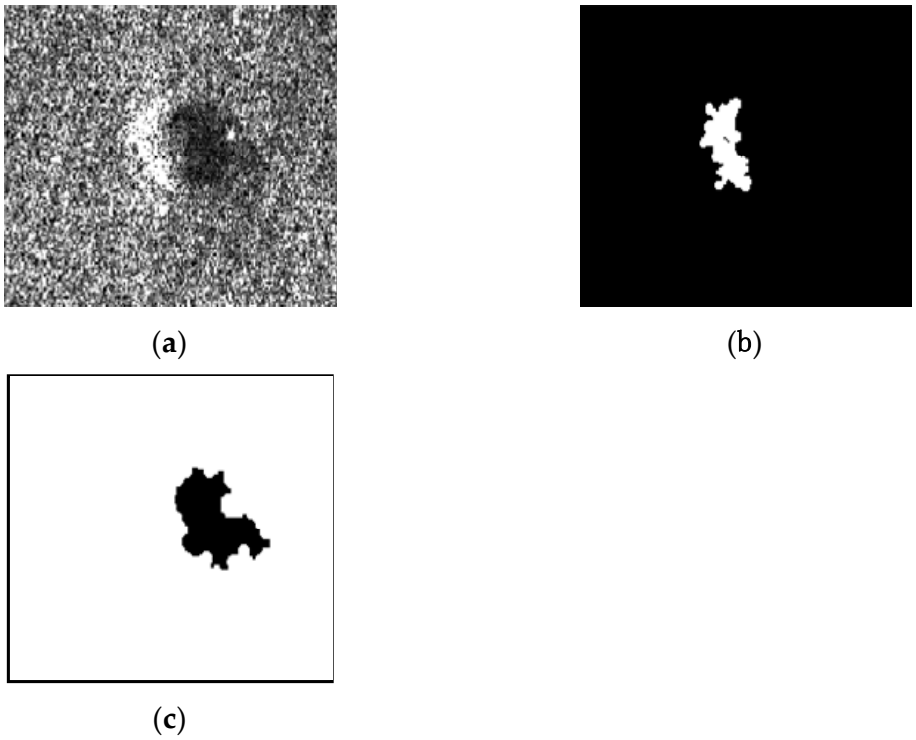

Figure 2. Synthetic aperture sonar image object segmentation results. (**a**) Original sonar image. (**b**) The highlight area of the object. (**c**) The shadow area of the object.

3.2. Single Feature Extraction Network

3.2.1. Depthwise Separable Convolution

The standard convolution operation has the effect of filtering features based on the convolutional kernels and combining features to produce a new representation. The filtering and combination steps can be split into two steps via the use of factorized convolutions called depthwise separable convolutions for substantial reduction in computational cost. Depthwise separable convolution proposed by Howard in [22] are made up of two layers: depthwise convolutions and pointwise convolutions. Depthwise convolutions are applied on each input channel, and pointwise convolutions are then used to create a linear combination of the output of the depthwise layer.

As shown in Figure 3a, a standard convolutional layer takes as input a $D_F \times D_F \times M$ feature map F and produces a $D_F \times D_F \times N$ feature map G where D_F is the width and height of the input feature map, M is the number of input channels, D_G is the width and height of the output feature map, and N is the number of the output channel. The standard convolutional layer is parameterized by convolution kernel K of size $D_K \times D_K \times M \times N$, where D_K is the size of the kernel, M is the number of input channels, and N is the number of output channels. Standard convolutions have the computational cost of:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F, \quad (3)$$

where the computational cost depends multiplicatively on the number of input channels M, the number of output channels N, the kernel size $D_K \times D_K$, and the feature map size $D_F \times D_F$. Depthwise separable convolution decomposes a standard convolution into a depthwise convolution and a pointwise convolution [22]. It can break the interaction between the number of output channels and the size of the kernel [23–25].

