remote sensing

MDPI

Article

S4Det: Breadth and Accurate Sine Single-Stage Ship Detection for Remote Sense SAR Imagery

Mingjin Zhang [1], Yingfeng Zhu [1], Longyi Li [1], Jie Guo [1,*], Zhengkun Liu [2] and Yunsong Li [1]

[1] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; mjinzhang@xidian.edu.cn (M.Z.); 21011210406@stu.xidian.edu.cn (Y.Z.); 23011210785@stu.xidian.edu.cn (L.L.); ysli@mail.xidian.edu.cn (Y.L.)
[2] China Academy of Aerospace Science and Innovation, Beijing 100088, China; 13910394841@163.com
* Correspondence: jguo@mail.xidian.edu.cn

Abstract: Synthetic Aperture Radar (SAR) is a remote sensing technology that can realize all-weather and all-day monitoring, and it is widely used in ocean ship monitoring tasks. Recently, many oriented detectors were used for ship detection in SAR images. However, these methods often found it difficult to balance the detection accuracy and speed, and the noise around the target in the inshore scene of SAR images led to a poor detection network performance. In addition, the rotation representation still has the problem of boundary discontinuity. To address these issues, we propose S4Det, a Sinusoidal Single-Stage SAR image detection method that enables real-time oriented ship target detection. Two key mechanisms were designed to address inshore scene processing and angle regression challenges. Specifically, a Breadth Search Compensation Module (BSCM) resolved the limited detection capability issue observed within inshore scenarios. Neural Discrete Codebook Learning was strategically integrated with Multi-scale Large Kernel Attention, capturing context information around the target and mitigating the information loss inherent in dilated convolutions. To tackle boundary discontinuity arising from the periodic nature of the target regression angle, we developed a Sine Fourier Transform Coding (SFTC) technique. The angle is represented using diverse sine components, and the discrete Fourier transform is applied to convert these periodic components to the frequency domain for processing. Finally, the experimental results of our S4Det on the RSSDD dataset achieved 92.2% mAP and 31+ FPS on an RTXA5000 GPU, which outperformed the prevalent mainstream of the oriented detection network. The robustness of the proposed S4Det was also verified on another public RSDD dataset.

Keywords: remote sensing SAR image; oriented object detection; context information; boundary discontinuity

check for updates

Received: 15 January 2025
Revised: 26 February 2025
Accepted: 3 March 2025
Published: 4 March 2025

Citation: Zhang, M.; Zhu, Y.; Li, L.; Guo, J.; Liu, Z.; Li, Y. S4Det: Breadth and Accurate Sine Single-Stage Ship Detection for Remote Sense SAR Imagery. *Remote Sens.* **2025**, *17*, 900. https://doi.org/10.3390/rs17050900

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ship detection is of great significance for monitoring military cruises, civil transportation, marine search and rescue, and marine fishing. Synthetic Aperture Radar (SAR) is a radar-based technology used for this purpose due to its ability to penetrate clouds and powerful perspective imaging capabilities, and thus, SAR can enable all-weather ocean ship monitoring [1]. Traditional SAR image ship detection methods encompass threshold-based approaches [2], saliency-driven techniques [3], and ship structure analyses [4]. However, these traditional algorithms often demand specialized knowledge and exhibit limitations in robustness to real-world requirements.

Recently, deep learning methods were widely used in the field of computer vision due to their excellent robustness [5–7]. Given the remarkable results achieved by deep learning technology in the field of natural image target detection and the increasing availability of large-scale SAR image datasets, such as RSSDD [8] and RSDD [9], some researchers initiated exploration into applying these advanced technologies for ship detection in SAR images. However, these methods [10,11] generally use Horizontal Bounding Boxes (HBBs) to locate the object, which faces several challenges. Especially in the detection of ship targets in SAR images, the HBBs of different targets often overlap, and for the detection of ship targets with large aspect ratios, there are many unnecessary background noises in the HBBs, which means it is difficult for the horizontal detection method [12,13] to meet the detection requirements. To address this limitation, the scholarly literature suggests employing Oriented Bounding Boxes (OBBs) as a means to represent targets. OBBs can effectively capture the orientation information of targets by including angle information, thereby providing a more accurate representation of the target's position and orientation. For instance, Zhou et al. [14] proposed angle prediction techniques to facilitate OBB implementation, but they may suffer from boundary discontinuity issues due to angle periodicity. Kun et al. [15] introduced a ship detection method based on a scattering keypoint guidance network, leveraging anchor-free networks to reduce computational burdens.

Despite considerable endeavors to enhance performance, certain limitations persist:

(1) The current oriented object detection network makes it difficult to achieve a balance between detection accuracy and speed.

(2) The existing detection networks have limited capability in inshore detection, often due to their inadequate utilization of context information surrounding the target, resulting in an increase in false positives and false negatives within the intricate inshore environment.

(3) The periodicity of the rotation angles in the OBB representation gives rise to challenges of boundary discontinuity problems during the detection process.

To alleviate these problems, we propose a **S**ine **S**ingle-Stage **S**AR image **S**hip oriented detection method called **S4Det** in this paper. To tackle the limited inshore detection capacity, we designed a Breadth Search Compensation Module (BSCM) consisting of Multi-scale Large Kernel Attention (MLKA) and Neural Discrete Codebook Learning (NDCL). The MLKA utilizes dilated convolutions with different kernel sizes to achieve multi-scale information searches, allowing for adaptation to different ship target sizes and facilitating the capture of contextual background information. The NDCL serves to rectify the information loss inherently incurred by dilated convolution, thus successfully reinstating crucial details essential for accurate ship detection. To mitigate the boundary discontinuity, we developed a sine Fourier transform encoding technique. Guided by the Nyquist sampling theorem, we employed discrete sinusoidal components to represent angles with periodicity. Drawing from the wave–particle duality principle, we used the Discrete Fourier Transform (DFT) to superimpose these sine wave components and convert them to the complex domain for processing. This approach enhanced our ability to capture angular information, and thus, effectively alleviated the issue of boundary discontinuity.

In summary, the contribution of this paper is threefold:

- We propose S4Det, a method for rapid and precise ship detection. Experimental results on the RSSDD and RSDD databases show that our method surpassed SOTA while maintaining a comparable speed.
- We developed a BSCM to acquire discrete feature representations suitable for dilated convolutions, thereby enhancing the efficiency of long-range contextual information.
- Our SFTC technique introduced a pioneering application of discrete periodic Fourier transform to manage angle periodicity. By effectively utilizing both the amplitude and phase information of angles, we achieved accurate angle regression.

2. Related Work

2.1. Oriented Object Detection Methods

Oriented object detectors are typically categorized into single-stage and two-stage frameworks based on their pipeline design. Two-stage detectors [16–18] first generate region proposals via a Region Proposal Network (RPN) [10], followed by proposal refinement and classification. While achieving high accuracy, their multi-step cascaded structure incurs significant computational overhead, limiting real-time performance. Single-stage detectors [19–21] eliminate the RPN by directly predicting bounding boxes and categories from feature maps, offering superior inference speed with a competitive accuracy. For SAR ship detection, recent advances highlight the growing preference for efficiency–accuracy balance. Long et al. [22] proposed a single-stage CNN-based detector using rotated boxes, though it struggled with complex inshore backgrounds. Pan et al. [23] introduced a graph convolutional network for global context modeling, but its computational complexity hinders scalability. Zhou et al. [24] proposed PVT-SAR and introduced a Transformer model into SAR image rotating target detection for the first time, but its complicated attention calculation brings a large time cost. Given the urgent demand for processing massive SAR data streams, we adopted a single-stage paradigm and rigorously optimized its speed–precision trade-off through targeted architectural innovations.

2.2. Large Kernel Convolution Methods

The Transformer model [25], which integrates self-attention mechanisms, obtains a wide range of receptive fields, exerting a substantial influence on computer vision. Revisiting Convolution Neural Networks (CNNs) has highlighted the limitation of small convolutional kernels in establishing extensive receptive fields. Hence, there has been a recent adoption of convolutional operations with larger kernels to attain wider receptive fields. RepLKNet [26] utilizes reparameterization to construct super-large 31×31 convolutional kernels. VAN [27] sequentially stacks large kernels with depth-wise convolution and depth-wise dilated convolution to a large receptive field. MAN [28] utilizes large kernel attention at different scales to obtain diverse and fine-grained feature information. Although these methods demonstrate the advantages of large kernel convolutions, they still have certain limitations. For instance, RepLKNet has higher computational complexity. On the other hand, VAN and MAN achieve large kernel convolutions through dilated convolutions, which reduce the network computation but also introduce a certain degree of information loss.

