information

Article

Vehicle Target Recognition in SAR Images with Complex Scenes Based on Mixed Attention Mechanism

Tao Tang [1,*], Yuting Cui [2], Rui Feng [3] and Deliang Xiang [3]

[1] College of Electronics Science and Technology, National University of Defense Technology, Changsha 410073, China
[2] Ceyear Technologies Co., Ltd., Qingdao 266555, China; cuiyuting@ceyear.com
[3] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100013, China; 2022210545@buct.edu.cn (R.F.); xiangdeliang@buct.edu.cn (D.X.)
* Correspondence: tangtao@nudt.edu.cn

Abstract: With the development of deep learning in the field of computer vision, convolutional neural network models and attention mechanisms have been widely applied in SAR image target recognition. The improvement of convolutional neural network attention in existing SAR image target recognition focuses on spatial and channel information but lacks research on the relationship and recognition mechanism between spatial and channel information. In response to this issue, this article proposes a hybrid attention module and introduces a Mixed Attention (MA) mechanism module in the MobileNetV2 network. The proposed MA mechanism fully considers the comprehensive calculation of spatial attention (SPA), channel attention (CHA), and coordinated attention (CA). It can input feature maps for comprehensive weighting to enhance the features of the regions of interest, in order to improve the recognition rate of vehicle targets in SAR images.The superiority of our algorithm was verified through experiments on the MSTAR dataset.

Keywords: mixed attention mechanism; MA-MobileNetV2; vehicle target recognition; synthetic aperture radar (SAR)

check for updates

Citation: Tang, T.; Cui, Y.; Feng, R.; Xiang, D. Vehicle Target Recognition in SAR Images with Complex Scenes Based on Mixed Attention Mechanism. *Information* **2024**, *15*, 159. https://doi.org/10.3390/info15030159

Academic Editor: Francesco Camastra

Received: 30 January 2024
Revised: 2 March 2024
Accepted: 5 March 2024
Published: 11 March 2024

1. Introduction

Many template-based and machine learning-based SAR image target recognition algorithms [1–6] have been proposed and achieved certain effectiveness. However, these traditional SAR image target recognition algorithms have the following drawbacks. (1) Insufficient feature extraction: traditional SAR image target recognition algorithms often use manually designed feature extraction methods, which can only extract local features of the image and cannot consider the global characteristics of the targets, resulting in insufficient feature extraction. (2) Feature redundancy: due to the manual feature extraction methods used in traditional SAR image target recognition algorithms, there are often a large number of redundant features, which not only reduce the recognition accuracy but also increase the computational complexity. (3) Difficult feature selection: Due to the presence of a large number of redundant features in traditional SAR image target recognition algorithms, feature selection is required to reduce the number of features. However, due to the complex interactions between features, it is often difficult to select the optimal subset of features. (4) Difficult feature combination: traditional SAR image target recognition algorithms typically use classifiers based on shallow models, which can only handle simple linear feature combinations and cannot handle complex nonlinear feature combinations, resulting in low recognition accuracy. These drawbacks pose significant challenges to the practical application of SAR image target recognition.

Convolutional neural network models can automatically extract important features of different targets; weight sharing using convolution layers, and spatial invariance using

pooling operations. By utilizing the powerful feature representation capability of convolutional neural network models, the workload in SAR image target recognition is significantly reduced. At the same time, it avoids the limitations of manually designing target features and greatly improves the recognition capability of SAR image targets [7]. The application of convolutional neural networks in SAR image target recognition has become a research hotspot [8–20].

In deep learning, models typically need to handle high-dimensional and complex input data. These inputs contain a large amount of information, but not all of it contributes to the model's output. Therefore, models need to focus more attention on the parts of the input data that are relevant to the task, in order to improve the accuracy and efficiency of the model's recognition. Attention mechanisms are methods for weighting the input data, highlighting the important features of the target. This allows the model to better focus on the parts of the image that are relevant to target recognition, thereby improving recognition accuracy. Researchers have proposed several new attention mechanism modules, which can be mainly divided into three categories: spatial attention [21], channel attention [22,23], and spatial–channel coordinated attention [24]. Meanwhile, in the field of SAR image target recognition, some researchers have also devoted themselves to applying attention mechanisms to improve the recognition performance of networks [25–30].

Zhang et al. [25] proposed an effectively lightweight attention mechanism convolutional neural network model (AM-CNN) for SAR automatic target recognition. Compared with traditional convolutional neural networks and state-of-the-art methods, this model has significant advantages in terms of performance and efficiency. Li et al. [28] proposed a fully convolutional attention block (FCAB), which can be combined with convolutional neural networks to refine important features in synthetic aperture radar (SAR) images and suppress unnecessary features, resulting in significant performance gains for SAR recognition. Wang et al. [29] proposed a non-local channel attention network for SAR image target recognition based on the GoogLeNet structure, which combines an asymmetric pyramid non-local block (APNB) and SENet. The use of SENet allows for channel dependencies based on feature fusion at different scales, improving recognition accuracy. Xu et al. [30] proposed a multi-scale capsule network with coordinate attention (CA-MCN), which deploys multiple dilation convolution layers to extract robust features and incorporates coordinate attention for target recognition at multiple scales.

However, the aforementioned methods for improving the attention part of convolutional neural networks only consider spatial information, channel information, or the coordination between spatial and channel information separately. They do not take into account the comprehensive weighting of spatial information, channel information, and the coordination between spatial and channel information. In response to this issue, a SAR image vehicle target recognition network based on a Mixed Attention (MA) mechanism, called MA-MobileNetV2, is proposed by introducing the Mixed Attention mechanism into the MobileNetV2 network [31]. The Mixed Attention mechanism can fully consider the computations of spatial attention (SPA), channel attention (CHA), and coordinate attention (CA), and weight the input feature maps complementarily to enhance the representation of features in the regions of interest. This presentation considers the computations of spatial attention, channel attention, and coordinate attention in deep neural network-based SAR image vehicle target recognition, effectively improving the accuracy of SAR image vehicle target recognition. Due to the incomplete SAR image vehicle target dataset in actual measurement scenarios, experiments were conducted on the MSTAR dataset. The results show that MA-MobileNetV2 has superior performance, with an average recognition accuracy of 99.85% for the 10 target classes. The average recognition accuracy has been improved by 3.1% compared to the unmodified MobileNetV2 network, and it also outperforms the recently reported SAR image vehicle target recognition algorithms based on attention-related improvements.

The structure of this paper is arranged as follows: Section 2 provides a detailed description of the basic structure of the MobileNetV2 model and each attention module. Section 3 presents the experimental parameter settings and analysis of the experimental results. Section 4 summarizes the work of this paper and discusses future research directions.

2. The Proposed Method

2.1. The MobileNetV2 Model

The MobileNetV2 model is a lightweight deep neural network proposed by Google. It significantly reduces the required number of computations and memory while maintaining the same level of accuracy. A key feature of MobileNetV2 is the use of inverted residuals and bottleneck residual blocks, which are composed of linear activation functions. The entire MobileNetV2 model is primarily composed of bottleneck structures. The structure of MobileNetV2 is shown in Table 1.

Table 1. Structure of MobileNetV2.

