remote sensing

MDPI

Article

Detection of Small Floating Target on Sea Surface Based on Gramian Angular Field and Improved EfficientNet

Caiping Xi [1] and Renqiao Liu [2,*]

[1] College of Automation, Jiangsu University of Science and Technology, Zhenjiang 212100, China
[2] Ocean College, Jiangsu University of Science and Technology, Zhenjiang 212100, China
* Correspondence: 202030029@stu.just.edu.cn

Abstract: In order to exploit the advantages of CNN models in the detection of small floating targets on the sea surface, this paper proposes a new framework for encoding radar echo Doppler spectral sequences into images and explores two different ways of encoding time series: Gramian Angular Summation Field (GASF) and Gramian Angular Difference Field (GADF). To emphasize the importance of the location of texture information in the GAF-encoded map, this paper introduces the coordinate attention (CA) mechanism into the mobile inverted bottleneck convolution (MBConv) structure in EfficientNet and optimizes the model convergence by the adaptive AdamW optimization algorithm. Finally, the improved EfficientNet model is used to train and test on the constructed GADF and GASF datasets, respectively. The experimental results demonstrate the effectiveness of the proposed algorithm. The recognition accuracy of the improved EfficientNet model reaches 96.13% and 96.28% on the GADF and GASF datasets, respectively, which is 1.74% and 2.06% higher than that that of the pre-improved network model. The number of parameters of the improved EfficientNet model is 5.38 M, which is 0.09 M higher than that of the pre-improved network model. Compared with the classical image classification algorithm, the proposed algorithm achieves higher accuracy and maintains lighter computation.

Keywords: radar target detection; sea clutter; image classification; deep learning

Citation: Xi, C.; Liu, R. Detection of Small Floating Target on Sea Surface Based on Gramian Angular Field and Improved EfficientNet. *Remote Sens.* **2022**, *14*, 4364. https://doi.org/10.3390/rs14174364

Academic Editor: Kacem Chehdi

Received: 25 July 2022
Accepted: 31 August 2022
Published: 2 September 2022

1. Introduction

The backscattered echoes from the sea surface received after a radar electromagnetic wave irradiation to the sea surface are defined as sea clutter [1]. Target detection in the context of sea clutter is of great importance in both military and civil fields, where the detection of small floating targets on the sea surface is one of the key research directions. The complex sea conditions and the working state of the radar have a great influence on the sea clutter, which makes the sea clutter waves have very complex space-time variation characteristics, such as non-uniform, non-Gaussian, and non-smooth characteristics [2], which presents a great challenge to the target detection technology in the sea clutter background. Small targets floating on the sea surface have two distinctive characteristics: small radar cross section (RCS) and slow-motion speed. The smaller RCS leads to weaker echoes, so the target information is often drowned in the sea clutter. Small targets floating on the sea surface have slower motion speed and simpler motion pattern, which makes the Doppler features of their echoes often submerged in the wider Doppler spectrum of sea clutter, so it is difficult to distinguish between target echoes and sea clutter on the Doppler spectrum.

Traditional techniques for target detection in the background of sea clutter are usually based on a particular test statistic, for example, Chen et al. [3] proposed a target detection method based on the non-extensive entropy of the Doppler spectrum by using the aggregation difference between sea clutter and the target on the Doppler spectrum. However, the detection effect of this method is not satisfactory when more clutter information is

mixed in the target Doppler spectrum, or when the aggregation difference between sea clutter and target Doppler spectrum is not obvious. According to the difference of texture change between sea clutter and target in time-Doppler spectra (TDS) images, a detection method based on local binary pattern (LBP) feature space is proposed in the literature [4]. However, this detection method requires a long time to accumulate observations and the detection results are unreliable when the target to clutter deviation ratio is low. A single statistical feature is often difficult to comprehensively describe the difference between the target echoes and sea clutter. The target detection method based on fusing multiple statistical features makes up for the shortcomings of the single feature-based detection method to a certain extent. For example, Xu et al. [5] proposed a method for detecting small targets at sea based on fusing multiple features in the time domain and frequency domain, which improves the detection accuracy to some extent; the literature [6] proposes a feature-expandable detection method in combination with spike suppression techniques, increasing the number of discrepancy features used to six. However, the multi-feature fusion method requires good complementary characteristics among statistical features, and it is also difficult to achieve the desired effect in the case of low signal-to-clutter ratio (SCR) and short-term observation accumulation. Moreover, it is difficult to achieve good generalization ability of the algorithm by using specific human-selected statistical features.

The rapid development of deep learning methods in recent years has provided a new research idea for target detection in the background of sea clutter. Convolutional neural network (CNN) can mine the abstract features contained in images; it not only has high accuracy classification ability and excellent generalization ability, but also does not require human intervention in the process of image feature extraction, so it has been gradually applied and developed in the field of radar. For example, a method based on CNN was proposed in the literature [7] for the detection and classification of micro-motion targets in the background of sea clutter, and the results verified that CNN has the advantage of high accuracy and intelligence in the recognition of radar echo signal time-frequency map feature extraction.

In order to take advantage of deep learning in the field of image classification, this paper explores a new framework for encoding Doppler spectral sequences of radar echoes into images, transforming the time-ordered target detection problem into an image classification problem. To find the best image encoding method for Doppler spectral sequences, two different encoding methods are explored in this paper: Gramian Angular Summation Field (GASF) and Gramian Angular Difference Field (GADF) [8]. To emphasize the importance of the texture information of the main locations in the encoded graph, this study attempts to add the CA [9] mechanism to the MBConv module of EfficientNet [10], a lightweight network model, and uses the adaptive AdamW optimization algorithm [11] to optimize the model convergence. Finally, the validation on GADF and GASF datasets shows that the proposed algorithm not only has high recognition accuracy but also has fewer model parameters, and the proposed algorithm has significantly better accuracy than similar optimal algorithms on the basis of lightweight computing.

The sections of this paper are organized as follows: Section 2 describes the experimental data and research methods used in the experiments; Section 3 presents the experimental details and findings; the discussion is organized in Section 4; the paper ends with a conclusion in Section 5.

2. Materials and Methods

This study aims to propose a framework for classifying radar echo data using deep learning techniques. In this study, the raw radar echo data are first segmented, and then the segmented data are fast Fourier transformed to obtain the corresponding Doppler spectral sequence, followed by applying GASF and GADF to encode the Doppler spectral sequence into an image. Then, the transformed images are fed into the CA-EfficientNet-B0 network model incorporating the CA attention module for training, and finally the images to be

detected are fed into the trained model for processing to identify the features in the images and perform classification.

This detection algorithm is divided into three main stages: (1) data preprocessing; (2) sequence coding; and (3) model training and prediction. Figure 1 shows the block diagram of the improved EfficientNet-based algorithm for detecting small targets floating on the sea surface. The details of this algorithm are described in the subsequent subsections.

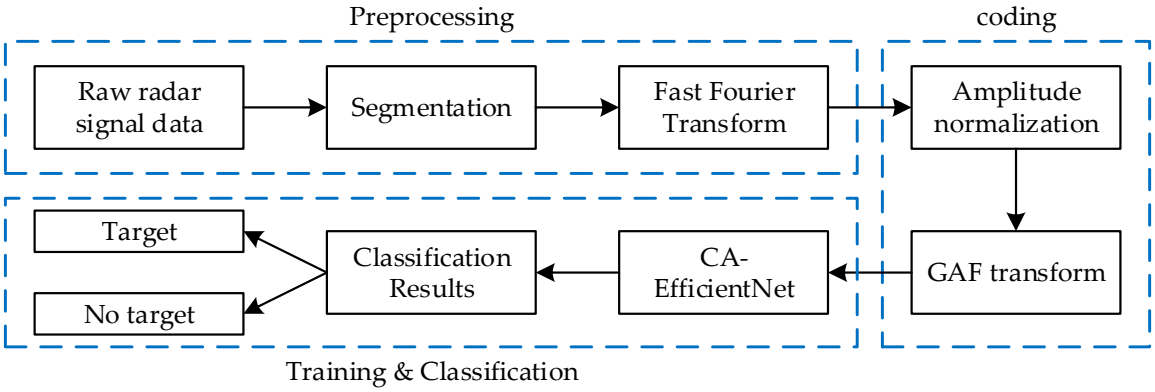


Figure 1. Block diagram of the proposed detection algorithm.