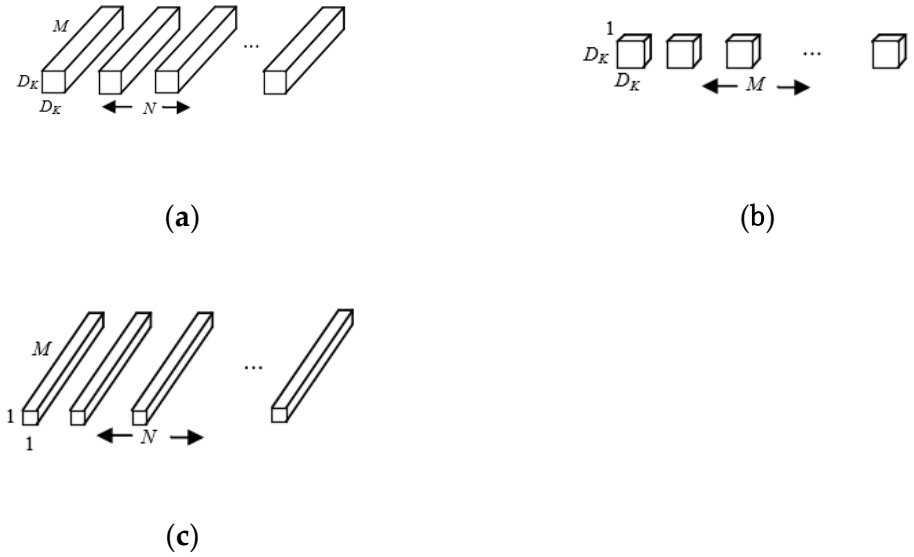

Figure 3. Schematic diagram of standard convolution and depth separable convolution. (**a**) Standard convolution process. (**b**) Depthwise convolution process. (**c**) Pointwise convolution process.

Depthwise convolution in Figure 3b has a computational cost of:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F, \tag{4}$$

It only filters input channels, it does not combine them to create new features. So, the pointwise convolution computes a linear combination of the output of depthwise convolution via 1×1 convolution and generates new features [11], as seen in Figure 3c. Pointwise convolution has a computational cost of:

$$M \cdot N \cdot D_F \cdot D_F, \tag{5}$$

Therefore, the cost of depthwise separable convolutions can be described as:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F, \tag{6}$$

which is the sum of the depthwise and pointwise convolutions. By expressing convolution as the process of filtering and combining, we obtain a reduction in the computation of:

$$(D_K \cdot D_K \cdot M \cdot D_F + M \cdot N \cdot D_F \cdot D_F)/(D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F) = 1/N + 1/D_K^2, \tag{7}$$

When 3×3 depthwise separable convolutions are used, the computation can be reduced to one-ninth of the standard convolutions. Compared with standard convolution, using depthwise separable convolution can save computing resources better, reduce classification time, and improve classification performance.

3.2.2. Improved Depthwise Separable Convolution

To extract features better, reduce the dependency between parameters, and alleviate the phenomenon of overfitting, the nonlinear activation function of Rectified Linear Unit (ReLU) [26], which is defined as Equation (8), is used after each depthwise convolution and pointwise convolution in the feature extraction network to strengthen the nonlinear

expression ability. At the same time, a Batch Normalization layer (BN) [27] is usually added before the activation function to accelerate the convergence of the network and prevent gradient explosion. It can also improve the accuracy of the model. The calculation formula of layer BN can be seen from Equation (9), where $\sum_{i=1}^{m} x_i$ is the output of the convolution layer, μ_B and σ_B^2 are mean and variance respectively, and y_i is the output result after normalization.

$$ReLU(x) = \max(x, 0), \tag{8}$$

$$\begin{cases} \mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2, \\ \hat{x}_i = (x_i - \mu_B)/\sqrt{\sigma_B^2 + \varepsilon} \\ y_i \leftarrow \gamma \hat{x}_i + \beta \end{cases} \tag{9}$$

However, the use of ReLU activation function may leave part of the neural network in a "dead" state, resulting in the loss of information. Suppose that there is a large gradient in the forward conduction process of the network so that the weight of the neural network is greatly updated, a negative value will be given by neuron for all inputs. Then the output of this negative value becomes zero after ReLU, and the gradient flowing through this neuron will always be zero. In this case, the neuron remains inactive, and the weight cannot be updated so that the network will not be able to learn normally.