Input	Operator	Expansion	Output Channels	Operator Repeat Times	Stride
$128 \times 128 \times 3$	Conv2d	-	32	1	2
$64 \times 64 \times 32$	Bottleneck	1	16	1	1
$64 \times 64 \times 16$	Bottleneck	6	24	2	2
$32 \times 32 \times 24$	Bottleneck	6	32	3	2
$16 \times 16 \times 32$	Bottleneck	6	64	4	2
$8 \times 8 \times 64$	Bottleneck	6	96	3	1
$8 \times 8 \times 96$	Bottleneck	6	160	3	2
$4 \times 4 \times 160$	Bottleneck	6	320	1	1
$4 \times 4 \times 320$	Conv2d	-	1280	1	1
$4 \times 4 \times 1280$	Avgpool	-	-	1	-
$1 \times 1 \times 1280$	Conv2d	-	k	-	-

The structure of the bottleneck, as shown in Figure 1, is composed of three parts: the expansion convolution part, the depthwise convolution part, and the projection convolution part. The entire structure uses ReLU6 as the activation function, as depicted in Figure 2.

The expansion convolution increases the number of channels in the input feature map using a 1×1 convolutional kernel. Its purpose is to enhance the ability of the depthwise convolution to extract meaningful information. By increasing the number of channels, the model can learn more feature representations, thereby improving its expressive power. The depthwise convolution performs convolutions separately for each input channel, using fewer parameters for computation, resulting in fewer parameters compared to traditional convolutions. On the other hand, the projection convolution, which is the opposite of the expansion convolution, uses a 1×1 convolutional kernel. Its output channels are smaller than the input channels, limiting the size of the model. The purpose of the projection convolution is to ensure that the number of channels does not increase excessively, thereby reducing the number of model parameters. The ReLU6 activation function is calculated according to the following formula:

$$f(x) = ReLU(x) = \min(\max(x, 0), 6) \quad (1)$$

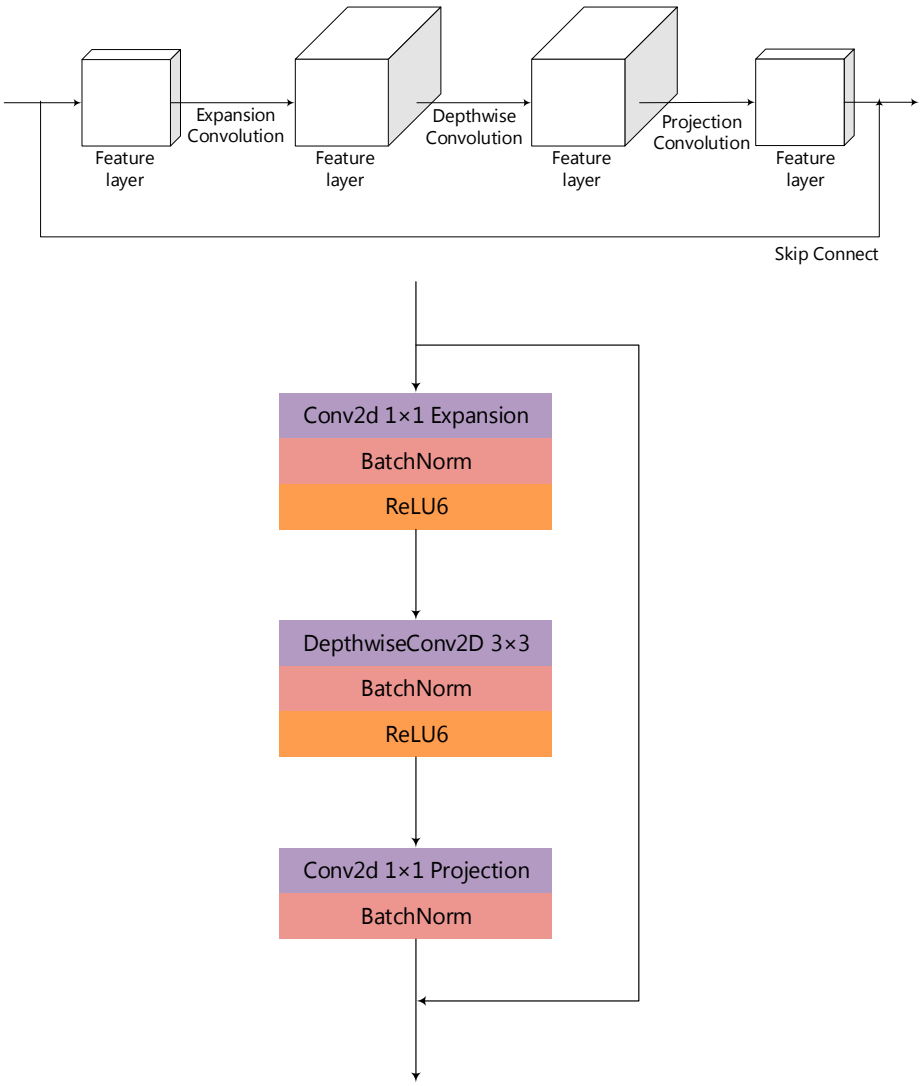


Figure 1. Structure of the bottleneck residual block.

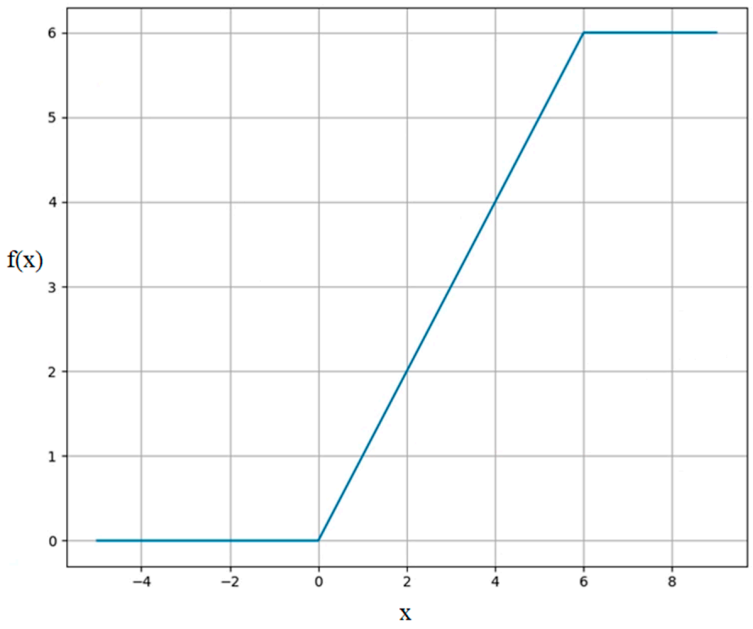


Figure 2. Figure of the ReLU6 function.

2.2. The Channel Attention Module

The channel attention mechanism can improve the recognition accuracy of the model by selecting channels that are more important for the recognition task and reducing the interference from irrelevant channels. The structure of the Channel Attention Module is shown in Figure 3. The channel attention mechanism first compresses the input feature map in the spatial dimension using average pooling and max pooling, obtaining two different vectors of size C × 1 × 1: F^c_{avg} and $F^c_{\max}$. They represent the average-pooled feature and max-pooled feature, respectively. Then, F^c_{avg} and $F^c_{\max}$ are inputted into a shared network to obtain the channel attention maps $MLP(F^c_{avg})$ and $MLP(F^c_{\max})$, where the shared network consists of a Multilayer Perceptron (MLP) with one hidden layer. After applying the shared network to F^c_{avg} and $F^c_{\max}$, the output feature vectors are merged using element-wise summation, and then passed through the sigmoid function to obtain the final channel attention weights $M_c(F)$. The calculation formula is as follows:

$$F^c_{avg} = AvgPool_c(F) \tag{2}$$

$$F^c_{\max} = MaxPool_c(F) \tag{3}$$

$$M_c(F) = sigmoid(MLP(F^c_{avg}) + MLP(F^c_{\max})) \tag{4}$$

where F represents the input feature, $M_c(F)$ represents the channel attention weights, $AvgPool_c$ represents average pooling operation along the channel dimension, and $MaxPool_c$ represents max pooling operation along the channel dimension.