2.1. Data Description

The measured sea clutter data used in this paper come from the IPIX (Intelligent Pixel processing X-band) radar database website of McMaster University in Canada [12]. The dataset was collected in 1993, measured with X-band fully coherent radar. The IPIX radar can transmit and receive electromagnetic waves of horizontal polarization (H-polarization) and vertical polarization (V-polarization) with center frequency of 9.39, peak power of 8, pulse repetition frequency (PRF) of 1 kHz, pulse width of 0.2, range resolution 30, and the number of pulses is 131072. Each group of data includes data of four polarization modes, VV, HH, VH and HV, and each polarization mode data includes echo signals of 14 range cells. The target is within one of the range cells, usually 2–3 range cells adjacent to the target cell are affected cells. The target is an airtight spherical container of diameter 1 wrapped in aluminum foil, capable of floating on the sea surface and moving randomly with the waves. The cell where the target is located is labeled as the primary target cell, the cell affected by the target is labeled as the secondary cell, and the rest of the cells are clutter-only cells. For ease of use, these 14 datasets are numbered and their file names, wind speeds, effective wave heights, primary target cells, and affected cells are listed in Table 1.

Table 1. 1993 IPIX radar data main parameter description.

Data Label	Name of Datasets	Primary Target Cell	Affected Cell	Wind Speed (km/h)	Wave Height (m)
#17	19931107_135603_starea	9	8, 10, 11	10	2.10
#18	19931107_141630_starea	9	8, 10, 11	10	2.08
#19	19931107_145028_starea	8	7, 9	12	2.05
#25	19931108_213827_starea	7	6, 8	9	1.01
#26	19931108_220902_starea	7	6, 8	9	1.03
#30	19931109_191449_starea	7	6, 8	19	0.89
#31	19931109_202217_starea	7	6, 8, 9	15	0.89
#40	19931110_001635_starea	7	5, 6, 8	9	0.91
#54	19931111_163625_starea	8	7, 9, 10	20	0.66
#280	19931118_023604_stareC0000	8	7, 9, 10	11	1.44
#283	19931118_035737_stareC0000	10	8, 9, 11, 12	0	1.30
#310	19931118_162155_stareC0000	7	6, 8, 9	33	0.90
#311	19931118_162658_stareC0000	7	6, 8, 9	33	0.90
#320	19931118_174259_stareC0000	7	6, 8, 9	27	0.91

2.2. Signal to Clutter Ratio Analysis

For the 14 sets of measured datasets in Table 1, the SCR in short time intervals fluctuated around the Average SCR(ASCR) [4]. Assuming that the target echo and sea clutter of the target cell are independent of each other, the ASCR of each dataset under any polarization can be calculated by the following equation [13]:

$$\mathrm{ASCR} = 10\log 10\left(\frac{2^{-17}\sum_{n=1}^{2^{17}}|x(n)|^2 - \overline{p}_c}{\overline{p}_c}\right) \tag{1}$$

where $\overline{p}_c$ represents the average power of sea clutter, estimated from the time series of all clutter cells of length 2^{17}.

In this paper, we analyzed the ASCR of the target cell for 14 datasets under four polarization models, and the results are shown in Figure 2. As can be seen from Figure 2, the 14 datasets have completely different ASCR. In the same dataset, the radar echoes of HH, HV and VH polarization have higher ASCR than those of VV polarization, which is related to the clutter level under different polarization modes, the polarization characteristics of the target, the sea state and the radar-illuminated area of the target during the data acquisition. The ASCRs of the 14 datasets under the four polarization methods are distributed in the range of $[-5\mathrm{dB},\ 20\mathrm{dB}]$, so these 14 datasets can be well used to measure the target detection performance of the proposed algorithm in this paper.

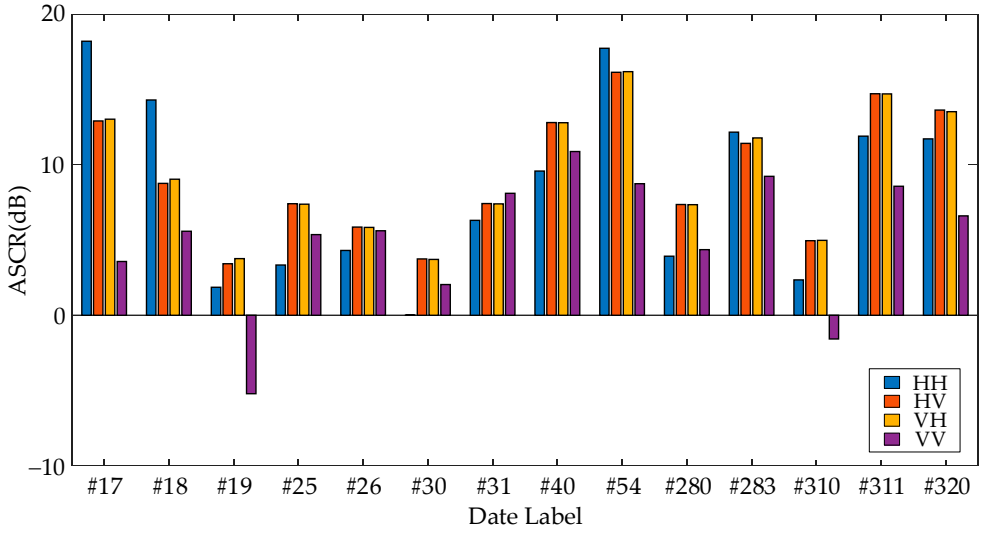


Figure 2. Average SCRs of the primary cells of the 14 datasets under four polarizations.

2.3. Analysis of Doppler Spectrum Characteristics

At each range cell, the received radar echo consists of the target, sea clutter, and noise. The echoes obtained by the IPIX radar can be divided into two categories: (1) If the measured range cell contains a target, the received echo consists of the target echo, noise, and sea clutter; (2) If the measured range cell does not contain a target, then the received echo contains only sea clutter and noise. At high SCR, the Doppler spectrum of target echo and sea clutter usually has an obvious difference: under the same sea condition, compared with sea clutter Doppler spectrum, the Doppler spectrum of target echo has stronger aggregation, lower central frequency, and more concentrated energy distribution.

Taking the data file #280VV as an example, this paper selects cell 1 (sea clutter cell) and cell 8 (target cell) as a comparison, and splits the point sequence of each distance cell into 255 segments, with the length of each segment being 1024 and the overlap rate of two adjacent segments being 50%, and then performs 1024-point fast Fourier transform (FFT) and normalization on the segmented data to obtain its normalized Doppler spectrum as shown in Figure 3.

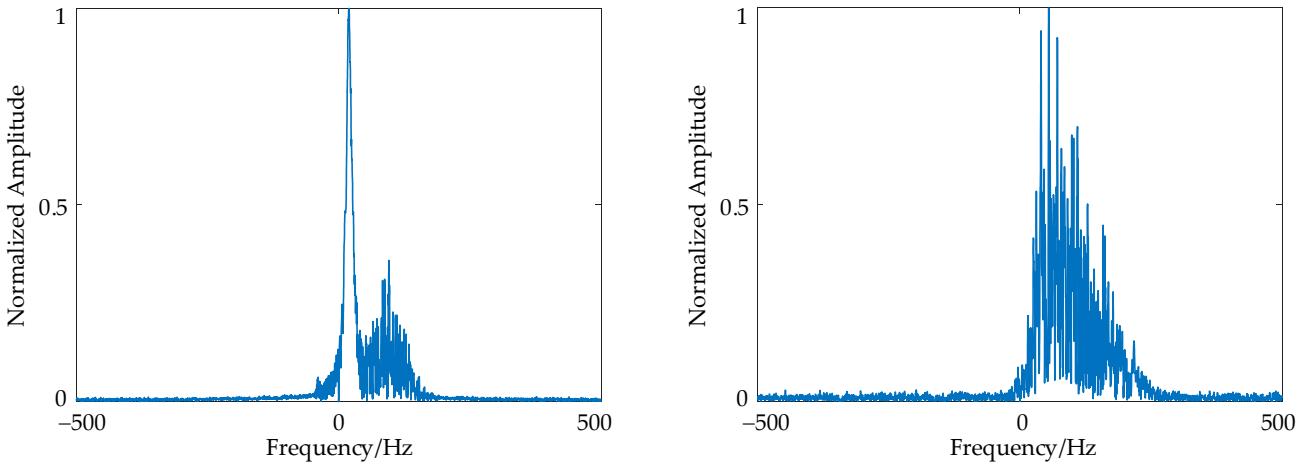


Figure 3. Comparison of Doppler Spectrum with Target Echo and Sea Clutter: (**a**) Target echo; (**b**) Sea clutter.