In order to avoid the occurrence of this phenomenon, we improve the network activation function, and replace the ReLU activation function with linear activation function after pointwise convolution [28,29]. The linear activation function is expressed as Equation (10), where x is the input, $f(x)$ represents the output, W represents the weight of the function, and b is the bias.

$$f(x) = Wx + b, \tag{10}$$

Linear activation does not increase the computation amount of depthwise separable convolution compared with ReLU activation function, so it retains the advantage in computation amount. More importantly, the linear output can preserve the information of each channel and provide more reliable features for subsequent object classification.

3.2.3. Structure of Single Feature Extraction Network

The feature extraction network designed in this paper mainly uses the improved depthwise separable convolution described in Section 3.2.2, and has the following considerations: (i) The dataset of sonar images is small, so fewer network layers are used to reduce the complexity of the network; (ii) The size of the convolution kernel is closely related to the amount of model parameters and computation. Larger convolution kernel will lead to excessive model parameters and loss of image details, so we tend to use a small convolution kernel in this paper; and (iii) Pooling layer will lose several original features while controlling for overfitting, so it is necessary to reduce the use of pooling layer in feature extraction.

All the above considerations are to make the network more suitable for underwater object classification and achieve better results. The structure of the single feature extraction network designed in this paper is shown in Figure 4, which has two traditional convolutions and seven depthwise separable convolution blocks (Block). Among the structure, each Block is composed of several improved depthwise separable modules. The network layer structure and parameters of the improved depthwise separable module are shown in Table 1, which is composed of three different convolutions. The structure and parameters of the whole feature extraction network are shown in Table 2, where t is the extension factor of the convolution layer in the depthwise separable module, c and n are the numbers of output channels of the network layer and the repetition times of the depthwise separable module, and s is the convolution stride.

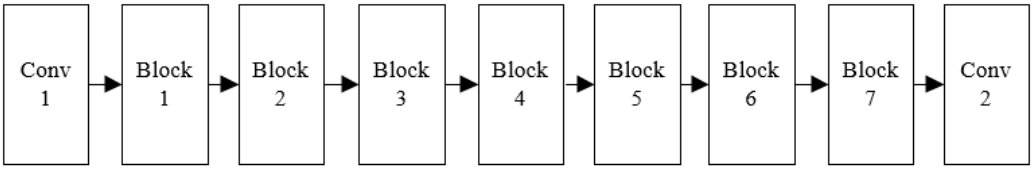

Figure 4. Structure of single feature extraction network.

Table 1. Structure of improved depth separation module.

Layer	Kernel Size	Activation Function
Convolution	1×1	ReLU
Depthwise Convolution	3×3	ReLU
Pointwise Convolution	1×1	Linear

Table 2. Structure of feature extraction network.

Layer	*t*	*c*	*n*	*s*
Conv1	-	32	1	2
Block1	1	16	1	1
Block2	6	24	2	2
Block3	6	32	3	2
Block4	6	64	4	2
Block5	6	96	3	1
Block6	6	160	3	2
Block7	6	320	1	1
Conv2	-	128	1	1

3.3. Fusion Classification Network

In order to make full use of the object features extracted by the single feature extraction network in Section 2.2 and achieve better results, we fuse the features of the two parts and use the fusion features for object classification. In the process of feature extraction, the shallow features of the network reflect more original information of the image, while the deep features of the network are more suitable for object classification. Therefore, we complete the fusion operation in the position of the last convolution layer of the single feature extraction network. In general, the classification of objects can be realized by using the full connection layer with fixed input channel. However, the structure of the image may be damaged if we adjust all images to the same size forcibly when the size of the sonar image is different. Spatial pyramid pooling [30,31] can convert a feature map of any size into a feature vector of fixed size, which can be sent to the full connection layer. Moreover, pyramid pooling can extract multi-scale features and improve the effect of object classification [32,33].