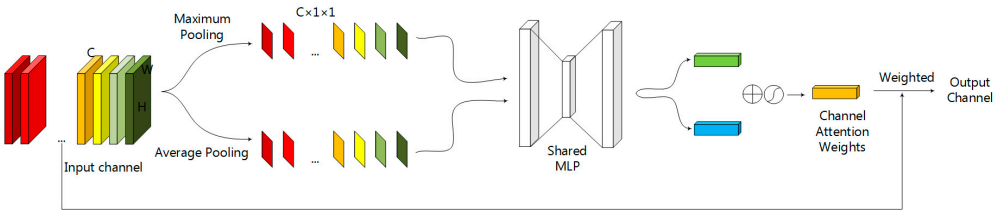


Figure 3. Structure diagram of the Channel Attention Module.

2.3. The Spatial Attention Module

The spatial attention mechanism can control the attention weights of different positions in order to focus more on important areas and locations with higher information content, thereby helping the model better capture important information in the image. The structure of the Spatial Attention Module is shown in Figure 4. To calculate spatial attention, average pooling and max pooling operations are applied along the channel axis, obtaining two different feature descriptors of size 1 × H × W: F^s_{avg} and $F^s_{\max}$. These feature descriptors are then concatenated to generate a 2 × H × W feature descriptor. The 2 × H × W feature descriptor is further convolved through a standard convolutional layer and passed through the sigmoid function to obtain the final spatial attention weights $M_s(F)$. The calculation formula is as follows:

$$F^s_{avg} = AvgPool_s(F) \tag{5}$$

$$F^s_{\max} = MaxPool_s(F) \tag{6}$$

$$M_s(F) = sigmoid(f(F^s_{avg}; F^s_{\max})) \tag{7}$$

where F represents the input feature, $M_s(F)$ represents the spatial attention weights, f represents the standard convolution operation, $AvgPool_s$ represents average pooling operation along the spatial dimension, and $MaxPool_c$ represents max pooling operation along the spatial dimension.

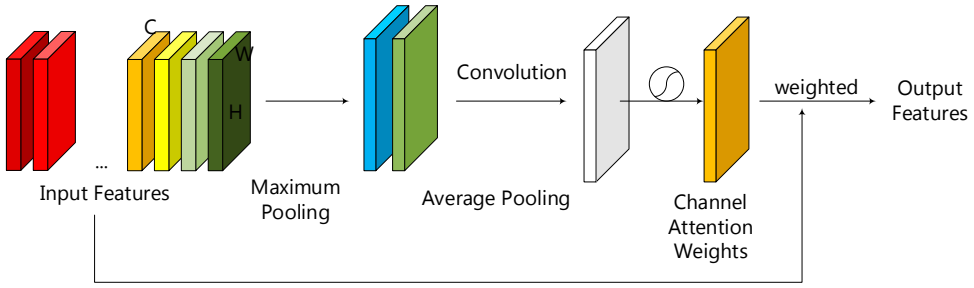


Figure 4. Structure diagram of the Spatial Attention Module.

2.4. The Spatial and Channel Coordinated Attention Module

The spatial and channel coordinated attention not only considers channel information but also takes into account position information related to the orientation. It can learn the dependency relationship between features based on the spatial location, thereby better capturing the relationship between different regions in the image. The structure of the Spatial and Channel Coordinated Attention Module is shown in Figure 5. To calculate the spatial and channel coordinated attention, average pooling is applied along the X-axis and Y-axis of the input feature map, resulting in F_{avg}^{h} of size C × 1 × W and F_{avg}^{w} of size C × H × 1. These feature maps are concatenated to generate a feature description of size C × 1 × (H + W). The C × 1 × (H + W) feature description is convolved through a standard convolutional layer and then decomposed into feature descriptions of sizes C × 1 × W and C × H × 1. The feature descriptions are then passed through sigmoid functions to obtain the final spatial attention weights $M_h(F)$ and $M_w(F)$. The calculation formula is as follows:

$$F_{avg}^{h} = AvgPool_h(F) \tag{8}$$

$$F_{avg}^{w} = AvgPool_w(F) \tag{9}$$

$$(M_h(F); M_w(F)) = sigmoid(f(M_{avg}^{h}; M_{avg}^{w})) \tag{10}$$

where F represents the input feature, $M_h(F)$ represents the attention weights along the spatial X-axis, $M_w(F)$ represents the attention weights along the spatial Y-axis, f represents the standard convolution operation, $AvgPool_h$ represents average pooling operation along the spatial X-axis, and $AvgPool_w$ represents average pooling operation along the spatial Y-axis.

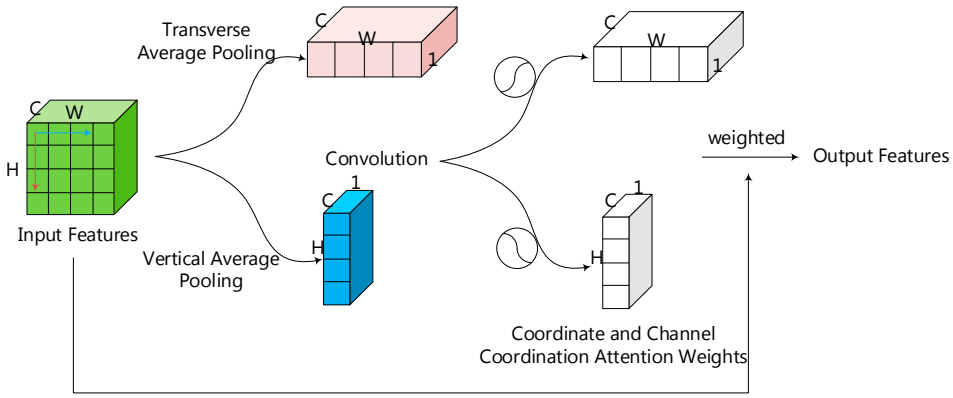


Figure 5. Structure diagram of the Spatial and Channel Coordinated Attention Module.

2.5. The Mixed Attention Convolutional Neural Network (MA-MobileNetV2)

Channel attention can automatically weight the channel dimension, spatial attention can automatically weight the spatial dimension, and the Spatial and Channel Coordinated Attention Module can weight the spatial distribution on channels. As shown in Figures 3–5, these three weighting methods have complementarity in different dimensions

and can complementarily weight the feature maps to improve the recognition accuracy of the network.

The structure of the Mixed Attention Module is shown in Figure 6. To calculate mixed attention, the input feature is weighted using channel attention (CHA), spatial attention (SPA), and coordination attention (CA), respectively. The final output feature map is obtained by applying the weighted mixed attention. The calculation of features and weights follows the broadcasting mechanism. The formula for computing the attention is as follows:

$$F' = F * M_c(F) * M_s(F) * (M_h(F) * M_w(F)) \tag{11}$$

where F represents the input feature, $M_c(F)$ represents the attention weights for channel attention, $M_s(F)$ represents the attention weights for spatial attention, $M_h(F)$ represents the attention weights along the spatial X-axis, $M_w(F)$ represents the attention weights along the spatial Y-axis, and F' represents the feature weighted by the Mixed Attention Module.