From Figure 3, it can be seen that the center frequency of the Doppler spectrum of both the target echo and the sea clutter is positive, which indicates that the ocean waves are moving closer to the radar. Compared with the target echoes, the sea clutter has a wider range of energy distribution in the frequency domain, this is because the ocean waves have more variable motion states and more complex scattering structures. In addition, although the motion state of the target will change with the change of sea state, the spatial geometry of the target is fixed, and the reflection of energy is more concentrated, so the Doppler spectrum of the target is more aggregated.

In fact, the differences between target echoes and sea clutter in the Doppler spectrum are not always as obvious as shown in Figure 3, and target echoes and sea clutter often overlap highly in the time and frequency domains. Some statistical features proposed at this stage cannot fully describe the difference information contained in the Doppler spectrum of target echo and sea clutter, so the detection accuracy of the target in sea clutter by the one-dimensional method is not high. In order to exert the superior performance of deep learning in the field of image classification and target recognition, this study encodes the Doppler spectrum sequence of radar echo data into a two-dimensional image, and applies a CNN model to classify the transformed image.

2.4. Time Series Coding Methods

Inspired by the great success of deep learning in computer vision, Wang. et al. [8] proposed two processing algorithms for encoding time series into images: GASF and GADF. These algorithms can convert a one-dimensional time series into a two-dimensional image, enabling the visualization of time series data. Meanwhile, the encoded image also retains the time dependence and correlation of the original data.

The GAF algorithm represents time series in polar coordinates rather than Cartesian coordinates, which evolved from the Gram matrix. The specific implementation of the steps of the GAF algorithm is as follows: Assuming that a time series $X = \{x_1, x_2, \ldots, x_n\}$ contains n real-valued observations, we can rescale all the values of X to the interval $[-1, 1]$ by the normalization method shown in Equation (2):

$$\tilde{x}_{-1}^{i} = \frac{(x_i - \max(X) + (x_i - \min(X)))}{\max(X) - \min(X)} \tag{2}$$

Similarly, we can readjust all values of X to the interval [0,1] using the normalization shown in Equation (3):

$$\tilde{x}_{0}^{i} = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{3}$$

We denote the rescaled time series as $\widetilde{X}$ and then map $\widetilde{X}$ to polar coordinates using the following equation:

$$\begin{cases} \phi_i = \arccos(\widetilde{x}_i), -1 \leq \widetilde{x}_i \leq 1, \widetilde{x}_i \in \widetilde{X} \\ r_i = \frac{i}{N}, i \in \mathbb{N} \end{cases} \tag{4}$$

where ϕ_i represents the arc cosine of $\widetilde{x}_i$ $(i = 1, 2, \ldots, n)$, corresponding to the angle in the polar coordinate system, r_i represents the polar coordinate radius corresponding to the timestamp i, and N plays the role of adjusting the span of the polar coordinate system.

The scaled data of the two normalization operations correspond to different angle ranges respectively when converted to the polar coordinate system. The data within the range of $[-1, 1]$ corresponds to the arc cosine function angle range of $[0, \pi]$, and the arc cosine value range corresponding to the data in the range of $[0, 1]$ is $[0, \pi/2]$. This polar coordinate system-based representation provides us with a new perspective for understanding time series, that is, the timescale changes of the series are mapped to the radius of the polar coordinate system over time, while the amplitude changes are mapped to the angle of the polar coordinate system. By calculating the sum and difference of trigonometric functions between sampling points, GASF and GADF are respectively defined as follows:

$$\text{GASF} = \begin{bmatrix} \cos(\phi_1 + \phi_1) & \cdots & \cos(\phi_1 + \phi_n) \\ \cos(\phi_2 + \phi_1) & \cdots & \cos(\phi_2 + \phi_n) \\ \vdots & \ddots & \vdots \\ \cos(\phi_n + \phi_1) & \cdots & \cos(\phi_n + \phi_n) \end{bmatrix} \tag{5}$$

$$\text{GADF} = \begin{bmatrix} \sin(\phi_1 - \phi_1) & \cdots & \sin(\phi_1 - \phi_n) \\ \sin(\phi_2 - \phi_1) & \cdots & \sin(\phi_2 - \phi_n) \\ \vdots & \ddots & \vdots \\ \sin(\phi_n - \phi_1) & \cdots & \sin(\phi_n - \phi_n) \end{bmatrix} \tag{6}$$

In summary, the GAF algorithm is used to convert the one-dimensional time series into a two-dimensional image through three steps: scaling, coordinate system conversion, and trigonometric function operations. Figure 4 shows the complete process of converting the rescaled Doppler spectrum sequence into a GAF-encoded map. First, the Doppler spectral sequence is normalized using Equation (2); then the normalized Doppler spectral sequence is converted to a polar coordinate system for representation using Equation (4); finally, the data matrices of the GASF and GADF images are obtained using Equations (5) and (6), respectively.

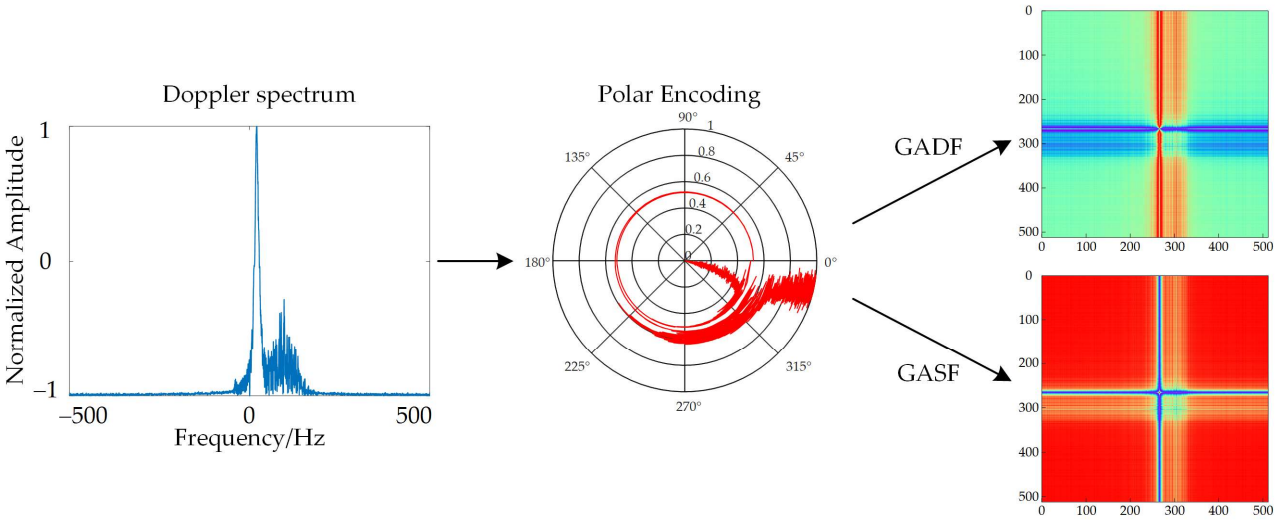


Figure 4. Schematic diagram of the conversion of Doppler spectral sequences into GAF maps.

2.5. CA-EfficientNet Network Model

To obtain better accuracy for network models, optimization of network depth, network width, and image resolution is usually adopted, such as ResNet [14], DenseNet [15], etc. However, these network models often change only 1 of the 3 dimensions of network depth, network width, and image resolution, and require tedious manual adjustment of parameters, and still yield sub-optimal accuracy and efficiency. In 2017, Google proposed depthwise separable convolution (DSC) as the basis for building the lightweight network MobileNet [16], which also allows users to modify the two parameters of network width and input resolution to adapt to different application environments; in 2019, Google Brain Team proposed the EfficientNet network model, which takes into account the depth, width, and resolution of the input image without increasing the number of parameters and operations, and scales these three parameters reasonably and efficiently to achieve optimal accuracy.

2.5.1. EfficientNet Model Selection

Compared with other network models, the EfficientNet family of networks can maintain high classification accuracy with a small number of model parameters at the same time. The scale of the network model is mainly determined by the scaling parameters in three dimensions: width, depth, and resolution. The higher the resolution of the input image, the deeper the network is needed to obtain a larger perceptual field of view, and similarly, the more channels are needed to obtain more accurate features. A total of eight models from B0 to B7 are constructed for the EfficientNet network according to the image resolution. The size of the B0~B7 model increases sequentially with higher accuracy and greater memory requirements. Among them, the accuracy of the B7 model is 84.4% and 97.1% for Top-1 and Top-5, respectively, on the natural image ImageNet dataset, which has reached the best accuracy at that time. The B0~B7 models have the least number of operations among the networks that achieve the same accuracy.