2.3. Methods for Boundary Discontinuity

Given the periodic nature of the rotation angle in OBBs as discussed [29], discrepancies emerge between the predicted and true angles at the period boundary. This discrepancy gives rise to a boundary discontinuity challenge, causing misalignment between predicted and ground truth areas. To tackle this issue, some works analyzed methods from the perspective of loss functions. SCRdet [30] introduced IoU-Smooth L1 to circumvent an abrupt increase in the loss function. KFIoU [31] adopted Gaussian distribution to model the loss function. Another strategy [29,32] involved transforming the angle prediction from a regression problem to an angle classification problem, but they introduced extra hyperparameters and burdened the network inference. Consequently, there emerges a demand for a novel method capable of accurately estimating angular values by employing minimal hyperparameters.

3. Methodology

We provide an overview of our approach illustrated in Figure 1. Our BSCM, consisting of MLKA and NDCL, is integrated into the top layer of the neck network. To improve the network detection head, a convolutional attention mechanism was introduced to reduce the background noise interference. Furthermore, an SFTC technique encodes angle information in the output, which is utilized for loss calculation and network training. In this part, we first introduce the general structure of S4Det, and in Sections 3.2 and 3.3, we focus on the BSCM and SFTC.

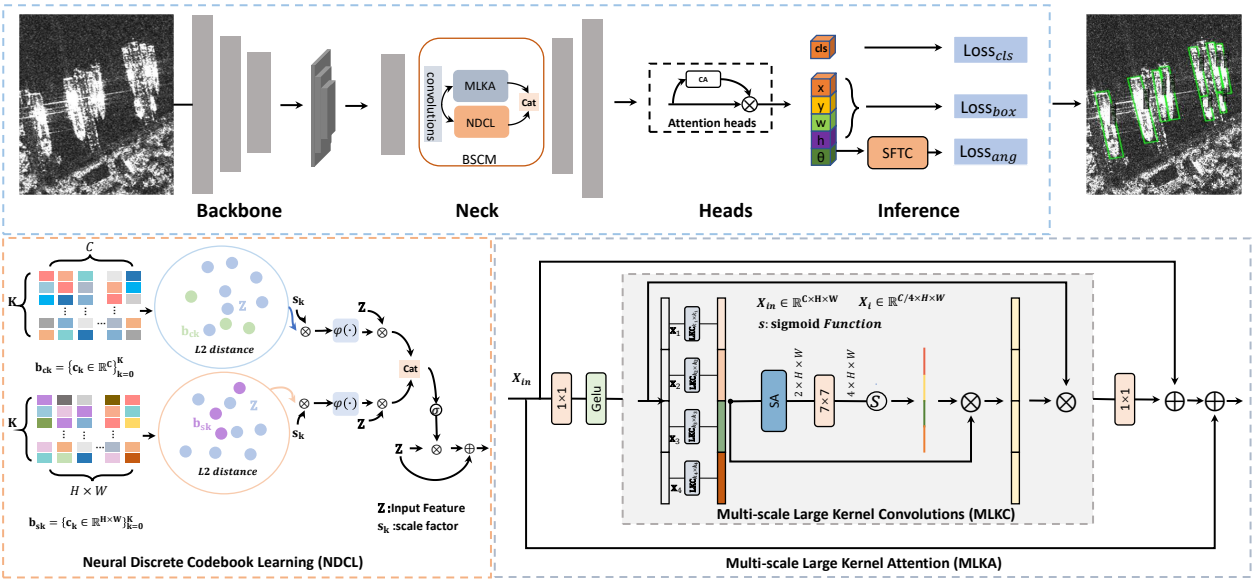

Figure 1. The overall framework of the proposed S4Det. BSCM (comprising MLKA and NDCL) is integrated into the top layer of the neck network, incorporating a convolutional attention mechanism for noise reduction and utilizing SFTC to encode angle information for loss calculation and training.

3.1. Overall Structure of S4Det

In the proposed S4Det method, we adopted a single-stage detector structure to achieve efficient and accurate oriented object detection. Unlike traditional methods that increase the depth of the backbone and neck networks, we used the same basic building block in both networks, effectively reducing the network depth and balancing the parameters. This design not only maintains structural consistency but also accelerates network convergence and inference speed. To address the issue of traditional models focusing solely on inter-layer feature flow and neglecting intra-layer feature selection, we integrated a BSCM into the feature layer. Utilizing dilated convolution, the BSCM effectively captures multi-scale features, reducing the number of network parameters and enhancing the inference efficiency. To mitigate feature information loss caused by dilated convolution, we incorporated dictionary learning [33,34] by representing high-dimensional feature vectors with low-dimensional codebook vectors in a novel approach. Furthermore, to handle the boundary discontinuity caused by angle periodicity, we employed a Fourier transform to process the angle information in the frequency domain. Through these designs, S4Det achieved an optimal performance in both detection accuracy and speed, successfully balancing the two aspects. Specifically, the backbone extracts three-layer feature maps $\{C_3, C_4, C_5\}$ with channel dimensions $\{256, 512, 1024\}$, respectively. Unlike traditional designs that increase the depth of both the backbone and neck, our method employs shared basic building blocks across these components. This design minimizes the number of network parameters while preserving feature diversity and expressive power. Additionally, the neck effectively fuses the extracted features to incorporate multi-scale information, enabling efficient infer-

ence without the computational burden of deeper structures. In addition, as depicted in Figure 2b, to improve the feature representation efficiency, we introduced a BSCM into the neck network. The BSCM effectively captures the target's contextual information within each layer through hollow convolution, extending the receptive field without increasing the number of parameters, thereby reducing the need for additional layers or complex operations. Furthermore, to address the regression challenges stemming from boundary discontinuity, we employed SFTC to encode the angle outputs of the detection head, facilitating frequency-based processing. These enhancements enable our S4Det method to achieve efficient and accurate rotated object detection.

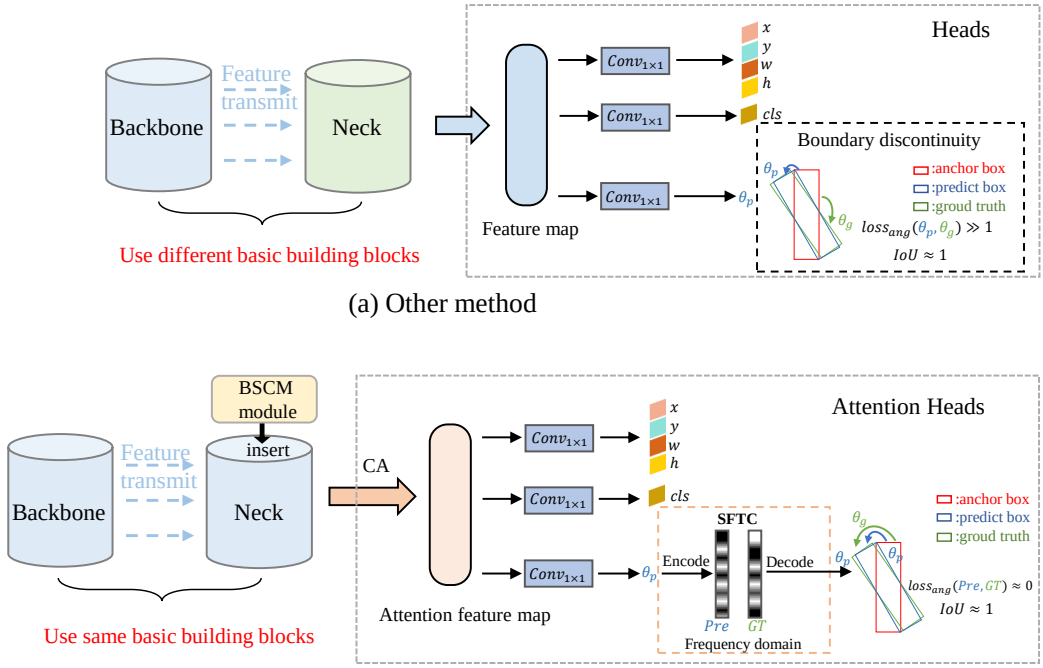

Figure 2. Comparison of the network design of the existing method (**a**) and our method (**b**). In our approach, the backbone and neck share a similar structure, while introducing the BSCM, attention detection head strategy, and angle SFTC.

3.2. Breadth Search Compensation Module (BSCM)

In the inshore scenes, complex background information introduces scattering noise in the SAR imaging mechanism, leading to interference and erroneous detections in the network. To tackle this concern, we propose a BSCM consisting of two main parts: MLKA and NDCL. It enables an extensive information search that leverages the contextual cues surrounding the targets to enhance the recognition, shape furnishing, and positional insights.

$$X_{in} = relu(BN(Conv_{7\times7}(X))). \tag{1}$$

First, we treat the input feature $X \in \mathbb{R}^{C\times H\times W}$ with a 7×7 convolutional layer, a Batch Normalization (BN) layer, and an activation function, where C is the channel number, and H and W give the spatial size of the input. This yields X_{in}, serving as the BSCM input.