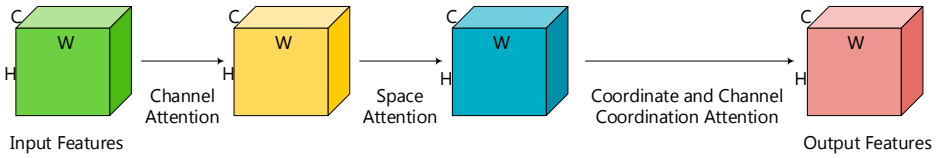


Figure 6. Structure diagram of the Mixed Attention Module.

The MA-MobileNetV2 network incorporates the mixed attention into the MobileNetV2 network, taking into account the comprehensive weighting of spatial attention, channel attention, and coordinated attention between spatial and channel dimensions on the input feature map. This enhances the representation of features in the region of interest and improves network performance. The structure of MA-MobileNetV2 is shown in Table 2, with bold sections indicating the locations where the Mixed Attention Module is introduced. The recognition process of MA-MobileNetV2 is illustrated in Figure 7. Firstly, the input image goes through the backbone network to obtain the feature map. Then, the obtained feature map is weighted using the Mixed Attention Module. The weighted feature map is then fed into pooling layers and fully connected layers to obtain the final recognition result. In the next section, the algorithm's performance will be validated through experiments.

Table 2. Structure of MA-MobileNetV2.

Input	Operator	Expansion	Output Channels	Operator Repeat Times	Stride
128 × 128 × 3	Conv2d	-	32	1	2
64 × 64 × 32	Bottleneck	1	16	1	1
64 × 64 × 16	Bottleneck	6	24	2	2
32 × 32 × 24	Bottleneck	6	32	3	2
16 × 16 × 32	Bottleneck	6	64	4	2
8 × 8 × 64	Bottleneck	6	96	3	1
8 × 8 × 96	Bottleneck	6	160	3	2
4 × 4 × 160	Bottleneck	6	320	1	1
4 × 4 × 320	Conv2d	-	1280	1	1
4 × 4 × 1280	**CHA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	**SPA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	**CA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	Avgpool	-	-	1	-
1 × 1 × 1280	Conv2d	-	k	-	-

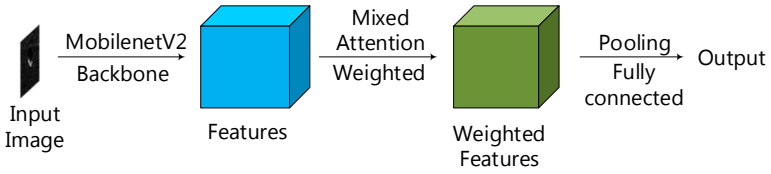


Figure 7. Recognition process diagram of MA-MobileNetV2.

3. Results and Analysis

3.1. Experimental Data and Parameter Settings

This article uses the MSTAR dataset as the training set for training and testing. The MSTAR dataset is a publicly available dataset developed jointly by the Advanced Research Projects Agency (DARPA) of the US Department of Defense and the Air Force Research Laboratory (AFRL) for synthetic aperture radar (SAR) target recognition. The slice of MSTAR data is the imaging result of X-band airborne SAR. The size of each image in the MSTAR dataset under SOC is 128 × 128, containing 10 types of targets. Figure 8 shows an example of each category in the MSTAR dataset, and Table 3 shows the partitioning of SOC in the MSTAR dataset. Each model of target in the dataset has a large number of images with different azimuth angles, ranging from 0 to 180°, with azimuth intervals of approximately 1 to 2°. In addition, there are two elevation angles available for each model, 17° and 15°. This article uses data with a pitch angle of 17° for training, and data with a pitch angle of 15° for testing, proving the superiority of the algorithm proposed in this article.

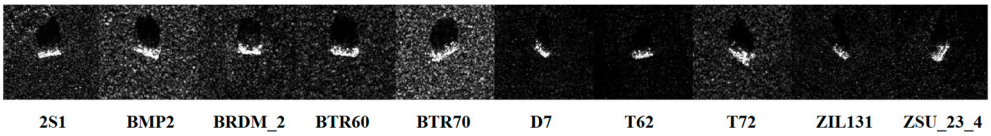


Figure 8. Sample of MSTAR Data.

Table 3. Number of ten-class military vehicles under SOC in MSTAR data.

Class	Number of Targets (17)	Number of Targets (15)	Sum. of Targets (17 and 15)
2S1	299	274	573
BMP2	233	195	428
BRDM_2	298	274	572
BTR60	256	195	451
BTR70	233	196	429
D7	299	274	573
T62	298	273	571
T72	232	196	428
ZIL131	299	274	573
ZSU_23_4	299	274	573
Sum	2746	2425	5171

The MA-MobileNetV2 network is trained and tested on the MSTAR target slice dataset introduced in Table 3 and Figure 8, using data at a 17° angle for training and data at a 15° angle for testing, to verify the recognition performance of the MA-MobileNetV2 network. Multiple sets of experiments are set up in this section to verify the performance improvement of the hybrid attention mechanism network in SAR image vehicle target recognition tasks. To ensure fairness in the experiments, all detection model training is conducted with the following settings:

1. The model parameters in the experiment are initialized using optical pre-trained recognition model parameters.
2. The features extracted by the backbone network of the model are universal. Freezing the training of the backbone network can speed up training efficiency and prevent

weight destruction. Therefore, for the first 5% of epochs in all experiments, the backbone network is frozen to adjust the parameters of the hybrid attention module and fully connected layers. At this stage, the feature extraction network remains unchanged, ensuring the stability of network training. The freezing is then lifted for the remaining 95% of epochs to adjust the overall parameters of the network. The batch size for the first 5% of epochs is set to 32, and the batch size for the remaining 95% of epochs is also set to 32.

3. The optimizer used for all models in the experiment is Stochastic Gradient Descent (SGD) optimizer. The learning rate is adjusted using the cosine annealing function. The initial learning rate is set to 0.01, and the minimum learning rate is set to 0.0001.
4. The computer configuration during the experiment is as follows: (1) CPU: AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz. (2) RAM: 16 GB. (3) GPU: NVIDIA Geforce RTX 3060 Laptop. (4) Operating System: Windows 11.

3.2. Performance Evaluation Metrics

To quantitatively compare the recognition performance of models, accuracy and recall are used as performance evaluation metrics for recognition results. The formulas for accuracy and recall calculation are as follows:

$$P = \frac{TP}{TP + FP} \tag{12}$$

$$R = \frac{TP}{NP} \tag{13}$$

where TP is the number of correctly predicted targets in the recognition results, FP represents the number of incorrectly predicted targets, NP represents the number of true targets, P is the precision (accuracy) of the recognition, and R represents the recall rate.

3.3. Performance Comparison between the MA-MobileNetV2 Network and the MobileNetV2 Network

The MA-MobileNetV2 network is an improved network that introduces the Mixed Attention (MA) Module based on the MobileNetV2 network. The training losses of the MA-MobileNetV2 network and the MobileNetV2 network on the MSTAR dataset as shown in Figure 9. The first 20 epochs represent the training process when the backbone network is frozen. During these epochs, the MA-MobileNetV2 network primarily adjusts the parameters of the Mixed Attention Module and the fully connected layer, while the MobileNetV2 network mainly adjusts the parameters of the fully connected layer.