For the characteristics of GAF images, we expect the selected base network model to have strong feature extraction capability and high recognition accuracy. We also need to consider the hardware implementability of the network, which requires that the number of parameters of the network model should not be too large. Higher image resolution can show more texture details, which is beneficial for the improvement of recognition accuracy. As the image resolution increases, the time and memory required to train the network model also increases exponentially. In order to select a suitable structure from the EfficientNet family of networks as our base network, we conducted comparison experiments on different network structures on the generated GADF and GASF datasets, respectively. We trained the network with the number of iterations set to 100 and the initial learning rate of 0.001, and the experimental results on the validation set are shown in Table 2.

Table 2. EfficientNetB0-B7 network comparison experiment results.

Network Model	Resolution (Pixel)	Parameters/M	Accuracy (%)		Loss		Model Training Time (h)	
			GADF	GASF	GADF	GASF	GADF	GASF
EfficienNet-B0	224 × 224	5.29	93.83	93.56	0.1734	0.1816	1.97	1.95
EfficienNet-B1	240 × 240	7.79	93.46	93.27	0.1831	0.1914	3.02	3.07
EfficienNet-B2	260 × 260	9.11	93.73	93.44	0.1774	0.1873	4.43	4.46
EfficienNet-B3	300 × 300	12.23	93.69	93.38	0.1776	0.1877	6.37	6.42
EfficienNet-B4	380 × 380	19.34	93.77	93.49	0.1770	0.1846	13.55	13.62
EfficienNet-B5	456 × 456	30.39	93.71	93.52	0.1763	0.1836	27.52	27.59
EfficienNet-B6	528 × 528	43.04	93.79	93.56	0.1789	0.1819	50.37	50.45
EfficienNet-B7	600 × 600	66.35	93.86	93.62	0.1757	0.1793	88.80	88.92

From Table 2, we can see that the computational effort increases exponentially as the complexity of the network model increases, which is manifested by a geometric increase in training time. The classification results on the GADF and GASF validation datasets do not improve with the increase of the complexity of the network. The accuracy of the B7

network model based on GADF validation datasets is higher than that of the B0 network model by 0.03%, and the accuracy of the B7 network model based on GASF validation datasets is higher than that of the B0 network model by 0.06. The accuracy of all other models is lower than that of the B0 network model. Therefore, we chose EfficientNet-B0, which has the least model parameters and the fastest training speed, as the base network for the next step.

The network structure of EfficientNet-B0 is shown in Table 3, which consists of 16 MBConv blocks, 2 convolutional layers, 1 global average pooling (GAP) layer, and 1 fully connected layer. The MBConv block contains the Depthwise (DW) Convolution, Squeeze-and-Excitation (SE) attention mechanism module, Swish activation function, and Dropout layer. The SE module compresses the input 3D feature matrix into a one-dimensional channel feature vector by a global average pooling operation, which reflects the importance of different channels of the input feature matrix.

Table 3. EfficientNet-B0 network structure.

Stage	Operator	Resolution	Channels	Layers
1	Conv 3 × 3	224 × 224	32	1
2	MBConv1, k 3 × 3	112 × 112	16	1
3	MBConv6, k 3 × 3	112 × 112	24	2
4	MBConv6, k 5 × 5	56 × 56	40	2
5	MBConv6, k 3 × 3	28 × 28	80	3
6	MBConv6, k 5 × 5	14 × 14	112	3
7	MBConv6, k 5 × 5	14 × 14	192	4
8	MBConv6, k 3 × 3	7 × 7	320	1
9	Conv 1 × 1 and Pooling and FC	7 × 7	1280	1

Note: Conv stands for convolution; MBConv stands for lightweight flip bottleneck convolution kernel; k stands for convolution kernel size.

2.5.2. Coordinate Attention

We conducted an observational study on a large sample of both coded maps and found that the texture features of the target and sea clutter are not randomly distributed. These texture features will only appear in specific regions, while most of the remaining regions will have duplicate texture information and no texture information. The focus on these regions with duplicated textures and regions without texture may be ineffective, so we need to emphasize the location information in the coded maps that is critical for target identification.

The channel attention mechanism has a significant effect on improving the model performance, but it only focuses on the long-term dependencies between channels and ignores the importance of location information. A new network attention mechanism is proposed in the literature [11], which embeds location information into channel attention, called the coordinate attention mechanism. Unlike channel attention that transforms feature tensor into a single feature vector, the coordinate attention mechanism decomposes channel attention into two one-dimensional feature encoding processes that aggregate features along two spatial directions, respectively. In this way, long-term dependencies are captured so that channel dependencies can be obtained while retaining precise location information. The generated feature maps are then encoded into direction-aware and position-aware attentional feature maps, respectively. This pair of attentional maps can be complementarily applied to the input feature maps to increase the representation of the object of interest. The coordinate attention mechanism is relatively simple and computationally small and is easy to insert into the network structure. A structure diagram of the CA module is shown in Figure 5.

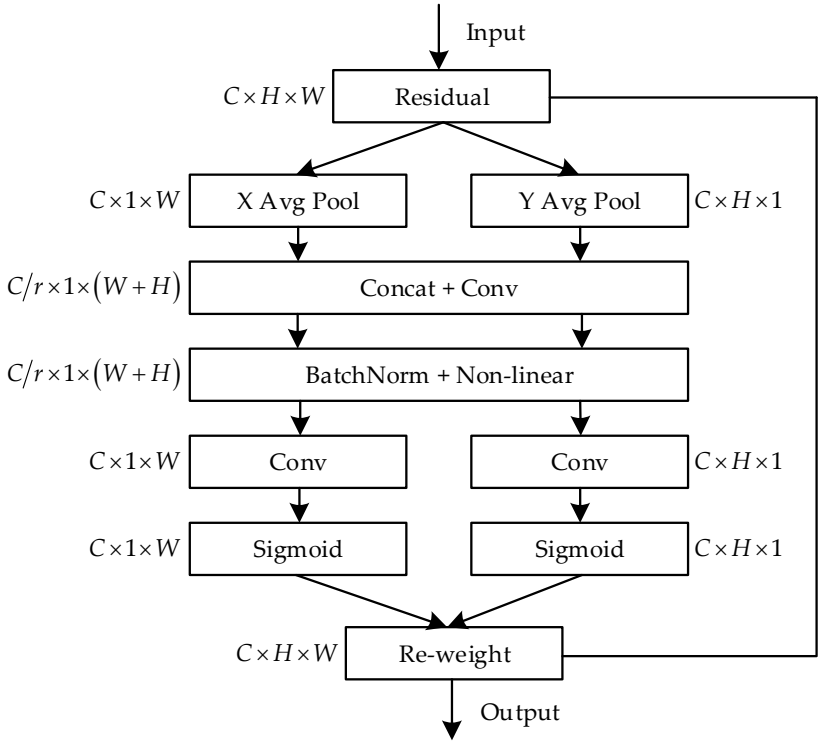


Figure 5. Coordinate Attention Block. Note: Pool stands for pooling; X/Y Avg Pool stands for average pooling in X/Y direction; FC stands for fully connected layer; Concat stands for stitching; BN stands for batch normalization; Non-linear, Sigmoid and Swish stand for nonlinear activation functions; *C* stands for number of channels; *H* stands for feature map height; *W* stands for feature map width; r is the channel scaling factor; each directly connected branch indicates the fusion of the input feature map with the feature map obtained through convolution and other operations.

2.5.3. Improvement of EfficientNet

To further improve the recognition accuracy of the sea clutter target detection model, this study introduces CA to further improve EfficientNet to enhance the learning of location information that plays an important role in the coded graph. Compared with the original EfficientNet network model, the following improvements to the network model are mainly made in this study.

1. Connecting CA modules in the form of residual structures after the first convolutional layer of the network. This is because the CA module changes the weight parameters by combining the features of the convolutional layers, applying larger weights to the important feature channels and smaller weights to other less important feature channels. The CA module enhances the global attention of the CNN so that the network does not lose more critical information due to the pre-convolutional operation, thus improving the model's ability to distinguish between targets and sea clutter.
2. The SE module within each MBConv module in the original EfficientNet network is replaced with a CA module. Specifically, for the MBConv module structure shown in Figure 6, the original SE module after the DW convolution module in it is replaced with a CA module. By this operation, the CA module can be used to capture the long-term dependency between network channels, so that the network can pay attention to the target-relevant region without losing the accurate position information.