Multi-scale Large Kernel Attention (MLKA): We employed MLKA to achieve an extensive information search. The pivotal component of MLKA is Multi-scale Large Kernel Convolution (MLKC). MLKC utilizes various sizes of Large Kernel Convolution (LKC) to create a multi-scale search window. This approach enables the effective selection of appropriate search windows for different-sized ship targets, thereby enhancing the target recognition capability. Specifically, *LKC* is achieved by decomposing the $k \times k$ convolution

kernel into three consecutive convolutional layers, namely, $[\frac{k}{d+1}-1]\times[\frac{k}{d+1}-1]$ depthwise convolution $DWConv$, $[\frac{k}{d+1}+1]\times[\frac{k}{d+1}+1]$ depthwise dilated convolution $DWDConv$ (d is the dilated rate), and 1×1 convolution $Conv_{1\times 1}$, formulated as

$$LKC = Conv_{1\times 1}(DWDConv(DWConv(\cdot))). \tag{2}$$

MLKC constructs four LKC_i with different kernel sizes: 3-5-1, 5-7-1, 7-9-1, and 9-11-1, where a-b-1 means cascading $a\times a$ depth-wise convolution, $b\times b$ depth-wise-dilated convolution, and point-wise convolution. Different from related work [28] that used different expansion rates to realize different scales of receptive fields, this study uniformly set the expansion rate to 3, reduced the setting of hyperparameters, and made the network easier to understand and adjust. Specifically, we first applied a 1×1 convolution and GELU activation function to X_{in}, which obtained $\tilde{X}_{in}$, while preserving both the spatial and channel dimensions. Subsequently, we evenly divided $\tilde{X}_{in}$ into n parts along the channel dimension $(\tilde{X}_1,\tilde{X}_2,\ldots,\tilde{X}_n)$. Each $\tilde{X}_i$ underwent processing through LKC_i, and their outcomes were concatenated along the channel dimension to construct feature information X_{mlkc} with different receptive fields, formulated as follows:

$$X_{mlkc} = Concat\left[LKC_1(\tilde{X}_1),\ldots,LKC_4(\tilde{X}_4)\right], \tag{3}$$

where $Concat(\cdot)$ denotes the feature map concatenation along the channel dimension.

To enhance the connection between different receptive fields, we employed average pooling and max pooling to process X_{mlkc}, which effectively extracts spatial relationships from different receptive fields.

$$SA = Concat\left[F_{avg}(X_{mlkc}),F_{min}(X_{mlkc})\right], \tag{4}$$

$$SA_i = Conv^{2\to 4}{}_{7\times 7}(SA),(i=1,2,3,4), \tag{5}$$

where F_{avg} and F_{min} are the max pooling and average pooling operators, which both reduce the channel dimension to 1. By concatenating these outcomes to yield spatial attention (SA) with a channel size of 2, we subsequently utilized a 7×7 convolution to expand the channel size to 4 to match the four distinct receptive fields. The sigmoid function processes SA_i to capture crucial information. Multiplying and summing the processed SA_i with LKC_i achieves effective spatial information fusion, detailed as

$$MSA = \sum_{i=1}^{4}(\sigma(SA_i)\cdot LKC_i(X_i)), \tag{6}$$

$$X_{mlkc} = \tilde{X}_{in}\cdot MSA. \tag{7}$$

where $\sigma()$ represents the sigmoid function. We employed SA_i to selectively extract feature information from different receptive fields and subsequently summed these values to derive the multi-head spatial attention (MSA). We multiplied $\tilde{X}_{in}$ with MSA to obtain the output of the MLKC component.

Finally, we performed convolutional processing on X_{mlkc} and applied skip connections to obtain the MLKA output Y_{mlka}.

Neural Discrete Codebook Learning (NDCL): MLKA adopts a strategy of employing dilated convolutions to achieve extensive information exploration. However, due to the presence of holes in convolutions, there might be a risk of information loss. To mitigate this loss, we introduced the NDCL method, which involves learning discrete information through the codebook, thereby compensating for the potential information deficiency that can arise from MLKA. As shown in Figure 3, the feature heatmap of MLKA had a

large receptive field but lacked local detail attention. NDCL made up for this shortage and made the heatmap show more sensitive and detailed features in local areas. Finally, BSCM combined the output of MLKA and NDCL to achieve an accurate capture of the global information.

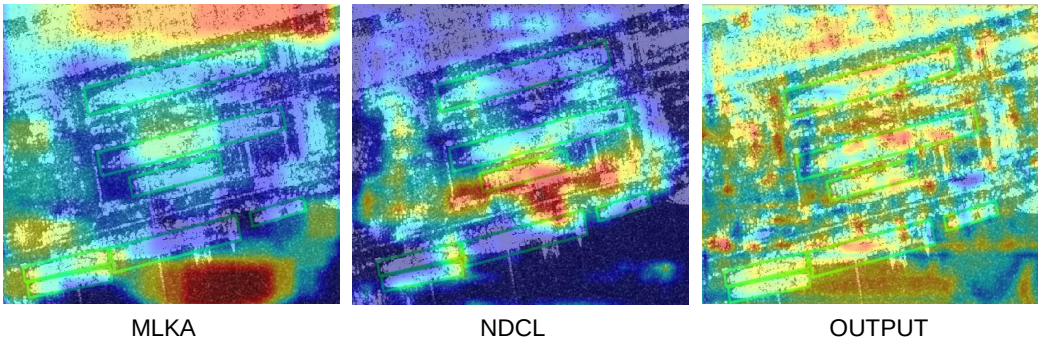

Figure 3. Feature heatmap visualization of the outputs from MLKA, NDCL, and BSCM.

For the input feature $X_{in} \in \mathbb{R}^{C\times H\times W}$, we first obtained $Z \in \mathbb{R}^{C\times H\times W}$ through the stem block, which was then integrated into our NDCL module. We utilized a learnable codebook $b = \{c_k \in \mathbb{R}^N\}_{k=1}^K$ to represent the dimensional information in Z, where K signifies the number of codewords c_k and N is the dimension of each codeword. By employing KN-dimensional codewords, we discretely represented Z, which effectively compensated for fine-grained information. Unlike previous dictionary-learning methods [35,36] that only establish codewords in the channel dimension, we extended this concept to include codewords within the spatial dimensions (HW) to achieve a three-dimensional representation of local information. This was accomplished as follows:

$$b_{ck} = \left\{c_k \in \mathbb{R}^C\right\}_{k=1}^{\mathrm{K}}, b_{sk} = \left\{c_k \in \mathbb{R}^{HW}\right\}_{k=1}^{\mathrm{K}}, \tag{8}$$

where b_{ck} and b_{sk} are the codebooks in the channel and spatial dimensions, respectively. c_k represents the k-th codeword. We replaced the corresponding dimensions of Z with the codewords to obtain the quantized feature v. Additionally, we employed a learnable scale factor s_k to adjust the similarity between the codeword and the dimensional information, whether in the channel or spatial dimensions.

$$v_{ck} = \sum_{i=1}^{C} \mathrm{S}\left(s_k||Z^i - b_{ck}||\right) \cdot \left(Z^i - b_{ck}\right), \tag{9}$$

$$v_{sk} = \sum_{j=1}^{H\times W} \mathrm{S}\left(s_k||Z^j - b_{sk}||\right) \cdot \left(Z^j - b_{sk}\right), \tag{10}$$

where Z^i is the information in the channel dimension. Z^j represents the feature information of each pixel in the spatial dimension. s_k is the k-th scaling factor. $S(\cdot)$ denotes the softmax function. v_{ck} and v_{sk} mean the k-th quantized channel and spatial information, respectively. We computed the $L2$ distance between Z and the codewords using $||\cdot||$ and subsequently employed the softmax function to yield smoothed features. Following this, we employed $\varphi(\cdot)$ to combine all v_{ck} and v_{sk}, where $\varphi(\cdot)$ comprises a BN layer with a ReLU activation layer and a mean layer. Based on this, the full information of the whole image with respect to the K codewords is calculated:

$$e_1 = \sum_{k=1}^{K} \varphi(v_{ck}), e_2 = \sum_{k=1}^{K} \varphi(v_{sk}). \tag{11}$$

We performed element-wise multiplication of the codewords and the input vector Z along the channel dimension, followed by summing the products. The output value a was obtained by applying the sigmoid function to the sum:

$$a = \delta(Concat(e_1 \cdot Z, e_2 \cdot Z)), \tag{12}$$

where $\sigma(\cdot)$ represents the sigmoid function. The outcome of the NDCL could be determined using the following equation:

$$Y_{ndcl} = a \cdot Z + Z, \tag{13}$$

where a aggregates the codeword information of the channel and space to adjust the required information by multiplying it with the feature Z.

Finally, we fused Y_{mlka} and Y_{ndcl} along the channel dimension to achieve the output of the BSCM:

$$X_{out} = Concat(Y_{mlka}, Y_{ndcl}). \tag{14}$$

3.3. Sine Fourier Transform Coding (SFTC)

In order to deal with the problem of boundary discontinuity caused by rotation angle periodicity, this section mainly introduces the encoding and decoding process of the detection box angle information predicted by the detection head. As shown in Figure 4, we sine-encoded the predicted angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$. The angle was encoded using a four-step phase shift method [37], where the initial phases were set at 0, 90, 180, and 270 degrees. This angle representation method complies with the sampling theorem and possesses encoding fault tolerance. The angle is represented by the sine function of two different frequencies:

$$\phi_1 = 2\theta, \quad \phi_2 = 4\theta, \tag{15}$$

where ϕ_1 and ϕ_2 are two frequencies representing the conversion relationship with the predicted angle θ.