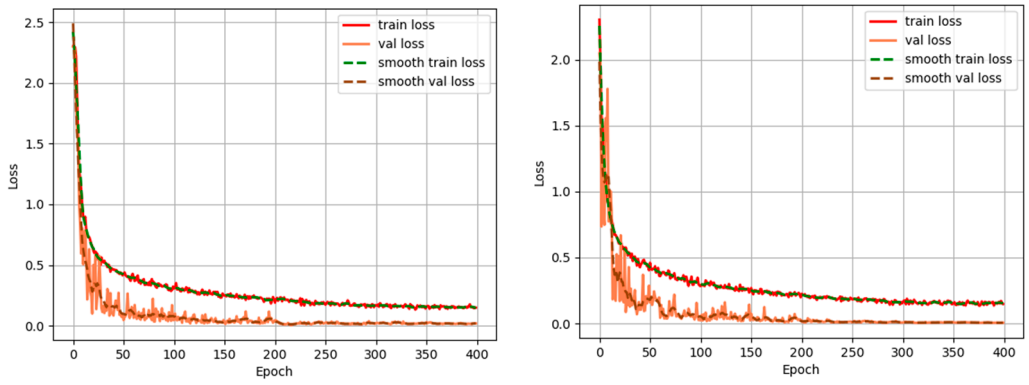


Figure 9. The training loss curves of the MA-MobileNetV2 network (**a**) and the MobileNetV2 network (**b**).

It can be observed that the MobileNetV2 network exhibits larger fluctuations in test loss during the first 20 epochs of training, whereas the MA-MobileNetV2 network shows a significant reduction in test loss fluctuations compared to the MobileNetV2 network. Adding the Mixed Attention Module to the MobileNetV2 network allows it to automatically suppress unimportant regions in the images and focus only on the relevant areas. This reduces the complexity of the model and improves its performance and stability.

Tables 4 and 5 present the recognition confusion matrices for the MA-MobileNetV2 network and the MobileNetV2 network, respectively, on the MSTAR dataset. It can be observed that compared to the MobileNetV2 network, the MA-MobileNetV2 network exhibits a significant improvement in recognition accuracy.

Table 4. The confusion matrix of the MA-MobileNetV2 network.

	2S1	BMP2	BRDM_2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4
2S1	274	0	0	0	0	0	0	0	0	0
BMP2	0	194	0	0	0	0	0	1	0	0
BRDM_2	0	0	274	0	0	0	0	0	0	0
BTR60	0	0	0	195	0	0	0	0	0	0
BTR70	0	0	0	2	194	0	0	0	0	0
D7	0	0	0	0	0	274	0	0	0	0
T62	0	0	0	0	0	0	273	0	0	0
T72	0	0	0	0	0	0	0	196	0	0
ZIL131	0	0	0	0	0	0	0	0	274	0
ZSU_23_4	0	0	0	0	0	0	0	0	0	274

Table 5. The confusion matrix of the MobileNetV2 network.

	2S1	BMP2	BRDM_2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4
2S1	252	6	0	3	3	0	0	8	2	0
BMP2	0	184	0	2	0	0	0	9	0	0
BRDM_2	0	0	265	0	0	3	0	1	4	1
BTR60	0	0	0	193	1	0	0	1	0	0
BTR70	0	2	0	6	187	0	0	1	0	0
D7	0	0	0	0	0	273	0	0	1	0
T62	0	0	0	0	0	1	267	2	0	3
T72	0	0	0	1	0	0	0	195	0	0
ZIL131	0	0	0	0	0	13	0	0	261	0
ZSU_23_4	0	0	0	0	0	1	0	0	0	273

The statistics of the class-wise recognition accuracy for the MA-MobileNetV2 network and the MobileNetV2 network on the MSTAR dataset are shown in Table 6. It can be observed that compared to the MobileNetV2 network, the MA-MobileNetV2 network demonstrates a significant improvement in the recognition accuracy of SAR image vehicle targets. The average recognition accuracy for the MA-MobileNetV2 network is 99.85%, while for the MobileNetV2 network it is 96.75%. This represents a 3.1% increase in average recognition accuracy for the MA-MobileNetV2 network over the MobileNetV2 network.

Table 6. The recognition and recall of MA-MobileNetV2 network and MobileNetV2 network.

	MA-MobileNetV2 Network	MobileNetV2 Network
Recognition accuracy	99.85%	96.75%
Recall	99.85%	96.92%

The statistics of the class-wise recall rate for the MA-MobileNetV2 network and the MobileNetV2 network on the MSTAR dataset are else shown in Table 6. Similar to the

recognition accuracy, the MA-MobileNetV2 network exhibits a notable improvement in the recall rate for SAR image vehicle targets compared to the MobileNetV2 network. The average recall rate for the MA-MobileNetV2 network is 99.85%, while for the MobileNetV2 network it is 96.92%. This represents a 2.93% increase in average recall rate for the MA-MobileNetV2 network over the MobileNetV2 network.

3.4. Performance Comparison between the MA-MobileNetV2 Network and State-of-the-Art Algorithms

The performance comparison of the MA-MobileNetV2 network with the latest attention-related improved SAR image vehicle target recognition algorithms is shown in Table 7. It can be concluded that the MA-MobileNetV2 network exhibits a significant improvement in average recognition accuracy compared to the latest attention-related improved SAR image vehicle target recognition algorithms. This demonstrates the superior performance of the MA-MobileNetV2 network in target recognition.

Table 7. The performance comparison between MA-MobileNetV2 and the latest attention-related improved recognition algorithms.

Accuracy	The Proposed Method	AM-CNN [25]	FCAB-CNN [28]	GoogleNet-APNB-ISEB [29]	CA-MCN [30]
2S1	100%	98.90%	100%	99.80%	99.27%
BMP2	100%	100%	98.46%	99.50%	100%
BRDM_2	100%	99.64%	99.27%	99.80%	99.64%
BTR60	98.98%	96.41%	98.98%	99.20%	99.49%
BTR70	100%	100%	100%	100%	100%
D7	100%	99.27%	100%	99.30%	99.27%
T62	100%	99.63%	99.27%	100%	99.63%
T72	99.49%	100%	100%	99.80%	100%
ZIL131	100%	99.64%	100%	99.80%	99.64%
ZSU_23_4	100%	100%	98.17	99.40%	100%
Average accuracy	**99.85%**	99.35%	99.51%	99.72%	99.59%

4. Discussion

In order to demonstrate the roles of the CHA module, SPA module, and CA module in the hybrid attention mechanism, this section conducts ablation experiments on the hybrid attention mechanism network to prove the necessity of the CHA module, SPA module, and CA module.