Pyramid pooling is added to the classification part, and the structure of the classification network is seen in Figure 5. Firstly, the output of the last convolution layer of the two single feature extraction networks is fused by Equation (11), where $input_1$ is the highlight images of the input, F_1 is the single feature extraction network, and $F_1(input_1)$ is the highlight feature map of the last layer of the feature extraction network. The shadow region is in the same way. F is the fusion feature map to realize the combination of highlight and shadow features.

$$F = Concate(F_1(input_1), F_2(input_2)), \tag{11}$$

Secondly, after fusing the features, a pyramid pooling layer is added to obtain the features of fixed size. Thirdly, full connection layer1 is added to reduce data dimension, and full connection layer2 is applied to realize object classification.

In order to improve the network fitting, the dropout function is added after both full connection layers to randomly drop out some cells; and the drop-out rate is set to 0.5. The output value of multiple categories can be converted into probability distribution in the

range of [0, 1] by Softmax function, as shown in Equation (12), where z_i is the output value of the ith node and C is the number of output nodes, the number of classification categories.

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_i}}, \tag{12}$$

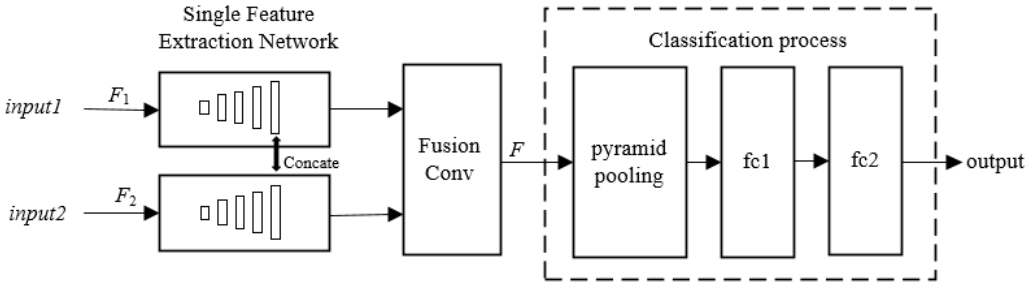

Figure 5. Structure of classification network.

4. Experiments and Analysis

The experimental dataset includes object images of three shapes, which are used for training and testing under different models to verify the accuracy of the proposed method in the underwater object classification task. The GPU of the experiment computer is RTX2070 and the CPU is 6-core i7-10750H. The network is built on Keras deep learning framework and accelerated by CUDNN.

4.1. Dataset

The dataset used includes real sonar images collected by lake and sea trials and simulated images obtained by three-dimensional modeling software, including sphere, cylinder, and truncated cone. The specific number of which is shown in Table 3. Simulated images are used for auxiliary training of the model. During the simulation, the grazing angle between sonar and the object is 30~45°, and the angle between the object axis and the incident acoustic wave is 0~180°. Some experimental images can be seen in Figure 6.