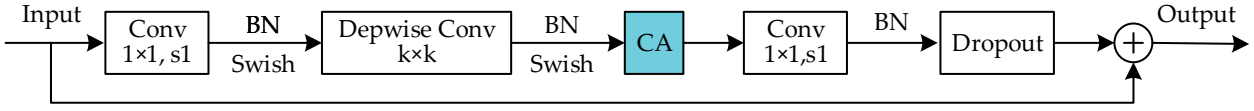


Figure 6. The improved MBConv schematic diagram. Note: ⊕ stands for channel-level addition.

The structure of the improved CA-EfficientNet-B0 network is shown in Figure 7: first, the input image is converted into a 224 × 224 × 32 matrix; then, the feature map after the first layer of the convolution operation is multiplied with the attention feature map enhanced by the CA module at the channel level to obtain the feature map with attention information; then, the higher-level features of the image are further extracted by seven MBConv modules embedded with CA in turn to obtain the 7 × 7 × 1280 feature map; finally, the result of the image recognition is obtained by the fully connected layer.

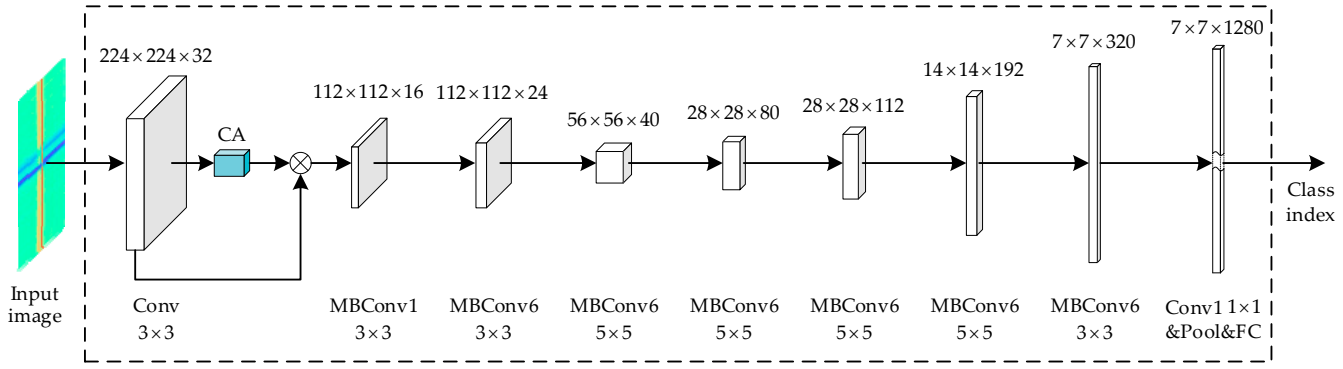
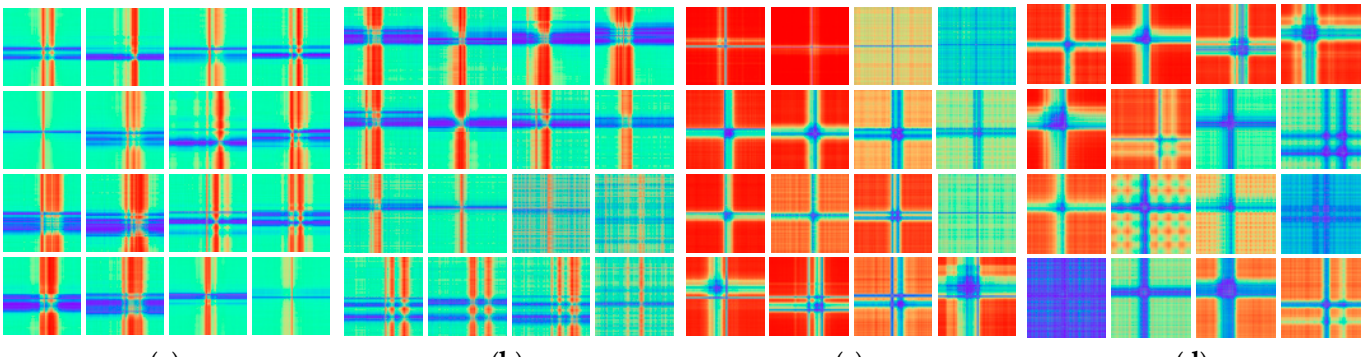


Figure 7. Diagram of network structure for CA-EfficientNet-B0. Note: ⊗ stands for channel-level multiplication.

3. Results

3.1. Construction of Training Dataset

The data used for model training and testing were converted from the measured sea clutter data. The 14 sets of measured data listed in Table 1 were converted into the corresponding two types of datasets according to the GASF and GADF coding methods mentioned in Section 2.4, respectively. First, the data of each distance cell are partitioned into subsequences of a length of 1024, and the interval overlap of adjacent subsequences is 50%. Then, the Doppler spectra of corresponding subsequences are calculated using 1024-point FFT, and then the normalized Doppler spectral sequences are encoded into GASF maps and GADF maps, respectively. Finally, they are used to construct the two datasets, respectively. We selected 85,680 images from the existing coding library for our test and validation sets, including 42,840 images for GADF and GASF, with a 4:1 ratio of training to validation and a 2:1 ratio of clutter to target. A total of 57,120 images were selected for the test set, including 28,560 images for GADF and GASF, with a 1:1 ratio of clutter to target. Figure 8 shows the sample of GADF and GASF coded images corresponding to the target and clutter in the dataset.



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

Figure 8. GADF and GASF encoded image samples: (**a**) GADF of sea clutter; (**b**) GADF of target; (**c**) GASF of sea clutter; (**d**) GASF of target.

3.2. Experimental Environment and Parameter Settings

All training and testing experiments were performed on the Linux operating system, and the open source Pytorch deep learning framework was used to build the network model. The CPU was AMD EPYC 7543 32-Core Processor with 30 GB of RAM, and the GPU was RTX A5000 with 24 G of video memory. The programming language was Python3.8, and the integrated development platform was PyCharm2021.1.3. In each batch, 64 images were trained, the iteration period was 100 times, the initial learning rate was set to 0.001, the AdamW optimizer was chosen to optimize the model parameters, the learning rate was adjusted by the cosine annealing algorithm strategy, the learning rate was adjusted from the initial learning rate to 0 within 100 epochs, and the loss function was the SoftMax cross-entropy loss function.

3.3. Test Evaluation Indicators

In order to evaluate the classification performance of the CA-EfficientNet-B0 network model proposed in this paper, the recognition accuracy (Accuracy), F1 value, number of parameters, and Floating-Point Operations (FLOPs) are selected as evaluation metrics in this study. The number of parameters is the total number of parameters that can be trained in the network model, measured in millions (M). The FLOPs are the number of floating-point operations, and are used to measure the complexity of the algorithm or model, and are measured in billions (B). The recognition accuracy is the probability value of the number of correctly predicted samples to the total number of tested samples. For image classification network models, recognition accuracy is the most important evaluation index, and its calculation formula is as follows:

$$\text{Accuracy} = \frac{P_C}{P_{ALL}} \times 100\% \tag{7}$$

where P_C denotes the number of correctly predicted samples and P_{ALL} denotes the total number of samples in the test set. To comprehensively evaluate the performance of the CA-EfficientNet model, this study also used the value that considers both the accuracy and recall of the classification model, which is calculated as follows:

$$F1 = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{8}$$

where Precision is the accuracy rate, which indicates how many of the totally predicted positive samples in the test set were correctly predicted; Recall is the recall rate, also called the full rate, which indicates how many of the positive samples in the original total positive sample sets were correctly predicted. The precision and recall rates are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative samples, respectively.

3.4. Training and Testing

For the CA-Efficient network model proposed in Section 2.5, the model was trained using the experimental setup given in Section 3.2. The training loss and validation loss curves are shown in Figure 9. From the figure, it can be seen that the training and validation loss values can drop steadily and quickly with the increase in the value of the epoch. It indicates the effectiveness and learnability of the model. At iteration number 70, the validation loss value was close to convergence and the classification accuracy tended to be stable. The training loss value basically converged at the 90th iteration, indicating that the

model reached saturation and the recognition accuracy no longer increases. This indicates that the experimental setup of this study is reasonable and feasible.