4.1. Ablation Experiments on CHA Module

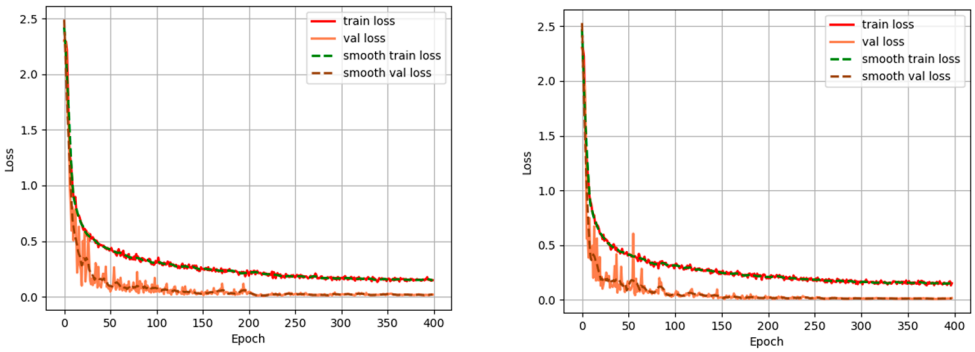
In this section, training was conducted on a modified version of the MobileNetV2 network with only the SPA module and CA module added. The network structure is shown in Table 8. The recognition performance of the perturbed network was compared and analyzed against the recognition performance of the hybrid attention network.

The training loss of the MA-MobileNetV2 network and the perturbed CHA module network on the MSTAR dataset is shown in Figure 10. It can be observed that the testing loss of the perturbed CHA module network exhibits more pronounced fluctuations compared to the MA-MobileNetV2 network. This suggests that the CHA module plays a certain role in improving the performance and stability of the model.

Tables 8 and 9 present the confusion matrices of the perturbed CHA module network and the MA-MobileNetV2 network on the MSTAR dataset, respectively. It can be observed that compared to the MA-MobileNetV2 network, the misclassification rate of the perturbed CHA module network significantly increases.

Table 8. Ablation network structure of the CHA module.

Input	Operator	Expansion	Output Channels	Operator Repeat Times	Stride
128 × 128 × 3	Conv2d	-	32	1	2
64 × 64 × 32	Bottleneck	1	16	1	1
64 × 64 × 16	Bottleneck	6	24	2	2
32 × 32 × 24	Bottleneck	6	32	3	2
16 × 16 × 32	Bottleneck	6	64	4	2
8 × 8 × 64	Bottleneck	6	96	3	1
8 × 8 × 96	Bottleneck	6	160	3	2
4 × 4 × 160	Bottleneck	6	320	1	1
4 × 4 × 320	Conv2d	-	1280	1	1
4 × 4 × 1280	**SPA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	**CA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	Avgpool	-	-	1	-
1 × 1 × 1280	Conv2d	-	k	-	-



(**a**) (**b**)

Figure 10. The training loss curves of (**a**) MA-MobileNetV2 network and (**b**) CHA module network.

Table 9. The confusion matrix of the ablation CHA module network.

	2S1	BMP2	BRDM_2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4
2S1	271	0	0	1	2	0	0	0	0	0
BMP2	0	191	0	1	2	0	0	1	0	0
BRDM_2	0	0	274	0	0	0	0	0	0	0
BTR60	0	0	0	195	0	0	0	0	0	0
BTR70	0	0	0	1	195	0	0	0	0	0
D7	0	0	0	0	0	273	0	0	1	0
T62	0	0	0	0	0	0	273	0	0	0
T72	0	0	0	0	0	0	0	196	0	0
ZIL131	0	0	0	0	0	0	0	0	274	0
ZSU_23_4	0	0	0	0	0	0	0	0	0	274

The statistical accuracies of various target classes in SAR vehicle recognition for the MA-MobileNetV2 network and the perturbed CHA module network are shown in Table 10. It can be observed that compared to the MA-MobileNetV2 network, the perturbed CHA module network exhibits a noticeable decrease in recognition accuracy for SAR vehicle targets. The average recognition accuracy of the MA-MobileNetV2 network is 99.85%, while that of the perturbed CHA module network is 99.56%, indicating a decrease of 0.29 percentage points in average recognition accuracy for the perturbed CHA module network compared to the MA-MobileNetV2 network.

Table 10. The recognition and recall of MA-MobileNetV2 network and the ablation CHA module network.

	MA-MobileNetV2 Network	CHA Module Network
Recognition accuracy	99.85%	99.56%
Recall	99.85%	99.60%

The recall rates of various target classes in SAR vehicle recognition for the MA-MobileNetV2 network and the perturbed CHA module network are also shown in Table 10. It can be observed that compared to the MA-MobileNetV2 network, the perturbed CHA module network exhibits a noticeable decrease in recall rate for SAR vehicle targets. The average recall rate of the MA-MobileNetV2 network is 99.85%, while that of the perturbed CHA module network is 99.60%, indicating a decrease of 0.25 percentage points in average recall rate for the perturbed CHA module network compared to the MA-MobileNetV2 network.

In conclusion, it can be inferred that the CHA module plays a significant role in improving the recognition accuracy, recall rate, and stability of the model.

4.2. Ablation Experiments on SPA Module

In this section, training was conducted on the MobileNetV2 network with only the CHA module and CA module added. The network structure is shown in Table 11. A comparative analysis was performed between the recognition performance of the ablation networks and the recognition performance of the hybrid attention network.

Table 11. Ablation network structure of the SPA module.

Input	Operator	Expansion	Output Channels	Operator Repeat Times	Stride
$128 \times 128 \times 3$	Conv2d	-	32	1	2
$64 \times 64 \times 32$	Bottleneck	1	16	1	1
$64 \times 64 \times 16$	Bottleneck	6	24	2	2
$32 \times 32 \times 24$	Bottleneck	6	32	3	2
$16 \times 16 \times 32$	Bottleneck	6	64	4	2
$8 \times 8 \times 64$	Bottleneck	6	96	3	1
$8 \times 8 \times 96$	Bottleneck	6	160	3	2
$4 \times 4 \times 160$	Bottleneck	6	320	1	1
$4 \times 4 \times 320$	Conv2d	-	1280	1	1
$4 \times 4 \times 1280$	**CHA**	**-**	**1280**	**1**	**-**
$4 \times 4 \times 1280$	**CA**	**-**	**1280**	**1**	**-**
$4 \times 4 \times 1280$	Avgpool	-	-	1	-
$1 \times 1 \times 1280$	Conv2d	-	k	-	-

The training losses of the MA-MobileNetV2 network and SPA module network on the MSTAR dataset are shown in Figure 11. The test loss of the SPA module network fluctuates slightly more than the MA-MobileNetV2 network, indicating that the SPA module has a certain effect on improving model stability.

Tables 5 and 12 present the recognition confusion matrices of the MA-MobileNetV2 network and the SPA module network on the MSTAR dataset. It can be observed that compared to the MA-MobileNetV2 network, the misclassification rate of the SPA module network significantly increases. The class-wise recognition accuracy statistics for the MA-MobileNetV2 network and SPA module network on the MSTAR dataset are shown in Table 13. It can be seen that the SPA module network exhibits a noticeable decrease in recognition accuracy for SAR image vehicle targets compared to the MA-MobileNetV2 network. The average recognition accuracy is 99.85% for the MA-MobileNetV2 network and 99.49% for the SPA module network, indicating a 0.36% decrease in average recognition accuracy for the SPA module network compared to the MA-MobileNetV2 network.