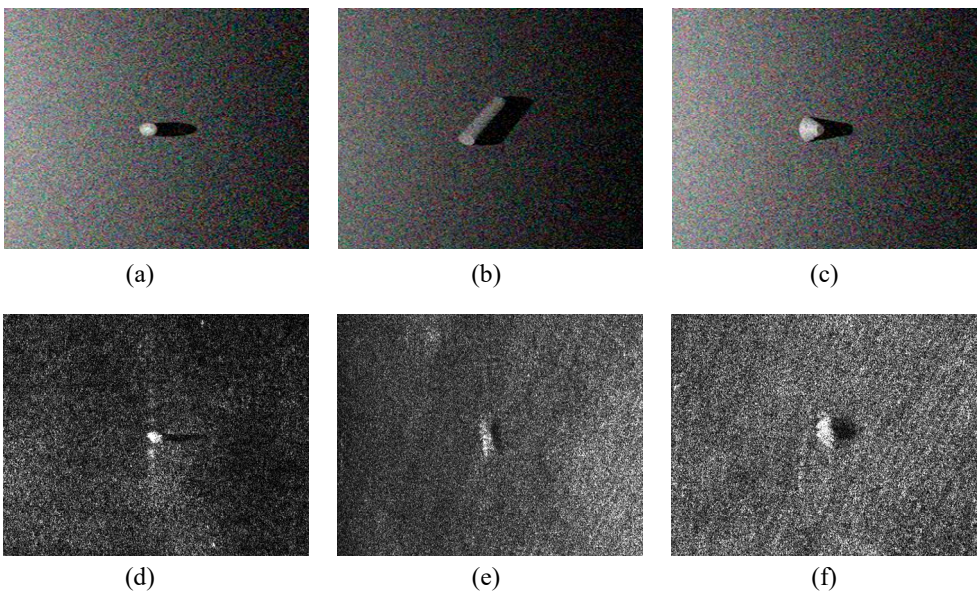

Figure 6. Image of experimental dataset. (**a**) Simulated sphere object. (**b**) Simulated cylinder object. (**c**) Simulated truncated cone object. (**d**) Real sphere object. (**e**) Real cylinder object. (**f**) Real truncated cone object.

Table 3. Number of experimental data set.

Shape	Simulated Images	Real Images
sphere	1200	24
cylinder	2410	18
truncated cone	2410	27

4.2. Experiment1: Performance of Single-Channel Classification Network with Different Feature Extraction Module

The whole network model proposed is a parallel fusion network, including two input layers, two parallel feature extraction networks, a fusion layer, and a classification layer. The two channels input different images respectively. However, the existing classification network is usually a single-channel network with only one input. Therefore, to objectively compare them with other models, we use the single-channel classification network composed of one input layer, one feature extraction network, and a classification layer compared with other models. Except that the feature extraction module is different (VGG16, Resnet50 and feature extraction network in this paper), the other structures are the same. The experimental data are highlight images segmented from simulated data of the object with three shapes. It is worth noting that the classification layer of VGG16 and Resnet itself is the same as that used in this paper, so the comparison of the feature extraction network is equivalent to the comparison of the whole single-channel model.

In this experiment, 80% of the images are randomly selected for training, and the remaining are used as a test set to verify network performance. The cost function of the network is classified as cross entropy to calculate the distance between the predicted value and the real label. Optimizer RMSProp is used to optimize the parameters of the whole network. During network training, the dropout rate is set to 0.5, the batch size is 16, the learning rate is 0.0001, and the number of iterations is 100. After feature extraction is completed, classification results are obtained by using the full-connection classification layer. The object classification performance on the single-channel classification network with different feature extraction modules is shown in Table 4.

Table 4. Classification performance of single-channel classification network with different feature extraction modules.

Model	Train Time/ms	Test Time/ms	Flops/M	Paras/M	Accuracy/%
VGG16	6	3	781	65.1	90.7
Renet50	8	2	549	45.7	80.7
This Paper	2	0.7	23	1.9	90.1

We use the network training time (Train Time), test time (Test Time), the amount of calculation (Flops), the number of parameters (Paras), and classification accuracy (Accuracy) to measure the classification performance of the network. The train time and test time of the network refers to the time that the network classifies each image in the training and testing process. The amount of computation and the number of parameters are two important indicators to measure the complexity of the model. The amount of computation corresponds to the time complexity of the model, that is, the length of the network execution time. The number of parameters corresponds to the space complexity of the model, that is, the amount of computer memory that needs to be occupied. Accuracy is the degree to which the model classifies objects correctly.

Results in Table 4 show that VGG16, as a feature extraction network, has the best classification accuracy of 90.7%, but its time cost is also high when the classification time is 6 ms during training and the test time is up to 3 ms per image. Moreover, its amount of calculation and the number of parameters are the largest of these models. When using Resnet50 for feature extraction, its time complexity and memory occupation are generally smaller than VGG16; however, its accuracy is also reduced to a certain extent. Due to

that, it is not suitable for the underwater object classification task. The model with feature extraction network designed in this paper reduces the classification time to some extent on average, compared with the other two networks, in achieving object classification. That is, the classification accuracy is only 0.6% worse than VGG16. Meanwhile, it has the least computational cost and the lightest model structure, which is more suitable for the object classification task.