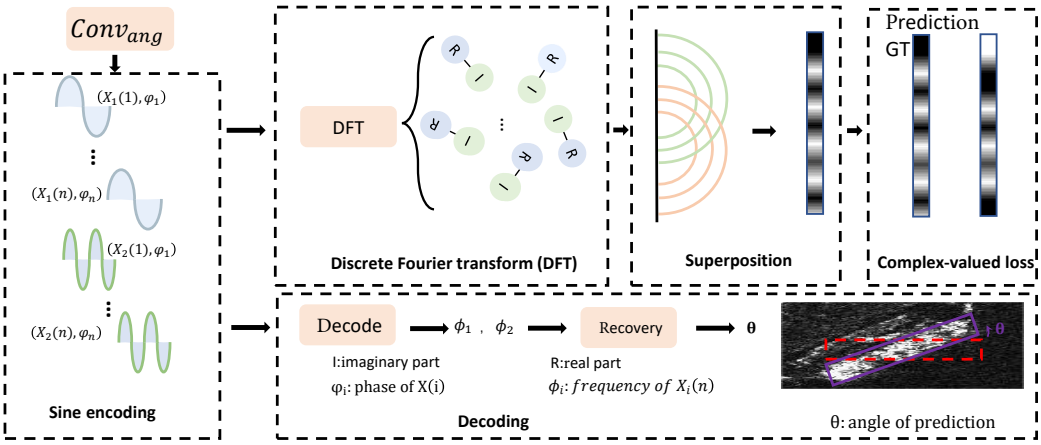

Figure 4. Coding and decoding process of SFTC. Sine encoding is applied to the predicted angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$, using a four-step phase shift method with initial phases set at 0, 90, 180, and 270 degrees.

Sine encoding: ϕ_1 and ϕ_2 are encoded as X_1 and X_2 using sine functions as follows:

$$\begin{cases} X_1(n) = \sin(\phi_1 + 2\pi n/M + \frac{\pi}{2}) \\ X_2(n) = \sin(\phi_2 + 2\pi n/M + \frac{\pi}{2}), \end{cases} \tag{16}$$

where $n = 1, 2, \ldots, M$ and M is the number of sine components. We can deduce that ϕ_1 represents a rotation period of π, while ϕ_2 corresponds to a rotation period of $\pi/2$, corresponding to the period of the rectangular OBB and square OBB.

Discrete Fourier transform (DFT): Directly performing regression calculations on these sinusoidal components will result in the loss of phase information of the components. Inspired by the wave–particle duality in quantum mechanics, a wave typically contains both amplitude and phase attributes, and the wave equation of a particle can fully describe its state from amplitude and phase. Therefore, we regard the sinusoidal wave components as the wave equation of free particles. Furthermore, there is a connection between the wave equation of free particles and the discrete Fourier transform. In the wave equation, the frequency is related to the oscillatory nature of the wave function, while in the Fourier transform, frequency represents the signal intensity. The superposition of free particle wave functions is analogous to spatial superposition, while superposition in the Fourier transform occurs in the frequency domain. Based on interference effects, the superposition state of wave functions reflects the phase relationship between particles. With the assistance of DFT, we realized the superposition of particles (sine components) to allow for a better observation of wave phenomena in the results and comprehensive utilization of amplitude and phase information to describe angles. The superposition of wave equations is shown as follows:

$$\begin{aligned} X(k) = &\sum_{n=1}^{\frac{N}{2}} X_1(n) e^{-j\frac{2\pi}{N}kn} + \\ &\sum_{n=\frac{N}{2}+1}^{N} X_2(n - \tfrac{N}{2}) e^{-j\frac{2\pi}{N}kn}, \end{aligned} \tag{17}$$

where $X(k)$ is the frequency domain representation. $X_1(n)$ and $X_2(n)$ represent the discrete values of particles at different frequencies. k denotes the wave number. N is the number of particles where $N = 8$. e means the base of the natural logarithm. j refers to the imaginary unit. $X_i(n)$ and $-\frac{2\pi}{N}kn$ are the amplitude and phase, respectively.

Decoding function: The formula for decoding ϕ from $X(n)$ can be described as

$$F(X(n)) = -\arctan \frac{\sum_{n=1}^{M} X(n) \sin(\frac{2n\pi}{M})}{\sum_{n=1}^{M} X(n) \cos(\frac{2n\pi}{M})}, \tag{18}$$

where ϕ_1 and ϕ_2 are calculated as follows:

$$\phi_1 = F(X_1(n)), \quad \phi_2 = F(X_2(n)), \tag{19}$$

where ϕ_2 has a twofold frequency relationship with respect to ϕ_1. We calculated the cosine of the difference between ϕ_1 and ϕ_2 to help restore the predicted angle:

$$\lambda = \cos\left(\phi_1 - \frac{\phi_2}{2}\right), \tag{20}$$

where λ denotes the cosine value of the angular difference between two phases and was utilized to recover the ultimate predicted angle.

The formula for restoring the predicted angle using λ is

$$\theta = \begin{cases} \frac{\pi}{2} + \frac{\phi_2}{4}, & if\ \lambda < 0 \\ \frac{\phi_2}{4}, & else \end{cases}, \tag{21}$$

where θ is the predicted angle output by the network.

3.4. Loss Function

The loss function consists of three components: the classification, bounding box, and angle loss as follows:

$$Loss = \lambda_1 Loss_{cls} + \lambda_2 Loss_{box} + \lambda_3 Loss_{ang}, \tag{22}$$

where λ_1, λ_2, and λ_3 are the loss weights for the classification, bounding box, and angle. Here, $\lambda_1 = 1$, $\lambda_2 = 2$, and $\lambda_3 = 0.1$. For the classification loss, we adopted the quality focal loss [38]:

$$Loss_{cls} = -|y - \tilde{p}|^{\beta}(1 - y)\log(1 - \tilde{p}) + y\log(\tilde{p}), \tag{23}$$

where y denotes the Intersection over Union (IoU) score. $\tilde{p}$ represents the predicted score. β is an adjustable factor. We set β to 2. For the bounding box loss, we applied the rotated IoU as the regression cost:

$$Loss_{box} = -\log(IoU). \tag{24}$$

For the angle loss, we calculated the sum of the absolute differences between the target and predicted real parts and the absolute differences between the target and predicted imaginary parts:

$$Loss_{ang} = |R_{tar} - R_{pre}| + |I_{tar} - I_{pre}|, \tag{25}$$

where R represents real parts and I represents imaginary parts.

4. Experiments and Results

4.1. Experimental Settings

Datasets: RSSDD serves as a challenging public benchmark for SAR ship detection, where it is characterized by OBB annotations. It encompasses SAR images with varied resolutions and polarizations, totaling 1160 images. The sizes of the SAR images vary, with an average size of 481 × 331. We utilized the preprocessing provided by MMRotate, where the images were scaled to 512 × 512 for training and testing. The dataset bifurcates into two scenarios: inshore and offshore. In the former, images often contain a significant amount of speckle noise in the background, necessitating robust noise reduction capabilities. In the latter, the background is relatively uniform, but most ship targets are small in size, which challenges the network's ability to recognize small objects. For evaluation, we partitioned the test set into inshore and offshore scenarios.

The RSDD dataset is a comprehensive SAR ship detection dataset covering a variety of imaging modalities, polarimetric modes, and resolutions. The dataset is structured into four main folders: JPEGImages, Annotations, Imagesets, and JPEGValidation. JPEGImages contains 7000 slices of 512 × 512 pixels. The annotations include the corresponding 7000 oblique box annotation files, which are annotated by the OpenCV long-edge definition method based on the COCO format and record the target center point, long and short edges, and rotation angle. Imagesets define training and test sets, including inshore and offshore sets. JPEGValidation contains images of both scenarios and their annotations.

Implementation details: The configuration of our single GPU training environment is as follows: Python 3.8.16, PyTorch 1.10.1, GCC 7.3, CUDA 11.3, and MMRotate 1.0. During the training, we used the AdamW optimizer. The momentum was set to 0.9, and the weight decreased to 0.0001. The initial learning rate was set to 2.5×10^{-4}. The maximum number of iterations of our proposed S4Det was 180 epochs on the RSSDD dataset and 36 epochs on the RSDD dataset. We adopted a cosine decay learning rate strategy, which started to decay from around half of the maximum epochs until reaching 0.05 of the initial learning rate. For evaluating the model performance, we adopted the same mAP calculation method as the VOC metric on the RSSDD dataset, and the COCO metric was used to compare the

RSDD dataset. To comprehensively evaluate the performance of the detection network, we also evaluated the performance of the network using the recall, precision, and frames per second (FPS). Recall measures the ability of the model to correctly detect positive samples. Precision reflects the accuracy of its detection results. FPS provides a measure of the processing speed of the network.