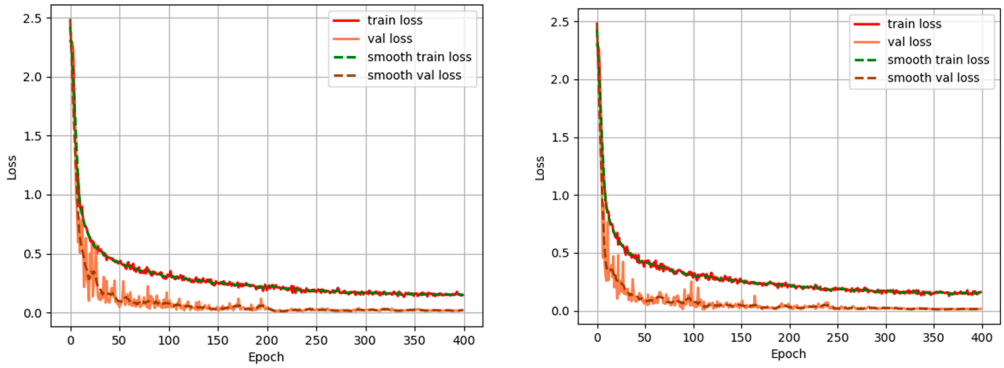


Figure 11. The training loss curves of the (**a**) MA-MobileNetV2 network and (**b**) SPA module network.

Table 12. The confusion matrix of the ablation SPA module network.

	2S1	BMP2	BRDM_2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4
2S1	268	0	0	5	0	0	0	0	1	0
BMP2	0	194	0	1	0	0	0	0	0	0
BRDM_2	0	0	273	0	0	1	0	0	0	0
BTR60	0	0	0	195	0	0	0	0	0	0
BTR70	0	0	0	3	193	0	0	0	0	0
D7	0	0	0	0	0	274	0	0	0	0
T62	0	0	0	0	0	0	273	0	0	0
T72	0	0	0	0	0	0	0	196	0	0
ZIL131	0	0	0	0	0	0	0	0	274	0
ZSU_23_4	0	0	0	0	0	0	0	0	0	274

Table 13. The recognition and recall of the MA-MobileNetV2 network and the ablation SPA module network.

	MA-MobileNetV2 Network	SPA Module Network
Recognition accuracy	99.85%	99.49%
Recall	99.85%	99.54%

The class-wise recognition recall statistics for the MA-MobileNetV2 network and SPA module network on the MSTAR dataset are also depicted in Table 13. It can be observed that the SPA module network shows a significant decrease in recall rate for SAR image vehicle targets compared to the MA-MobileNetV2 network. The average recognition recall rate is 99.85% for the MA-MobileNetV2 network and 99.54% for the SPA module network, indicating a 0.31% decrease in average recognition recall rate for the SPA module network compared to the MA-MobileNetV2 network.

In conclusion, it can be inferred that the SPA module has a significant impact on improving the recognition accuracy and recall rate of the model, as well as some effect on enhancing model stability.

4.3. Ablation Experiments on CA Module

In this section, training was conducted on the MobileNetV2 network with only the CHA module and SPA module added. The network structure is shown in Table 14. The recognition performance of the ablation networks was compared and analyzed with that of the hybrid attention network.

Table 14. Ablation network structure of the CA module.

Input	Operator	Expansion	Output Channels	Operator Repeat Times	Stride
128 × 128 × 3	Conv2d	-	32	1	2
64 × 64 × 32	Bottleneck	1	16	1	1
64 × 64 × 16	Bottleneck	6	24	2	2
32 × 32 × 24	Bottleneck	6	32	3	2
16 × 16 × 32	Bottleneck	6	64	4	2
8 × 8 × 64	Bottleneck	6	96	3	1
8 × 8 × 96	Bottleneck	6	160	3	2
4 × 4 × 160	Bottleneck	6	320	1	1
4 × 4 × 320	Conv2d	-	1280	1	1
4 × 4 × 1280	**CHA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	**SPA**	**-**	**1280**	**1**	**-**
4 × 4 × 1280	Avgpool	-	-	1	-
1 × 1 × 1280	Conv2d	-	k	-	-

The training losses of the MA-MobileNetV2 network and CA module network on the MSTAR dataset are shown in Figure 12. The test loss of the CA module network fluctuates slightly more than the MA-MobileNetV2 network, indicating that the CA module plays a certain role in improving model stability.

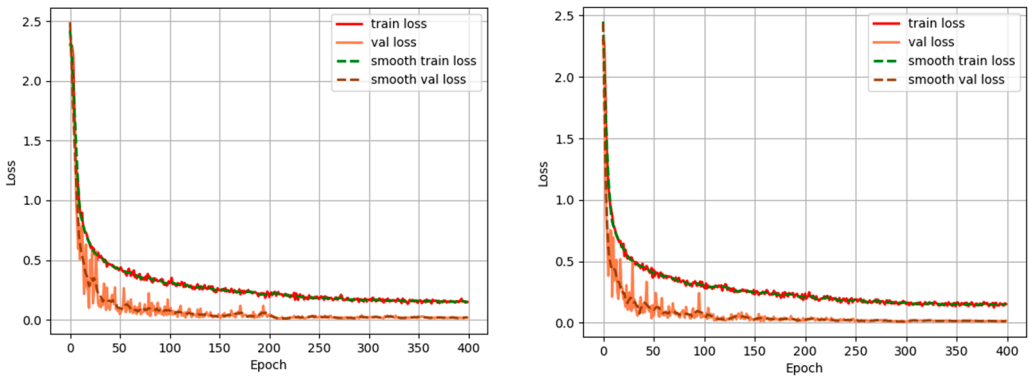


Figure 12. The training loss curves of the (**a**) MA-MobileNetV2 network and (**b**) CA module network.

Table 15 presents the recognition confusion matrices of the CA module network and the MA-MobileNetV2 network on the MSTAR dataset. It can be observed that compared to the MA-MobileNetV2 network, the misclassification rate of the CA module network significantly increases.

Table 15. The confusion matrix of the ablation CA module network.

	2S1	BMP2	BRDM_2	BTR60	BTR70	D7	T62	T72	ZIL131	ZSU_23_4
2S1	272	0	0	1	0	0	0	1	0	0
BMP2	0	195	0	0	0	0	0	0	0	0
BRDM_2	0	0	274	0	0	0	0	0	0	0
BTR60	0	1	0	194	0	0	0	0	0	0
BTR70	0	0	0	1	195	0	0	0	0	0
D7	0	0	0	0	0	271	0	0	3	0
T62	0	0	0	0	0	0	269	0	0	4
T72	0	2	0	0	0	0	0	194	0	0
ZIL131	0	0	0	0	0	0	0	0	274	0
ZSU_23_4	0	0	0	0	0	0	0	0	0	274

The class-wise recognition accuracy statistics for the MA-MobileNetV2 network and CA module network on the MSTAR dataset are shown in Table 16. It can be observed that compared to the MA-MobileNetV2 network, the CA module network exhibits a noticeable decrease in recognition accuracy for SAR image vehicle targets. The average recognition accuracy is 99.85% for the MA-MobileNetV2 network and 99.44% for the CA module network, indicating a decrease of 0.41 percentage points in average recognition accuracy for the CA module network compared to the MA-MobileNetV2 network.

Table 16. The recognition and recall of MA-MobileNetV2 network and CA module network.

	MA-MobileNetV2 Network	CA Module Network
Recognition accuracy	99.85%	99.44%
Recall	99.85%	99.47%

The class-wise recognition recall statistics for the MA-MobileNetV2 network and CA module network on the MSTAR dataset are depicted in Table 16. It can be observed that the CA module network shows a significant decrease in recall rate for SAR image vehicle targets compared to the MA-MobileNetV2 network. The average recognition recall rate is 99.85% for the MA-MobileNetV2 network and 99.47% for the CA module network, indicating a decrease of 0.38 percentage points in average recognition recall rate for the CA module network compared to the MA-MobileNetV2 network.