4.3. Experiment2: Validation of Joint Classification Network

Using the network proposed in Figures 4 and 5, experiments are conducted on simulated images and real data. The parameters of the network are the same as previous experiments and the experimental process is as follows: (i) Using the joint classification method proposed in this paper, the binary images of highlight and shadow are input in pairs into the feature extraction network shown in Figure 4 for feature extraction; the network shown in Figure 5 is used to complete feature fusion and classification; (ii) the feature extraction network shown in Figure 4 is used to input the segmented highlight, shadow, and original images respectively, and realize classification process through this single network.

The average classification accuracy obtained after ten experiments is shown in Table 5. Results show that the classification accuracy obtained by feature extraction using the original image of the object is the lowest with 67.1% of real data, followed by the shadow image of the object with 81.3%. When using highlight images for object classification, the accuracy of simulated data is increased by 11.4% and 6.6%, respectively. Compared with original and shadow images, the accuracy of real data is increased by 18.6% and 4.4%, respectively. The joint classification method proposed in this paper has the highest classification accuracy, reaching 93.1% and 90.8% respectively. The reason may be that the segmented images are binary, and the extracted features are shape features. In some angles, different objects will have the same shape, which is easy to cause the possibility of misclassification. The method proposed in this paper uses the feature information of both two parts at the same time to get higher classification accuracy.

Table 5. Classification accuracy of simulation and real data.

Input	Simulated Images/%	Real Images/%
Original Image	78.7	67.1
Shadow Image	86.7	81.3
Highlight Image	90.1	85.7
Joint of Both	93.1	90.8

4.4. Experiment 3: Robustness of Joint Classification Network

In the acquisition of sonar images, the object may not have shadows due to angle or other reasons. In order to verify the classification performance of the proposed network, classification experiments are conducted on simulated and real data. The network training uses a pair of the complete highlight-shadow image. The test process only uses a single highlight/shadow image, and the missing image is recorded as an empty array. The classification results of this experiment are shown in Table 6.

Table 6. Classification accuracy of test sets.

Input	Simulation Images/%	Real Images/%
Shadow image Only	83.6	80.0
Highlight image Only	86.7	83.1

Under the absence of the highlight region, the joint classification method proposed in this paper only uses a shadow image to realize the classification. The accuracy of a simulated and real image is approximately 83.6% and 80%. Under the absence of a shadow region, the classification accuracy of the proposed joint classification network is 86.7 and

83.1%, respectively. Compared with the results of the experiment on normal data, the above classification accuracy of using only one kind of image is relatively reduced by using the proposed joint classification network. This may be due to the certain interference of the empty array input to the network. Nevertheless, the lowest accuracy of the network can still reach about 80%, which has certain robustness for the underwater object classification task compared with the single network that loses the classification effectiveness under such circumstances.

5. Conclusions

A new method for underwater object classification is studied in this paper. In this method, Markov random field theory is used to segment the highlight and shadow regions of the underwater sonar image respectively, which preserves the complete information of the object compared with some methods that segment highlight the region and shadow simultaneously. Secondly, the depthwise separable convolution is used to automatically extract the features of the two parts, which reduces the computation and complexity of the network, avoids the problem of incomplete manual feature extraction, reduces the classification time, and improves efficiency. In addition, the feature fusion is realized by parallel network structure, which makes good use of the features of shadow and highlight regions. According to the characteristics of sonar imaging and the real shape of the objects, the simulation data set is established, which is more effective for underwater target classification tasks when using transfer learning.

In conclusion, this method makes use of the advantages of depthwise separable convolution in feature extraction, avoids the defects of manual feature extraction, could adapt to different image sizes of the dataset, and makes full use of the effective information in the images. The experimental results are as follows: (1) Compared with VGG16 and Resnet50 models, the classification model with feature extraction network designed in this paper has better overall classification performance. The classification accuracy reaches 90.1%, the average calculation cost is reduced by dozens of times, and the model is the lightest with only 1.9 M. (2) The classification results of the proposed classification method based on depthwise separable convolution feature fusion is higher than that of the single-channel network. The accuracy is improved by 7.9% and 12.7% on average in the simulation and real sonar images, which also proves the effectiveness of the parallel network architecture. (3) In the case of partial data missing, the method proposed in this paper is still effective, and the classification accuracy can reach at least 80%, indicating that the network has certain robustness.