Evaluation metrics: Precision (P_d), recall (R_d), average recall (AR), and mAP are commonly employed metrics for model evaluation. For the precision and recall, they can be calculated as follows:

$$\begin{cases} P_d = \frac{TP}{TP+FP} \\ R_d = \frac{TP}{TP+FN} \end{cases}, \tag{26}$$

where TP represents the number of correctly detected targets. FP represents the false alarms. FN denotes the missed targets.

The AR is the average of the IoU over all recalls on $[0.5, 0.95]$, which is defined as

$$AR = \int_{0.5}^{0.95} Recall(IoU)d(IoU). \tag{27}$$

The AP metric quantitatively evaluates the overall detection performance of a detector by calculating the area under the precision–recall curve. It can measure the overall detection performance of the detector at different thresholds and is defined as follows:

$$AP = \int_0^1 P_d(R_d)dR_d, \tag{28}$$

where FPS stands for how many images the model can process per second and is used to measure the detection speed of a detector, which is defined as

$$FPS = \frac{1}{Times}, \tag{29}$$

where $Times$ is the average detection time for each image.

4.2. Ablation Study

To enhance the inshore detection accuracy and reduce false detections and omissions, we introduce the BSCM to employ multi-scale convolution kernels with dilation to extract object features of varying scales and capture contextual information, followed by an attention detection head strategy to suppress background noise effectively. Additionally, we propose an SFTC method to tackle angle periodicity boundary issues. Exhaustive ablation experiments validated the effectiveness of our approach by demonstrating significant accuracy improvement and robustness across various hyperparameter configurations and module combinations.

Impact of each component in S4Det: To rigorously evaluate the contribution of each component within the S4Det framework, we executed a comprehensive set of ablation studies on the RSSDD dataset. The outcomes of these ablation experiments are meticulously documented in Table 1. When individual parts were used separately, there were substantial enhancements in both the precision and mAP on the baseline RTMDet [39], underscoring their effectiveness. Notably, the synergistic integration of all components resulted in a remarkable enhancement of the network performance. Specifically, the accuracy increased by 10.9%, where it improved from 77.1% to 88.0%. This integration notably curtailed the instances of false positives and false negatives, which signified a robust reduction in

erroneous detections. Concurrently, the mAP witnessed an uplift of 1.6% to reach the SOTA level.

Table 1. Impacts of each component in S4Det.

Baseline	Att-Head	BSCM	SFTC	All Scenes	
				Precision	mAP
RTMDet				77.1	90.6
	✓			82.7	90.8
		✓		78.9	91.6
			✓	78.2	91.2
	✓	✓	✓	**88.0**	**92.2**

Impact of each component in the BSCM: As shown in Table 2, the MLKA and NDCL were individually applied to the baseline network. Compared with the use of the MLKA and NDCL together, it can be concluded that the combination of the MLKA and NDCL could effectively improve the detection performance.

Table 2. Impacts of each component in BSCM.

Baseline	MLKA	NDCL	All Scenes	
			Precision	mAP
RTMDet	✓		77.1	88.8
		✓	74.4	89.1
	✓	✓	**78.9**	**91.6**

Impact of kernel and dilation rate design: MLKC can be composed of different LKC scales, which can be represented by $a - b - 1$, where $a = b - 2$. To observe the effect of different settings of b, we first performed ablation experiments on b. As shown in Table 3, we obtained the best results under the settings of $(5, 7, 9, 11)$. In addition, we performed ablation experiments with related dilation rate settings in Table 4, including the same dilation rate and different dilation rate settings. It can be seen from this table that the dilation rate setting did not increase the amount of calculation and network parameters. In addition, according to the calculation formula of the expanded receptive field, the formula is as follows:

$$RF = 1 + d \times (k - 1), \tag{30}$$

where RF represents the receptive field size, d represents the dilation rate, and k represents the convolution kernel size. Finally, the best effect was achieved with an expansion rate of $(3, 3, 3, 3)$. This setting enabled the kernel sizes of 5, 7, 9, and 11 to generate receptive fields of 13, 19, 25, and 31. Visualization of the target size distribution in Figure 5 demonstrates that these receptive fields aligned well with the target size and captured the surrounding information effectively. As shown in Figure 6, we also provide the learning curves of the loss values during training for the numerical results presented in Table 4.

Table 3. Multi-scale b kernel design. The grey background represents the optimal choice.

b Kernel Design	FLOPs (G)	Params (M)	Precision	mAP
(3, 5, 7, 9)	67.37	68.31	76.5	89.2
(5, 7, 9, 11)	67.43	68.36	78.9	91.6
(7, 9, 11, 13)	67.50	68.44	68.2	91.3

Table 4. Dilation design. The grey background represents the optimal choice.

Dilation Design	FLOPs (G)	Params (M)	Precision	mAP
(1, 1, 1, 1)	67.43	68.36	76.3	91.4
(2, 2, 2, 2)	67.43	68.36	77.7	90.0
(3, 3, 3, 3)	67.43	68.36	78.9	91.6
(4, 4, 4, 4)	67.43	68.36	77.2	90.9
(1, 2, 3, 4)	67.43	68.36	78.4	90.6

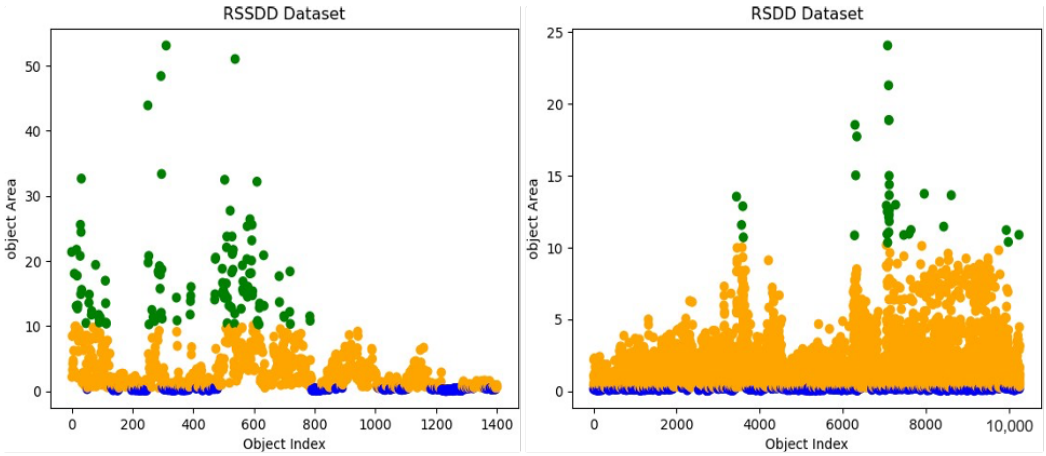

Figure 5. Target area distributions.

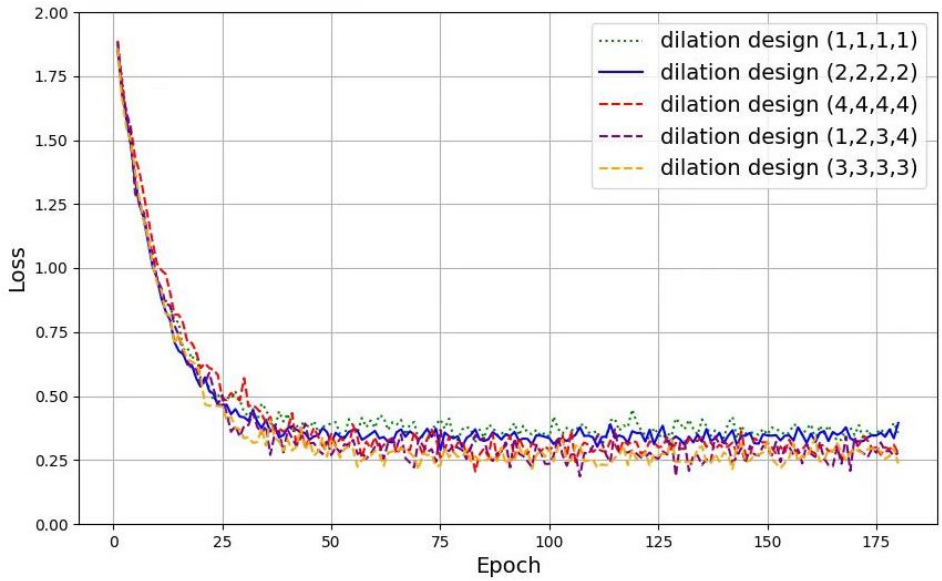

Figure 6. Learning curves of the loss values for different dilation designs.