In conclusion, it can be inferred that the CA module has a significant impact on improving the recognition accuracy and recall rate of the model, as well as some effect on enhancing model stability.

5. Conclusions

This paper proposes a Hybrid Attention Mechanism Module and applies it to improve the MobileNetV2 network. The hybrid attention mechanism comprehensively considers spatial attention, channel attention, and coordinated attention between spatial and channel dimensions. It can effectively and complementarily weight the input feature maps to enhance the representation of features in the regions of interest, thus improving the accuracy of vehicle target recognition in SAR images by the MobileNetV2 network.

Experiments are conducted on the MSTAR dataset to validate the superiority of the proposed algorithm. The results show that the recognition accuracy of the MA-MobileNetV2 algorithm is significantly improved compared to the original MobileNetV2 network and the latest attention-based SAR image vehicle target recognition algorithms. Additionally, ablation experiments are conducted to verify the necessity of applying spatial attention, channel attention, and coordinated attention between spatial and channel dimensions in the proposed module.

At present, the deep learning-based vehicle target detection algorithm in SAR image has great advantages in the case of sufficient computing resources, but it is still difficult to deploy it on some devices that need edge computing. The next step is to study the lightweight implementation of the network while ensuring recognition accuracy, reducing hardware dependencies, and improving algorithm efficiency to achieve engineering applications.

Author Contributions: Conceptualization, Y.C. and T.T.; methodology, Y.C.; software, Y.C.; validation, T.T. and Y.C.; resources, T.T.; writing—original draft preparation, T.T. and Y.C.; writing—review and editing, T.T., R.F. and D.X.; supervision, D.X.; project administration, T.T.; funding acquisition, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Hunan province, China under Projects 2021JJ30780.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please contact author for data requests.

Conflicts of Interest: Author Yuting Cui was employed by the Ceyear Technologies Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Dong, G.; Kuang, G.; Wang, N.; Zhao, L.; Lu, J. SAR Target Recognition via Joint Sparse Representation of Monogenic Signal. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3316–3328. [CrossRef]
2. Huang, X.; Qiao, H.; Zhang, B. SAR Target Configuration Recognition Using Tensor Global and Local Discriminant Embedding. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 222–226. [CrossRef]
3. El-Darymli, K.; Gill, E.W.; Mcguire, P.; Power, D.; Moloney, C. Automatic Target Recognition in Synthetic Aperture Radar Imagery: A State-of-the-Art Review. *IEEE Access* **2016**, *4*, 6014–6058. [CrossRef]
4. Yang, N.; Zhang, Y. A Gaussian Process Classification and Target Recognition Algorithm for SAR Images. *Sci. Program.* **2022**, *2022*, 9212856. [CrossRef]
5. Ding, B. Model-driven Automatic Target Recognition of SAR Images with Part-level Reasoning. *Optik* **2022**, *252*, 168561. [CrossRef]
6. Hu, J. Automatic Target Recognition of SAR Images Using Collaborative Representation. *Comput. Intell. Neurosci.* **2022**, *2022*, 3100028. [CrossRef]
7. Du, L.; Wang, Z.; Wang, Y.; Di, W.; Lu, L.I. Survey of research progress on target detection and discrimination of single-channel SAR images for complex scenes. *J. Radars* **2020**, *9*, 34–54. [CrossRef]
8. Xu, F.; Wang, H.; Jin, Y. Deep Learning as Applied in SAR Target Recognition and Terrain Classification. *J. Radars* **2017**, *6*, 136–148. [CrossRef]
9. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional Neural Network With Data Augmentation for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 364–368. [CrossRef]
10. Chen, S.; Wang, H.; Xu, F.; Jin, Y.-Q. Target Classification Using the Deep Convolutional Networks for SAR Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4806–4817. [CrossRef]
11. Shao, J.; Qu, C.; Li, J.; Peng, S. A Lightweight Convolutional Neural Network Based on Visual Attention for SAR Image Target Classification. *Sensors* **2018**, *18*, 3039. [CrossRef]
12. Chen, H.; Zhang, F.; Tang, B.; Yin, Q.; Sun, X. Slim and Efficient Neural Network Design for Resource-constrained SAR Target Recognition. *Remote Sens.* **2018**, *10*, 1618. [CrossRef]
13. Min, R.; Lan, H.; Cao, Z.; Cui, Z. A Gradually Distilled CNN for SAR Target Recognition. *IEEE Access* **2019**, *7*, 42190–42200. [CrossRef]
14. Zhang, F.; Liu, Y.; Zhou, Y.; Yin, Q.; Li, H.-C. A lossless lightweight CNN design for SAR target recognition. *Remote Sens. Lett.* **2020**, *11*, 485–494. [CrossRef]
15. Pei, J.; Huang, Y.; Sun, Z.; Zhang, Y.; Yang, J.; Yeo, T.-S. Multiview Synthetic Aperture Radar Automatic Target Recognition Optimization: Modeling and Implementation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6425–6439. [CrossRef]
16. Zhao, P.; Liu, K.; Zou, H.; Zhen, X. Multi-Stream Convolutional Neural Network for SAR Automatic Target Recognition. *Remote Sens.* **2018**, *10*, 1473. [CrossRef]
17. Wang, N.; Wang, Y.; Liu, H.; Zuo, Q.; He, J. Feature-Fused SAR Target Discrimination Using Multiple Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1695–1699. [CrossRef]
18. Cho, J.H.; Park, C.G. Multiple Feature Aggregation Using Convolutional Neural Networks for SAR Image-Based Automatic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1882–1886. [CrossRef]
19. Tian, Z.; Wang, L.; Zhan, R.; Hu, J.; Zhang, J. Classification via weighted kernel CNN: Application to SAR target recognition. *Int. J. Remote Sens.* **2018**, *39*, 9249–9268. [CrossRef]
20. Kwak, Y.; Song, W.J.; Kim, S.E. Speckle-Noise-Invariant Convolutional Neural Network for SAR Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 549–553. [CrossRef]
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
22. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the Computer Vision—ECCV, Munich, Germany, 8–14 September 2018; pp. 3–19.
24. 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.
25. Zhang, M.; An, J.; Yang, L.D.; Wu, L.; Lu, X.Q. Convolutional Neural Network with Attention Mechanism for SAR Automatic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 1–5.
26. Wang, D.; Song, Y.; Huang, J.; An, D.; Chen, L. SAR Target Classification Based on Multiscale Attention Super-Class Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9004–9019. [CrossRef]

27. Lang, P.; Fu, X.; Feng, C.; Dong, J.; Qin, R.; Martorella, M. LW-CMDANet: A Novel Attention Network for SAR Automatic Target Recognition. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 6615–6630. [CrossRef]
28. Li, R.; Wang, X.; Wang, J.; Song, Y.; Lei, L. SAR Target Recognition Based on Efficient Fully Convolutional Attention Block CNN. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
29. Wang, Z.; Xin, Z.; Liao, G.; Huang, P.; Xuan, J.; Sun, Y.; Tai, Y. Land-Sea Target Detection and Recognition in SAR Image Based on Non-Local Channel Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
30. Xu, H.; Xu, F. Multi-Scale Capsule Network with Coordinate Attention for SAR Automatic Target Recognition. In Proceedings of the 2021 7th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR), Virtual Conference, 1–3 November 2021; pp. 1–5.
31. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.