However, there are also some limitations and deficiencies in the research process. For example, there is a certain gap between the classification accuracy of the simulated and the real sonar image. It is the essential difference between the two images that causes this difference, which also provides an important idea for future research. Therefore, our subsequent research will focus on the following aspects. Firstly, the characteristics of an object in sonar images will be studied to find a more reliable image simulation method. Secondly, a more effective feature extraction network will be designed according to the sonar images. Finally, a classification method suitable for the complex underwater environment will be proposed to achieve object classification.

Author Contributions: J.T. gave academic guidance to this research work and put forward feasible suggestions for the research content. At the same time, the manuscript has been modified by J.L., in the content and structure of the article has given a reasonable method of modification. W.G. designed the core method proposed in this paper, wrote the program, carried out relevant experimental verification, and drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Institute of Acoustics, Chinese Academy of Sciences, under a project entitled, "Intelligent Classification of Underwater Objects in Sonar Images". The number of the funding is E1511301.

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. Zhai, H.; Jiang, Z.; Zhang, P.; Tian, J.; Liu, J. Underwater object highlight segmentation in SAS image using Rayleigh mixture model. In Proceedings of the 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 27–29 November 2015; pp. 418–423.
2. Lu, Z.; Chen, Y.C.; Zhang, T.D.; Yu, J. A Sonar Image Segmentation Algorithm based on Two-Dimensional Spatio-Temporal Fuzzy Entropy. In Proceedings of the 2018 IEEE 8th International Conference on Underwater System Technology: Theory and Applications (USYS), Wuhan, China, 1–3 December 2018; pp. 1–5. [CrossRef]
3. Pramunendar, R.A.; Wibirama, S.; Santosa, P.I. Fish Classification Based on Underwater Image Interpolation and Back-Propagation Neural Network. In Proceedings of the 2019 5th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 30–31 July 2019; pp. 1–6. [CrossRef]
4. Dura, E.; Bell, J.; Lane, D. Superellipse Fitting for the Recovery and Classification of Mine-Like Shapes in Sidescan Sonar Images. *IEEE J. Ocean. Eng.* **2008**, *33*, 434–444. [CrossRef]
5. Kumar, N.; Mitra, U.; Narayanan, S.S. Robust Object Classification in Underwater Sidescan Sonar Images by Using Reliability-Aware Fusion of Shadow Features. *IEEE J. Ocean. Eng.* **2015**, *40*, 592–606. [CrossRef]
6. Xu, W.H.; Xu, Y.J.; Dong, L.L.; Ying, L. Level-set and SVM based target recognition of image sonar. *Chin. J. Sci. Instrum.* **2012**, *33*, 49–55.
7. Fandos, R.; Zoubir, A.M. Optimal Feature Set for Automatic Detection and Classification of Underwater Objects in SAS Images. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 454–468. [CrossRef]
8. Lopera, O.; Dupont, Y. Automated target recognition with SAS: Shadow and highlight-based classification. In Proceedings of the 2012 Oceans, Yeosu, Korea, 21–24 May 2012; pp. 1–5. [CrossRef]
9. Reed, S.; Petillot, Y.; Bell, J. An automatic approach to the detection and extraction of mine features in sidescan sonar. *IEEE J. Ocean. Eng.* **2003**, *28*, 90–105. [CrossRef]
10. Sinai, A.; Amar, A.; Gilboa, G. Mine-Like Objects detection in Side-Scan Sonar images using a shadows-highlights geometrical features space. In Proceedings of the OCEANS 2016 MTS/IEEE Monterey, Monterey, CA, USA, 19–23 September 2016; pp. 1–6. [CrossRef]
11. Hou, B.; Luo, X.H.; Wang, S.; Jiao, L.; Zhang, X. Polarimetric SAR images classification using deep belief networks with learning features. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 2366–2369. [CrossRef]
12. Jia, F.; Xu, J.; Sun, X.; Ma, Y.; Ni, M. Blind Image Separation Method Based on Cascade Generative Adversarial Networks. *Appl. Sci.* **2021**, *11*, 9416. [CrossRef]
13. Chen, E.Z.; Wu, X.M.; Wang, C.Y.; Du, Y. Application of Improved Convolutional Neural Network in Image Classification. In Proceedings of the 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 8–10 November 2019; pp. 109–113. [CrossRef]
14. Zhu, K.Q.; Tian, J.; Huang, H.N. Underwater objects classification method in high-resolution sonar images using deep neural network. *Acta Acust.* **2019**, *44*, 595–603.
15. William, D.P. Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. In Proceedings of the 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2497–2502.
16. Tian, X.D.; Liu, Z.; Li, L. Study of Sonar Image Segmentation Based on Markov Random Field. In Proceedings of the 2006 6th World Congress on Intelligent Control and Automation, Dalian, China, 21–23 June 2006; pp. 9618–9622. [CrossRef]
17. Li, J.; Jiang, P.; Zhu, H. A Local Region-Based Level Set Method with Markov Random Field for Side-Scan Sonar Image Multi-Level Segmentation. *IEEE Sens. J.* **2021**, *21*, 510–519. [CrossRef]
18. Cao, J.Z.; Song, A.G. Research on the texture image segmentation method based on Markov random field. *Chin. J. Sci. Instrum.* **2015**, *36*, 776–786.
19. Song, Y.T.; Ji, Z.X.; Sun, Q.S. Brain MR Image Segmentation Algorithm Based on Markov Random Field with Image Patch. *Acta Autom. Sin.* **2014**, *40*, 1754–1763.
20. Liu, A.P.; Fu, K.; You, H.J.; Liu, Z. SAR Image Segmentation Based on Multiscale Auto Regressive and Markov Random Field Models. *J. Electron. Inf. Technol.* **2009**, *31*, 2557–2562.
21. Rebhi, A.; Abid, S.; Fnaiech, F. Fabric defect detection using local homogeneity and morphological image processing. In Proceedings of the 2016 International Image Processing, Applications and Systems (IPAS), Hammamet, Tunisia, 5–7 November 2016; pp. 1–5. [CrossRef]

22. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
23. Hung, K.W.; Zhang, Z.; Jiang, J. Real-Time Image Super-Resolution Using Recursive Depthwise Separable Convolution Network. *IEEE Access* **2019**, *7*, 99804–99816. [CrossRef]
24. Srivastava, H.; Sarawadekar, K. A Depthwise Separable Convolution Architecture for CNN Accelerator. In Proceedings of the 2020 IEEE Applied Signal Processing Conference (ASPCON), Kolkata, India, 7–9 October 2020; pp. 1–5. [CrossRef]
25. Hoang, V.; Hoang, V.; Jo, K. Realtime Multi-Person Pose Estimation with RCNN and Depthwise Separable Convolution. In Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 6–7 April 2020; pp. 1–5.
26. Bousbai, K.; Merah, M. A Comparative Study of Hand Gestures Recognition Based on MobileNetV2 and ConvNet Models. In Proceedings of the 2019 6th International Conference on Image and Signal Processing and their Applications (ISPA), Mostaganem, Algeria, 24–25 November 2019; pp. 1–6.
27. Thakkar, V.; Tewary, S.; Chakraborty, C. Batch Normalization in Convolutional Neural Networks—A comparative study with CIFAR-10 data. In Proceedings of the 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, India, 12–13 January 2018; pp. 1–5.
28. Stursa, D.; Dolezel, P. Comparison of ReLU and linear saturated activation functions in neural network for universal approximation. In Proceedings of the 2019 22nd International Conference on Process Control (PC19), Strbske Pleso, Slovakia, 11–14 June 2019; pp. 146–151. [CrossRef]
29. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
31. Jose, A.; Lopez, R.D.; Heisterklaus, I.; Wien, M. Pyramid Pooling of Convolutional Feature Maps for Image Retrieval. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 480–484. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *IEEE Access* **2015**, *11*, 234–241.
33. Nie, Q.Q.; Xiao, B.; Bi, X.L.; Li, W. Multi-focus Image Fusion Algorithm Based on Super Pixel Level Convolutional Neural Network. *J. Electron. Inf. Technol.* **2021**, *43*, 965–973.