Impacts of different angle-encoding methods: We compared our angle-encoding method SFTC with other methods, namely, CSL [29] and PSCD [32], on the baseline network RTMDet, as shown in Table 5. The results indicate that the CSL encoding performed poorly in the SAR image detection due to its complex parameters. Although the PSCD encoding method is simple and effective, it lacked sufficient emphasis on the phase and amplitude information of the angle, which led to inaccurate regression. In contrast, our SFTC encoding achieved the best performance among similar methods for efficiently solving the boundary discontinuity. The precision and mAP increased from 77.1 and 90.6 to 78.2 and 91.2 when using angle extraction compared with not using angle extraction. In Figure 7, we include the training loss curves corresponding to the numerical results in Table 5.

Table 5. Performance comparison of different angle coding methods.

Method	All Scenes		
	Recall	Precision	mAP
RTMDet	92.5	77.1	90.6
+CSL	73.7	48.0	68.9
+PSCD	92.3	74.3	90.6
+SFTC	**92.5**	**78.2**	**91.2**

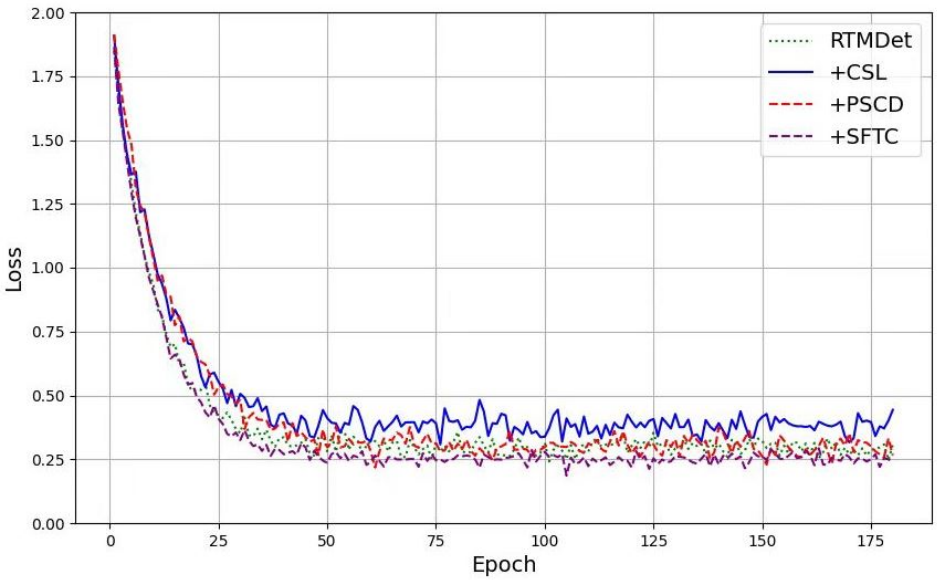

Figure 7. Learning curves of the loss values for different modules.

Impacts of the number of sine components: As shown in Figure 8, we evaluated different numbers of components. Theoretically, setting the number of components to 3 or more was sufficient to meet the Nyquist sampling theorem. We can observe that the performance was the best when M was set to 4.

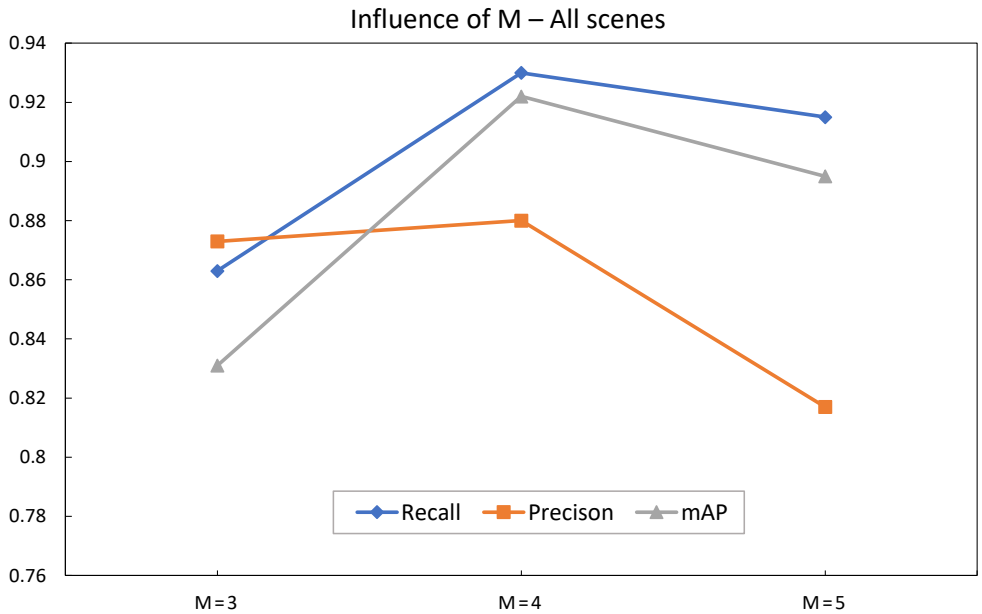

Figure 8. Detection performance for different M values in all scenes.

4.3. Comparisons with the State of the Art

In this section, we compare our method with the current leading detectors, namely, two-stage detectors, like Oriented RCNN [17] and RoI Transformer [16], as well as one-stage detectors, like R3Det [19], S2ANet [20], and KFIoU [31]. To maintain uniformity in the network scale, other detectors also utilized a 152-layer ResNet [40] as the backbone.

Comparisons on different datasets: From Table 6, it can be observed that our single-stage detection approach S4Det achieved the highest levels of recall, precision, and mAP in inshore scenarios on RSSDD dataset. Moreover, in offshore scenarios, we emerged as the best-performing single-stage detector, with mAP and recall standing at a near-optimal level. Due to the relatively consistent background in maritime scenes, which typically includes ocean backgrounds with minimal detection interference, the advantage of the RPN of the two-stage network became evident. As a result, the two-stage detector performed exceptionally well in distant shore scenarios. Without dividing the test set into different scenes, our S4Det achieved an outstanding mAP of 92.2%, which was the best performance. In addition, we include the training loss curves (Figure 9) corresponding to the numerical results in Table 6.

Table 6. Comparison of different methods on the RSSDD test set. Bold items indicate optimal values in a column, and underlined items indicate suboptimal values in a column.

Method	Stage	Inshore			Offshore			All Scenes
		Recall	Precision	mAP	Recall	Precision	mAP	mAP
Oriented RCNN	Two	84.0	<u>82.7</u>	80.6	92.0	**93.9**	91.6	89.1
RoI Transformer	Two	<u>84.7</u>	73.5	<u>81.3</u>	**94.8**	<u>92.2</u>	**94.4**	<u>91.2</u>
S2ANet	One	81.7	53.0	77.0	93.0	58.7	92.0	88.6
R3Det	One	80.9	35.2	76.8	91.5	32.9	90.6	87.4
KFIoU	One	81.7	38.8	76.0	93.8	52.1	92.9	89.2
Our method	One	**91.6**	**90.2**	**89.8**	<u>94.3</u>	87.3	<u>93.2</u>	**92.2**

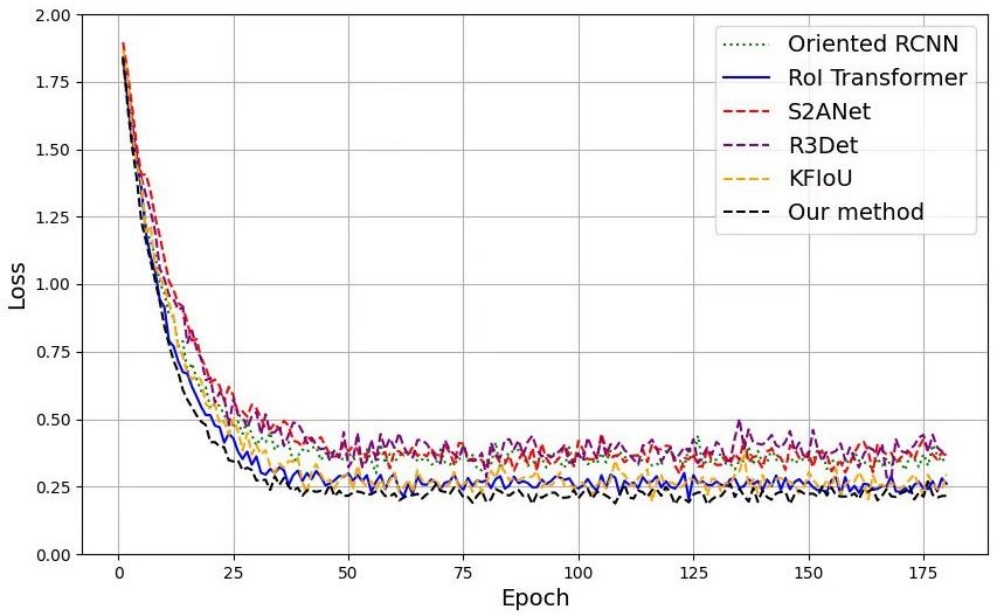

Figure 9. Learning curves of the loss values for different methods.