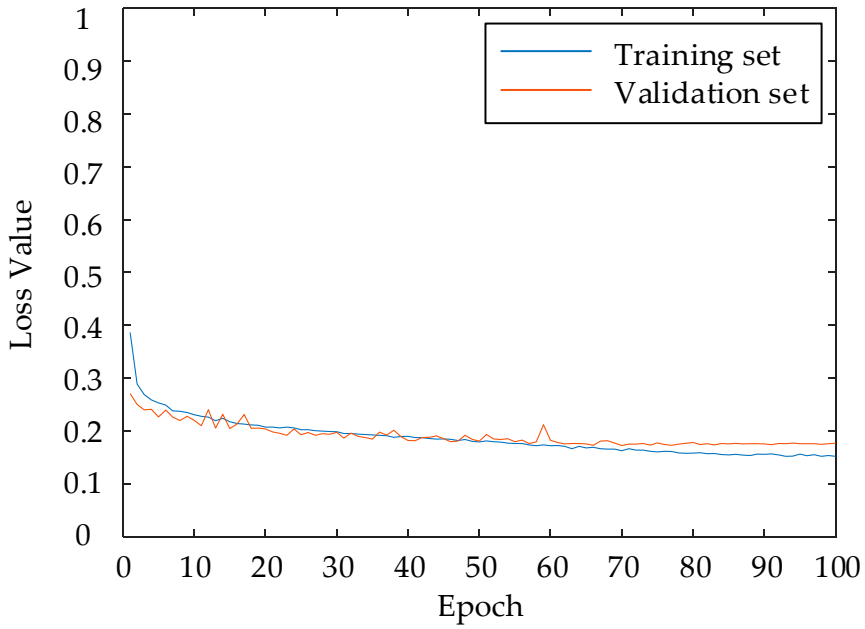


Figure 9. Curves for training and validation loss.

3.5. CA-EfficientNet Model Ablation Trials

In order to verify the effectiveness of the CA-EfficientNet-B0 model proposed in this study, four ablation test schemes were used in this study as follows.

Scheme 1: Using EfficientNet-B0 only, with the optimizer choosing stochastic gradient descent (SGD) + momentum and the learning rate adjustment strategy choosing cosine annealing algorithm.

Scheme 2: Replacing the optimizer with the AdamW optimization algorithm on the basis of Scheme 1.

Scheme 3: Replacing the SE module of the MBConv module in the EfficientNet-B0 network with the CA module on the basis of Scheme 2.

Scheme 4: Based on Scheme 3, the CA module is introduced after the first convolutional layer to form the lightweight network model CA-EfficientNet-B0 proposed in this study.

The experimental results of the above ablation test schemes are shown in Table 4. From the experimental results of Scheme 1 and Scheme 2, it is clear that the AdamW optimizer can robustly improve the convergence of the model. Comparing the experimental results of Scheme 2 and Scheme 3, it is clear that the CA module has stronger attention learning ability than the SE module. Comparing the experimental results of Scheme 3 and Scheme 4, it is clear that adding the CA module after the first module is effective. From the experimental results of Scheme 4, it can be seen that the improved CA-EfficientNet-B0 model proposed in this study achieves 96.13% and 96.28% recognition accuracy on the GADF and GASF test datasets, respectively. Compared with the baseline model, the accuracy of the improved network model proposed in this study was improved by 1.74% and 2.06% on the GADF and GASF datasets, respectively. The combined experimental results of Schemes 1, 2, 3, and 4 show that the model improvement scheme and the network training strategy in this study are feasible and effective. In summary, the CA-EfficientNet-B0 model is a lightweight and deep network model with good recognition accuracy.

Table 4. Ablation study results for CA-EfficientNet model.

No.	Improvement Strategy	Parameters/M	Accuracy/%		F1 Value/%	
			GADF	GASF	GADF	GASF
1	EfficientNet-B0	5.29	94.39	94.22	91.04	90.79
2	EfficientNet-B0 + AdamW	5.29	94.75	94.69	91.68	91.55
3	EfficientNet-B0 + AdamW (SE→CA)	5.38	95.25	95.23	92.46	92.43
4	CA-EfficientNet-B0 + AdamW	5.38	96.13	96.28	93.92	94.18

4. Discussion

4.1. Performance Comparison with Similar Excellent Deep Network Models

In order to fully evaluate the recognition performance of the CA-EfficientNet-B0 network proposed in this study on GADF and GASF datasets, some classical CNN models were selected for performance comparison in this study, including AlexNet [17], VGG [18], GoogleNet [19], ResNet [14], DenseNet [15], SqueezeNet [20], MobileNet [16,21,22], ConvNeXt [23], and EfficientNet [10]. For a fair comparison, the experiments were conducted by, firstly, all using the open-source deep learning framework Pytorch; and secondly, keeping the basic architecture of these deep learning models unchanged. The recognition accuracies of the selected comparative network models and the CA-EfficientNet-B0 network model proposed in this study on the GADF and GASF datasets are given in Table 5.

Table 5. Performance comparison of each network model on GADF and GASF datasets.

Model Type	Accuracy/%		F1 Value/%		Parameters/M	FLOPs/B
	GADF	GASF	GADF	GASF		
AlexNet [17]	93.78	93.49	90.10	89.73	61.10	0.71
VGG-11 [18]	93.86	93.98	90.15	90.48	132.86	7.62
VGG-16 [18]	93.97	94.01	90.22	90.51	138.36	15.48
GoogleNet [19]	94.13	94.03	90.67	90.47	7.00	1.59
ResNet-34 [14]	94.67	94.14	91.56	90.64	21.80	3.67
ResNet-50 [14]	94.04	93.67	90.43	89.89	25.56	4.11
DenseNet-121 [15]	94.46	94.16	91.17	90.71	7.98	2.87
SqueezeNet [20]	92.75	92.97	88.40	88.75	87.51	15.36
MobileNetV1 [16]	94.00	93.63	90.44	89.79	3.21	0.58
MobileNetV2 [21]	94.35	94.00	90.98	90.37	3.50	0.31
MobileNetV3 [22]	93.88	93.72	90.18	89.93	2.54	0.06
ConvNeXt-tiny [23]	91.92	92.16	86.99	87.46	28.57	4.46
EfficientNet-B0	94.39	94.22	91.04	90.79	5.29	0.40
CA-EfficientNet-B0	96.13	96.28	93.92	94.18	5.38	0.41

As can be seen from Table 5, the benchmark network EfficientNet-B0 selected in this study achieved the best recognition accuracy compared to other classical classification networks (networks other than ResNet-34 and DenseNet on the GADF dataset). ResNet-34 and DenseNet outperformed the EfficientNet-B0 model on the GADF dataset. The CA-EfficientNet-B0 network proposed in this study can be improved by 1.46 and 1.67 percent compared to ResNet-34 and DenseNet, respectively.

The model parameters of CA-EfficientNet-B0 are only 5.38 M, and the FLOPs are 0.41 B. As shown in Table 5, except for MobileNetV1, MobileNetV2, and MobileNetV3, the model parameters and the FLOPs of CA-EfficientNet-B0 are much smaller than those of other classical network models. Compared with the classical classification networks VGG-16, ResNet-50, and GoogleNet, the number of CA-EfficientNet-B0 model parameters is only 3.89%, 21.05%, and 76.86% of the number of parameters of these network models. In summary, the CA-EfficientNet-B0 model proposed in this study has the features of high accuracy, low number of parameters, easy migration, and easy deployment. Based on this, the subsequent application to mobile can be attempted to meet the requirements of accurate detection of floating small targets in the background of sea clutter in complex situations and promote the development of intelligent detection methods.

4.2. Visual Analysis of Class Activation Map

In order to visualize the effectiveness of the CA-EfficientNet-B0 model proposed in this experiment, a class activation map visualization comparison analysis was performed on some data by the EfficientNet-B0 model and the improved CA-EfficientNet-B0 model based on the Grad-CAM [24] technique, and here we only show the analysis results for the GADF dataset, as shown in Figure 10.