From Table 7, our proposed S4Det also performed best in the inshore scene on the RSDD dataset. In the far-sea scene, the RPN part of the two-stage network had the advantage of accurate detection because most of the background in the far-sea was simply without complex speckle noise, and the performance of our S4Det was slightly lower than

the best-performing two-stage network RoI Transformer. In all the test scenarios, our S4Det achieved a competitive AP_{75} score of 50.8%, slightly below the RoI Transformer's 51.1%. This difference was attributed to the RoI Transformer's RPN, which offers advantages in detecting small targets. However, S4Det excelled with optimal scores in AP_{50} (91.4%), mAP (50.0%), and average recall (AR) (55.4%). Figure 10 shows the AP (average precision) curve comparison between our S4Det method and other methods under different thresholds during the training process with the same number of iterations. It is evident from the figure that the S4Det method achieved the highest detection accuracy. In addition, we include an analysis of the method's performance on the SSDD dataset with high-resolution images (608 × 608) in Table 8. It indicates that S4Det had a stronger generalization ability.

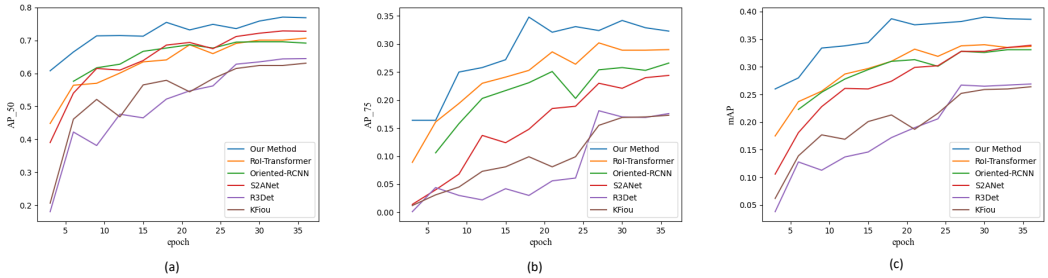

Figure 10. *AP* curves on RSDD. (**a**) AP_{50} curve. (**b**) AP_{75} curve. (**c**) mAP curve.

Table 7. Comparison of different methods on the RSDD test set. Bold items indicate optimal values in a column, and underlined items indicate suboptimal values in a column.

Method	Stage	Inshore			
		AP_{50}	AP_{75}	$mAP_{(50:95)}$	$AR_{(50:95)}$
Oriented RCNN	Two	69.2	26.6	33.1	40.3
RoI Transformer	Two	70.7	<u>29.0</u>	<u>33.7</u>	<u>40.4</u>
S2ANet	One	<u>72.9</u>	24.0	33.5	39.7
R3Det	One	64.5	17.6	26.9	34.1
KFIoU	One	63.1	17.3	26.4	35.5
Our method	One	**77.1**	**32.9**	**38.7**	**45.8**
Method	Stage	Offshore			
		AP_{50}	AP_{75}	$mAP_{(50:95)}$	$AR_{(50:95)}$
Oriented RCNN	Two	93.6	53.5	52.3	57.1
RoI Transformer	Two	**95.4**	**55.7**	**53.4**	**58.2**
S2ANet	One	94.2	49.4	50.1	55.5
R3Det	One	93.2	43.3	47.9	53.5
KFIoU	One	91.9	40.4	45.7	53.3
Our method	One	<u>94.3</u>	<u>54.7</u>	<u>52.6</u>	<u>57.8</u>
Method	Stage	All Scenes			
		AP_{50}	AP_{75}	$mAP_{(50:95)}$	$AR_{(50:95)}$
Oriented RCNN	Two	89.9	48.8	48.8	53.8
RoI Transformer	Two	<u>90.9</u>	**51.1**	<u>49.7</u>	<u>54.6</u>
S2ANet	One	90.7	44.5	47.0	52.3
R3Det	One	88.4	38.5	44.1	49.6
KFIoU	One	86.6	35.8	42.1	49.8
Our method	One	**91.4**	<u>50.8</u>	**50.0**	**55.4**

Table 8. Comparison of different methods on the SSDD test set. Bold items indicate optimal values in a column.

Method	All Scenes		
	Precision	Recall	mAP
RoI Transformer	88.8	89.57	86.98
S2ANet	87.48	89.56	85.47
LPST-Det	93.32	93.17	92.1
Our method	**93.73**	**95.68**	**94.30**

Visual results: To provide a visual comparison between the different methods, we present the visual results of the various methods on different RSSDD dataset images in Figure 11. Upon observation, it is evident that apart from our method, the other approaches show instances of false positives and missed detections in these images. Specifically, focusing on densely distributed ship targets, the other methods incorrectly identified two closely positioned ships as a single target, whereas only our method accurately identified both targets. Similarly, in other images with densely distributed ships, the issue of missed detections persisted for the other networks, which our method effectively addressed. This underscores the effectiveness of the proposed BSCM for global information retrieval and SFTC for precise angle representation. In specific instances, such as the first image, ship scattering resulted in artifacts, which were erroneously identified as targets by S2ANet and KFIoU. In the second image, except for the proposed S4Det, all the other methods incorrectly identified two densely distributed ships as one target. For the far-off scenario depicted in the fourth image, only our method avoided recognizing the scattering noise from underwater reefs as targets. In summary, the proposed S4Det achieved accurate detection in both the inshore and far-off scenarios, thus affirming the effectiveness of our approach.

We also make a visual comparison of the related methods on the RSDD dataset in Figure 12 and select representative error-prone images. The visualization results show that other methods always produced a false detection or missed detection in these complex scenes, and the performance of our method showed the ability of our network to accurately detect in complex situations.

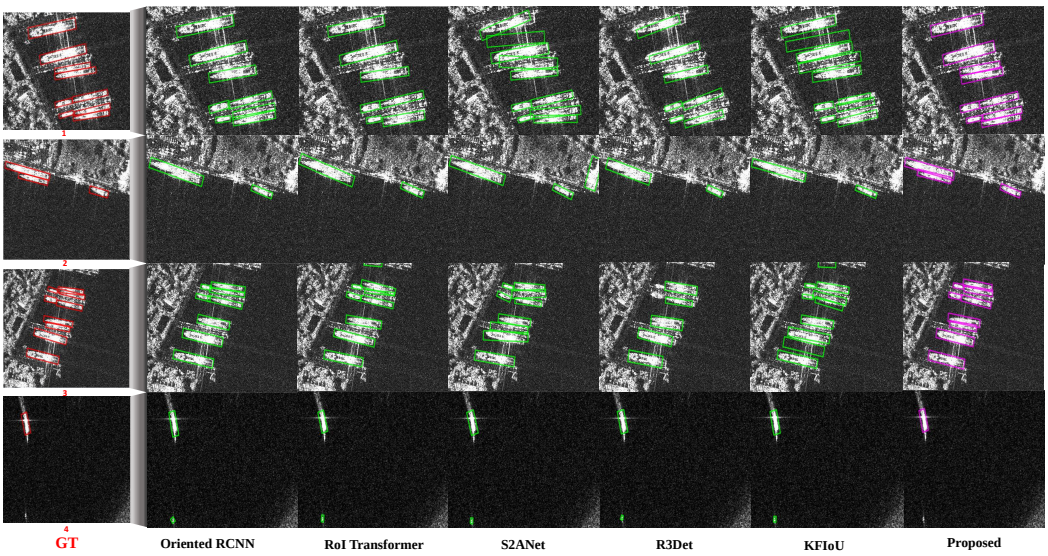

Figure 11. Visualization of the detection results of different methods on RSSDD. Red rectangles indicate the actual ship targets. Green and purple rectangles represent the detection results of five comparative methods and our method, respectively.

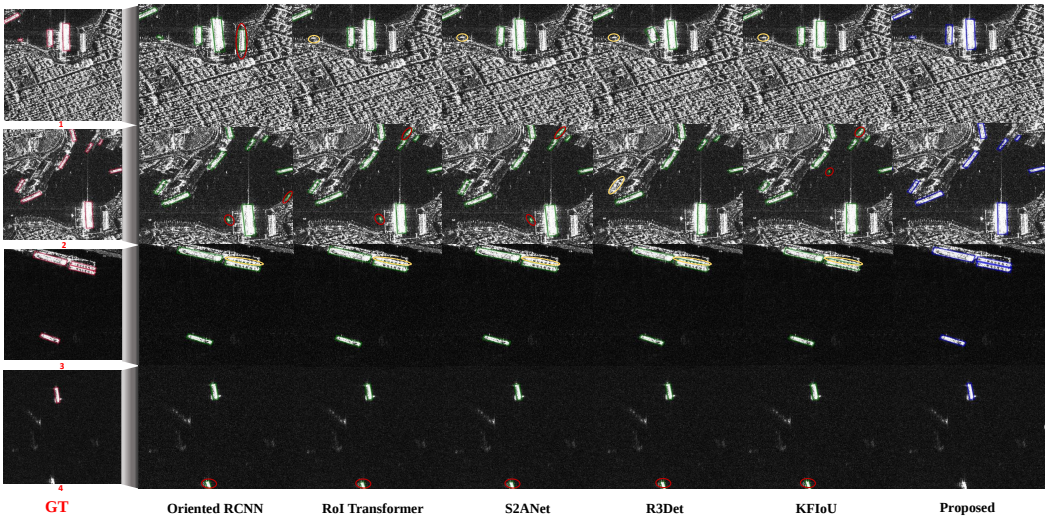

Figure 12. Visualization of the detection results of different methods on RSDD. The red rectangular detection boxes represent the actual ground truth annotation. The green and blue rectangular detection boxes denote the detection results of the five comparison methods and the proposed method, respectively. The red ellipses in the picture represent false detection, and the yellow ellipses represent missed detection.