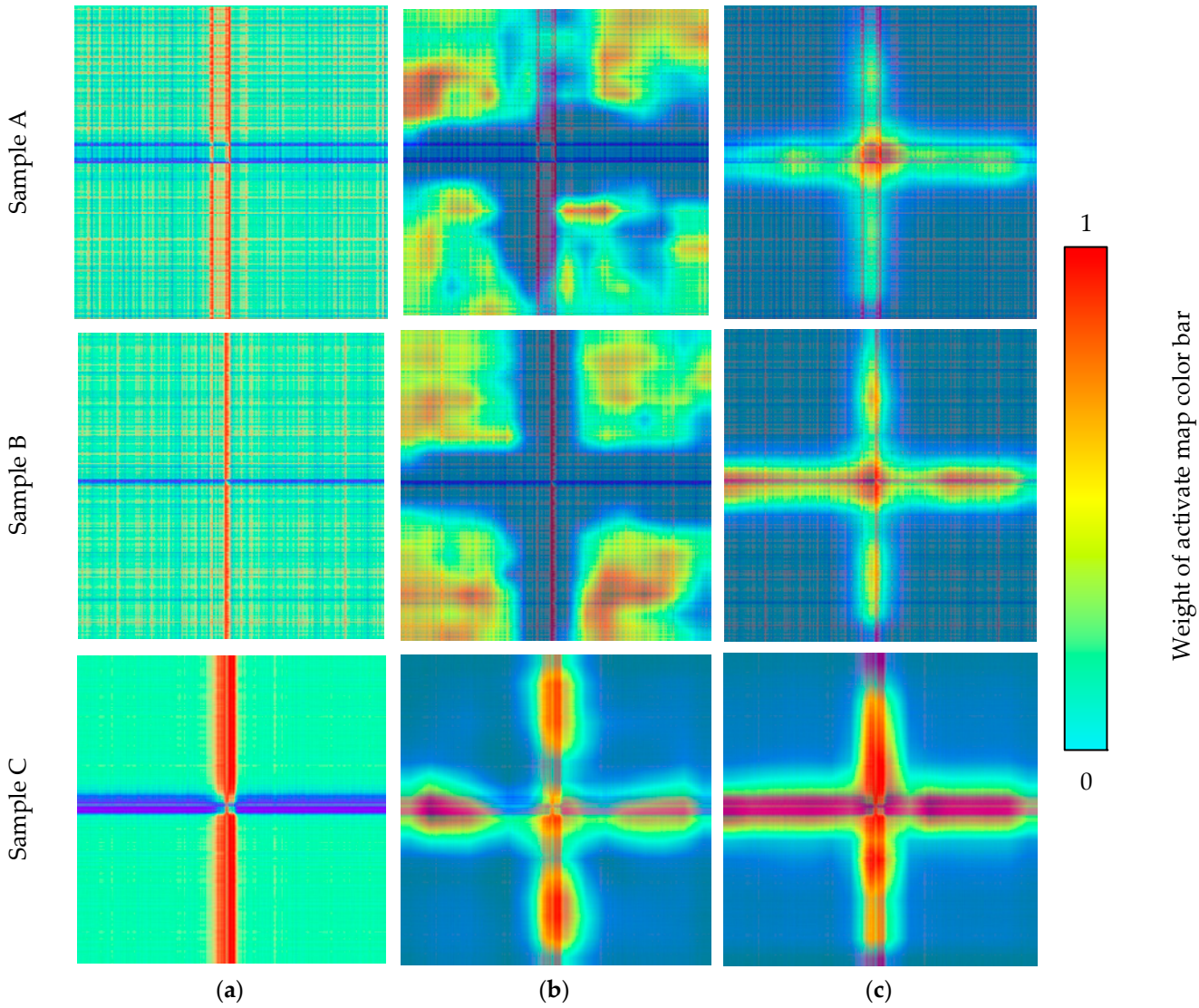


Figure 10. Class Activation Maps for GADF Encoded Graphs for Different Models: (**a**) Original image; (**b**) EfficientNet-B0; (**c**) CA-EfficientNet-B0.

From Figure 10, it can be seen that for the same original coded map, when there is complex background interference information in the coded map, the benchmark network EfficientNet-B0 selected in this study cannot focus on the regions related with the texture information corresponding to the target and clutter, such as Sample A and Sample B. In addition, EfficientNet-B0 is not accurate enough for the target and sea clutter in the coded map, and the focused areas are usually scattered. EfficientNet-B0 focuses less on the intermediate areas with rich texture information. The improved CA-EfficientNet-B0 model can precisely locate the position of the effective texture information in the image. It pays less attention to the background area, so it receives little interference from the complex background. This shows that the coordinate attention module can effectively extract the key regions of the features and suppress the interference of the background regions. This

also indicates that the proposed method in this study has stronger attention learning ability and can effectively improve the recognition accuracy.

4.3. The Impact of SCR on Target Classification

In addition, we tested the data under different polarization methods for different datasets, respectively, by the proposed CA-EfficientNet-B0 model, and the test results are shown in Figure 11. The recognition correct rate of each data shows a weak correlation with the ASCR of the corresponding data in Section 2.2. It indicates that the detection performance of the target detection method proposed in this study does not depend on the ASCR in the dataset. It has a good detection performance even at low SCR, such as the #19VV data and #310VV data with ASCR below 0 dB, and their corresponding correct recognition rates for both types of coded maps exceed 97.5%. In addition, comparing the results in Figure 11a,b, it can be seen that the two coding methods show some differences in the same numbered data, and the average detection accuracy of the dataset based on the GASF coding method is slightly higher than that of the dataset based on the GADF coding method.

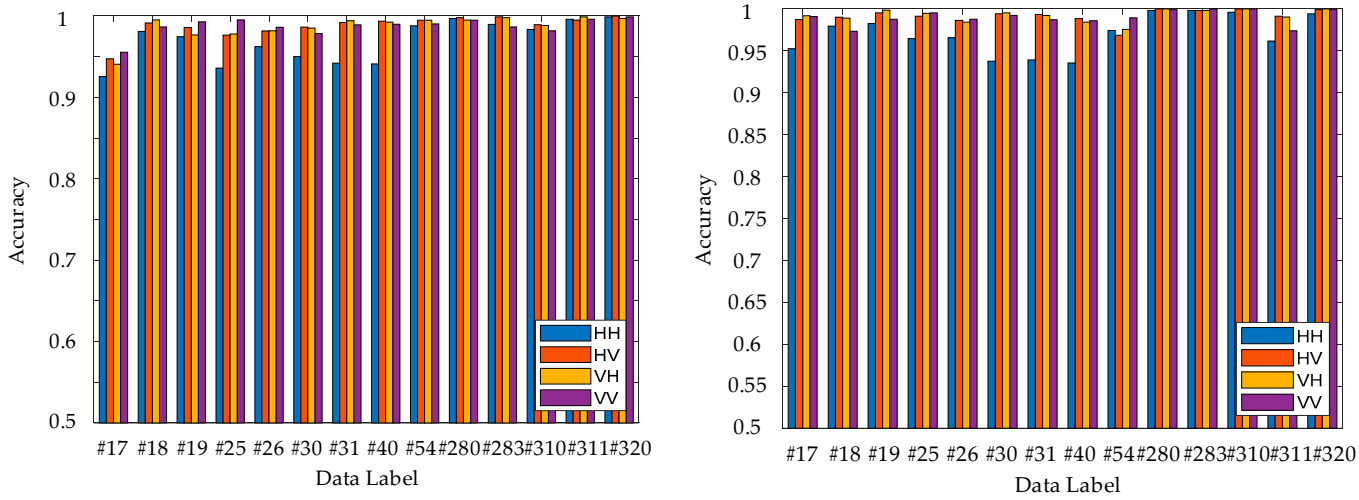


Figure 11. Detection results by CA-EfficientNet-B0 model: (**a**) GADF coding; (**b**) GASF coding.

4.4. Comparison with Other Detectors

In this section, we compare GAF-based detectors with tri-feature-based detectors [13] and iForest-based detectors [5], choosing ten datasets listed in Table 1 to evaluate the performance of the detectors. The observation time was set to 1.024 s (L = 1024), and the false alarm probability (FAR) was 0.001. As shown in Figure 12, under the four polarization modes, the proposed detector achieves higher detection probability than the other two existing detectors.

The amount of computational cost of the proposed algorithm in this paper mainly includes three parts. The first part is the computational cost of the preprocessing of time series and fast Fourier transform. The second part is the computational cost of the process of transforming the sequence of the Doppler spectrum to GAF map. The last part is the computational cost needed to train model. The model training time mainly depends on the size of the training set, the epoch of training, and the computational performance of the equipment used for training. Details of the model training time are listed in Table 2. The main computational cost is consumed by the preliminary preparatory work, because we need a considerable amount of training data to train our model, and the computational cost in this phase may be significant. Compared with algorithms based on statistical features, our proposed algorithm has no advantage in terms of computational cost. However, if our model has been trained, when we use it to detect new data, the detection process becomes

extremely fast. The significant advantage of the proposed method is that we only need to train new datasets on the basis of the trained network model, so that our model can have stronger generalization ability. In Section 4.1, we compare the proposed algorithm with similar deep network models in terms of model parameters and the FLOPs, and it can be seen that our model has great advantages.