4.4. Speed and Complexity Analysis

Table 9 provides a comparison between our approach and other methods in terms of the inference speed and complexity. Compared with the two-stage networks, Oriented RCNN and RoI Transformer, the proposed S4Det demonstrated a significant advantage in inference speed. Moreover, our method achieved optimal levels of the network FLOPs and model parameter count. When compared with the single-stage networks S2ANet and KFIoU, our approach maintained a speed advantage, and the 68.69 M parameter count was also superior to that of S2ANet (70.81 M). In summary, our method outperformed the SOTA in both inference speed and model complexity.

Table 9. Inference time and complexity of different methods.

Method	FPS (img/s)	FLOPs (G)	Params (M)
Oriented RCNN	23.8	102.25	75.99
RoI Transformer	23.6	116.22	89.76
S2ANet	24.3	88.02	70.81
R3Det	22	150.92	81.67
KFIoU	24.4	121.32	76.25
Our method	**31.1**	**64.25**	**68.69**

5. Conclusions

In this paper, we introduce a single-stage detector, S4Det, for ship detection in SAR images. We first propose a BSCM to enable broader and more detailed information search capabilities, thereby reducing the false positive and false negative rates in target detection. We also designed SFTC to improve the accuracy of target positioning by using amplitude and phase information to jointly determine angle predictions. Through experiments on RSSDD and RSDD databases, we achieved the best balance between speed and precision. These improvements not only improved the detection performance, but also provided valuable insights and directions for our future research.

Author Contributions: Methodology, M.Z.; Software, Y.Z.; Validation, Z.L.; Writing—original draft, L.L.; Writing—review & editing, J.G.; Supervision, Y.L. All authors read and agreed to the published version of this manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62272363 and Grant 92470108, in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2021QNRC001, and in part by the Joint Laboratory for Innovation in Satellite-Borne Computers and Electronics Technology Open Fund 2023 under Grant 2024KFKT001-1.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

SAR	Synthetic Aperture Radar
BSCM	Breadth Search Compensation Module
SFTC	Sine Fourier Transform Coding
OBBs	Oriented Bounding Boxes
CA	Convolutional attention
MLKA	Multi-scale Large Kernel Attention
NDCL	Neural Discrete Codebook Learning
CNN	Convolution Neural Network

References

1. Zhang, M.; He, C.; Zhang, J.; Yang, Y.; Peng, X.; Guo, J. SAR-to-Optical Image Translation via Neural Partial Differential Equations. In Proceedings of the IJCAI, Vienna, Austria, 23–29 July 2022; pp. 1644–1650.
2. Xing, X.; Chen, Z.; Zou, H.; Zhou, S. A fast algorithm based on two-stage cfar for detecting ships in sar images. In Proceedings of the 2009 2nd Asian-Pacific Conference on Synthetic Aperture Radar, Xi'an, China, 26–30 October 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 506–509.
3. Wang, Y.; Liu, H. A hierarchical ship detection scheme for high-resolution sar images. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 4173–4184. [CrossRef]
4. Duan, C.; Hu, W.; Du, X. SAR image based geometrical feature extraction of ships. In Proceedings of the 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 24–29 July 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 2547–2550.
5. Zhang, M.; Zhang, R.; Zhang, J.; Guo, J.; Li, Y.; Gao, X. Dim2Clear Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
6. Zhang, M.; Bai, H.; Zhang, J.; Guo, J. Rkformer: Runge-kutta transformer with random-connection attention for infrared small target detection. In Proceedings of the 30th ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1730–1738.
7. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 877–886.
8. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
9. Xu, C.; Su, H.; Li, J.; Li, Y.; Ya, L.; Ga, L.; Ya, W.; Wa, T. RSDD-SAR: Rotated ship detection dataset in SAR images. *J. Radar* **2022**, *11*, 581–599.
10. Liu, T.; Yang, Z.; Gao, G.; Marino, A.; Chen, S.W.; Yang, J. A general framework of polarimetric detectors based on quadratic optimization. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5237418. [CrossRef]
11. Liu, T.; Yang, Z.; Gao, G.; Marino, A.; Chen, S.W. Simultaneous diagonalization of Hermitian matrices and its application in PolSAR ship detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5220818. [CrossRef]
12. Liu, S.; Huang, D.; Wang, Y. Pay attention to them: Deep reinforcement learning-based cascade object detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 2544–2556. [CrossRef]

13. Lin, Q.; Zhao, J.; Fu, G.; Yuan, Z. CRPN-SFNet: A high-performance object detector on large-scale remote sensing images. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 416–429. [CrossRef]
14. Zhou, Y.; Jiang, X.; Li, Z.; Liu, X. Arbitrary-oriented sar ship detection via frequency learning. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4552–4555.
15. Fu, K.; Fu, J.; Wang, Z.; Sun, X. Scattering-keypoint-guided network for oriented ship detection in high-resolution and large-scale SAR images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11162–11178. [CrossRef]
16. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 2849–2858.
17. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented r-cnn for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3520–3529.
18. Wang, Y.; Zhang, Z.; Xu, W.; Chen, L.; Wang, G.; Yan, L.; Zhong, S.; Zou, X. Learning oriented object detection via naive geometric computing. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 10513–10525. [CrossRef]
19. Yang, X.; Yan, J.; Feng, Z.; He, T. R3det: Refined single-stage detector with feature refinement for rotating object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 3163–3171.
20. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align deep features for oriented object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5602511. [CrossRef]
21. Huang, Z.; Li, W.; Xia, X.G.; Wang, H.; Tao, R. Task-wise sampling convolutions for arbitrary-oriented object detection in aerial images. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 5204–5218. [CrossRef] [PubMed]
22. Wang, S.; Cai, Z.; Yuan, J. Automatic SAR Ship Detection Based on Multifeature Fusion Network in Spatial and Frequency Domains. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4102111. [CrossRef]
23. Pan, D.; Gao, X.; Dai, W.; Fu, J.; Wang, Z.; Sun, X.; Wu, Y. SRT-Net: Scattering Region Topology Network for Oriented Ship Detection in Large-Scale SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5202318. [CrossRef]
24. Zhou, Y.; Jiang, X.; Xu, G.; Yang, X.; Liu, X.; Li, Z. PVT-SAR: An Arbitrarily Oriented SAR Ship Detector With Pyramid Vision Transformer. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 291–305. [CrossRef]
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
26. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31 $\times$ 31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
27. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *arXiv* **2022**, arXiv:2202.09741. [CrossRef]
28. Wang, Y.; Li, Y.; Wang, G.; Liu, X. Multi-scale attention network for single image super-resolution. *arXiv* **2022**, arXiv:2209.14145.
29. Yang, X.; Yan, J. Arbitrary-oriented object detection with circular smooth label. In *Computer Vision—ECCV 2020: Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part VIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 677–694.
30. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8232–8241.
31. Yang, X.; Zhou, Y.; Zhang, G.; Yang, J.; Wang, W.; Yan, J.; Zhang, X.; Tian, Q. The kfiou loss for rotated object detection. *arXiv* **2022**, arXiv:2201.12558.
32. Yu, Y.; Da, F. Phase-shifting coder: Predicting accurate orientation in oriented object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13354–13363.
33. Zhang, M.; Wang, N.; Li, Y.; Gao, X. Deep latent low-rank representation for face sketch synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3109–3123. [CrossRef]
34. Zhang, S.; Gao, X.; Wang, N.; Li, J.; Zhang, M. Face Sketch Synthesis via Sparse Representation-Based Greedy Search. *IEEE Trans. Image Process.* **2015**, *24*, 2466–2477. [CrossRef] [PubMed]
35. Van Den Oord, A.; Vinyals, O. Neural discrete representation learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6309–6318
36. Zhou, S.; Chan, K.; Li, C.; Loy, C.C. Towards robust blind face restoration with codebook lookup transformer. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 30599–30611.
37. Zuo, C.; Feng, S.; Huang, L.; Tao, T.; Yin, W.; Chen, Q. Phase shifting algorithms for fringe projection profilometry: A review. *Opt. Lasers Eng.* **2018**, *109*, 23–59. [CrossRef]
38. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.

39. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. Rtmdet: An empirical study of designing real-time object detectors. *arXiv* **2022**, arXiv:2212.07784.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.