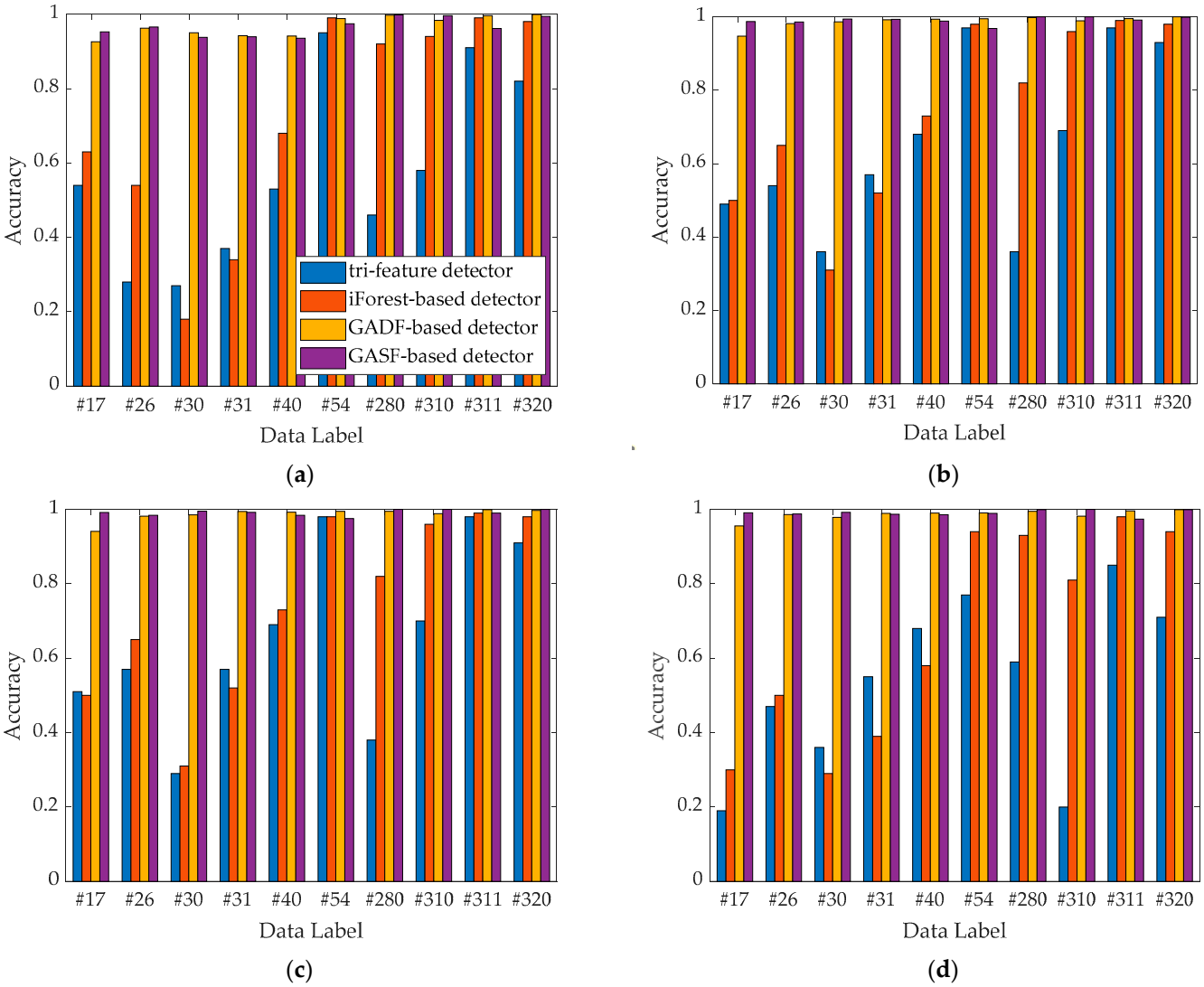


Figure 12. Detection probability of the tri-feature detector, the iForest-based detector, proposed method at (**a**) HH, (**b**) HV, (**c**) VH, and (**d**) VV polarization.

5. Conclusions

In this paper, we convert the target recognition problem into a classification problem for Doppler spectral encoded maps. We use the improved CA-EfficientNet-B0 model to train and test the experimentally generated dataset for the detection of targets on the sea surface. The detection and classification results show that the CA-EfficientNet-B0 network model proposed in this paper has the advantage of high accuracy and intelligence in the detection of small floating targets on the sea surface, and our method is effective in the classification of both targets and clutter. In addition, by comparing with other classical classification algorithms, it is shown that the CA-EfficientNet-B0-based method proposed in this paper has more advantages in detection and recognition probability, and also provides a new idea for radar sea moving target detection and classification. Both the proposed GADF and GASF coded graph detection methods have good recognition performance, and for the same Doppler spectral sequence, the two coding methods exhibit different

characteristics. It leads to the datasets constructed based on the two coding methods exhibiting different detection performance in the same network model, and the average detection accuracy of the dataset based on the GASF coding method is higher than that of the GADF coding method based on the datasets.

Since there are some differences in the description and details of the same Doppler spectral sequence between GASF and GADF coding, how to fuse the two different coding methods and apply them to the deep learning model is a problem that needs to be solved in the future. In addition, finding a more suitable time series coding algorithm for this research and further improving the detection accuracy of the network model are also the future exploration directions of this research.

Author Contributions: Conceptualization, C.X.; methodology, R.L.; software, R.L.; validation, C.X. and R.L.; formal analysis, R.L.; investigation, R.L.; resources, R.L.; data curation, R.L.; writing—original draft preparation, R.L.; writing—review and editing, C.X.; visualization, R.L.; supervision, C.X.; project administration, C.X.; funding acquisition, C.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was sponsored by the National Natural Science Foundation of China under Grant number 61901195.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the data also forming part of ongoing study.

Acknowledgments: The authors are grateful to the anonymous reviewers for helpful comments and are grateful to Simon Haykin of the McMaster University of Canada for sharing the IPIX radar datasets with common users.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Farina, A.; Gini, F.; Greco, M.V.; Verrazzani, L. High Resolution Sea Clutter Data: Statistical Analysis of Recorded Live Data. *IEE Proc.-Radar Sonar Navig.* **1997**, *144*, 121–130. [CrossRef]
2. Xu, S.; Bai, X.; Guo, Z.; Shui, P. Status and Prospects of Feature-Based Detection Methods for Floating Targets on the Sea Surface. *J. Radars* **2020**, *9*, 684–714. [CrossRef]
3. Chen, S.; Luo, F.; Hu, C.; Nie, X. Small Target Detection in Sea Clutter Background Based on Tsallis Entropy of Doppler Spectrum. *J. Radars* **2019**, *8*, 344–354. [CrossRef]
4. Zhou, Y.; Cui, Y.; Xu, X.; Suo, J.; Liu, X. Small-Floating Target Detection in Sea Clutter via Visual Feature Classifying in the Time-Doppler Spectra. *arXiv* **2020**. [CrossRef]
5. Xu, S.; Zhu, J.; Jiang, J.; Shui, P. Sea-Surface Floating Small Target Detection by Multifeature Detector Based on Isolation Forest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 704–715. [CrossRef]
6. Gu, T. Detection of Small Floating Targets on the Sea Surface Based on Multi-Features and Principal Component Analysis. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 809–813. [CrossRef]
7. Su, N.; Chen, X.; Guan, J.; Mou, X.; Liu, N. Detection and Classification of Maritime Target with Micro-motion Based on CNNs. *J. Radars* **2018**, *7*, 565–574. [CrossRef]
8. Wang, Z.; Oates, T. Imaging Time-Series to Improve Classification and Imputation. In Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; AAAI Press: Palo Alto, CA, USA; pp. 3939–3945. [CrossRef]
9. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717. [CrossRef]
10. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114. [CrossRef]
11. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**. [CrossRef]
12. McMaster IPIX Radar. Available online: http://soma.ece.mcmaster.ca/ipix/dartmouth/datasets.html (accessed on 7 July 2022).
13. Shui, P.-L.; Li, D.-C.; Xu, S.-W. Tri-Feature-Based Detection of Floating Small Targets in Sea Clutter. *IEEE Trans. Aerosp. Electron. Syst.* **2014**, *50*, 1416–1430. [CrossRef]

14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
15. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [CrossRef]
16. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**. [CrossRef]
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
18. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**. [CrossRef]
19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
20. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
21. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv* **2019**. [CrossRef]
22. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. *arXiv* **2019**. [CrossRef]
23. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. *arXiv* **2022**. [CrossRef]
24. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]