remote sensing

MDPI

Article

An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation

Xiangkai Xu, Zhejun Feng, Changqing Cao *, Mengyuan Li, Jin Wu, Zengyan Wu, Yajie Shang and Shubing Ye

School of Physics and Optoelectronic Engineering, Xidian University, 2 South TaiBai Road, Xi'an 710071, China; xkxu@stu.xidian.edu.cn (X.X.); zhjfeng@mail.xidian.edu.cn (Z.F.); myli151024@stu.xidian.edu.cn (M.L.); jinw9824@stu.xidian.edu.cn (J.W.); zywu_21@stu.xidian.edu.cn (Z.W.); 20051212174@stu.xidian.edu.cn (Y.S.); sbye@stu.xidian.edu.cn (S.Y.)

* Correspondence: chqcao@mail.xidian.edu.cn

Abstract: Remote sensing image object detection and instance segmentation are widely valued research fields. A convolutional neural network (CNN) has shown defects in the object detection of remote sensing images. In recent years, the number of studies on transformer-based models increased, and these studies achieved good results. However, transformers still suffer from poor small object detection and unsatisfactory edge detail segmentation. In order to solve these problems, we improved the Swin transformer based on the advantages of transformers and CNNs, and designed a local perception Swin transformer (LPSW) backbone to enhance the local perception of the network and to improve the detection accuracy of small-scale objects. We also designed a spatial attention interleaved execution cascade (SAIEC) network framework, which helped to strengthen the segmentation accuracy of the network. Due to the lack of remote sensing mask datasets, the MRS-1800 remote sensing mask dataset was created. Finally, we combined the proposed backbone with the new network framework and conducted experiments on this MRS-1800 dataset. Compared with the Swin transformer, the proposed model improved the mask AP by 1.7%, mask AP_S by 3.6%, AP by 1.1% and AP_S by 4.6%, demonstrating its effectiveness and feasibility.

Keywords: instance segmentation; object detection; Swin transformer; remote sensing image; cascade mask R-CNN

Citation: Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation. *Remote Sens.* **2021**, *13*, 4779. https://doi.org/10.3390/rs13234779

Academic Editors: Fahimeh Farahnakian, Jukka Heikkonen and Pouya Jafarzadeh

Received: 19 October 2021
Accepted: 22 November 2021
Published: 25 November 2021

1. Introduction

With the continuous advancement of science and technology, remote sensing technology is eagerly developing. The feature information contained in remote sensing images has become more abundant, and a large amount of valuable information can be extracted from it and used for scientific and technological research. Machine learning based on probability and statistics usually requires complex feature description and suffers from obvious deficiencies when dealing with complex object detection and segmentation problems [1,2]. The deep structure and feature learning capabilities of deep learning achieved great success in the field of image processing, and a large number of scholars also applied it to the field of remote sensing object detection and instance segmentation [3,4]. Remote sensing image object detection and segmentation tasks have an important research significance and value for the development of aviation and remote sensing fields, and have broad application prospects in many practical scenarios, such as marine monitoring, ship management and control, and ground urban planning. In urban planning, the extraction of relevant urban metrics is important for characterizing urban typologies, and image segmentation based on deep learning is optimal for the extraction of road features in marginal areas located in urban environments [5].

Instance segmentation has become an important, complex and challenging field in machine vision research. Instance segmentation can be defined as a technology that simultaneously solves the problem of object detection and semantic segmentation. As with

semantic segmentation, it not only has the characteristics of pixel level classification, but also has the characteristics of object detection, where different instances must be located, even if they are of the same type. Figure 1 shows the differences and relationships among object detection, semantic segmentation and instance segmentation.

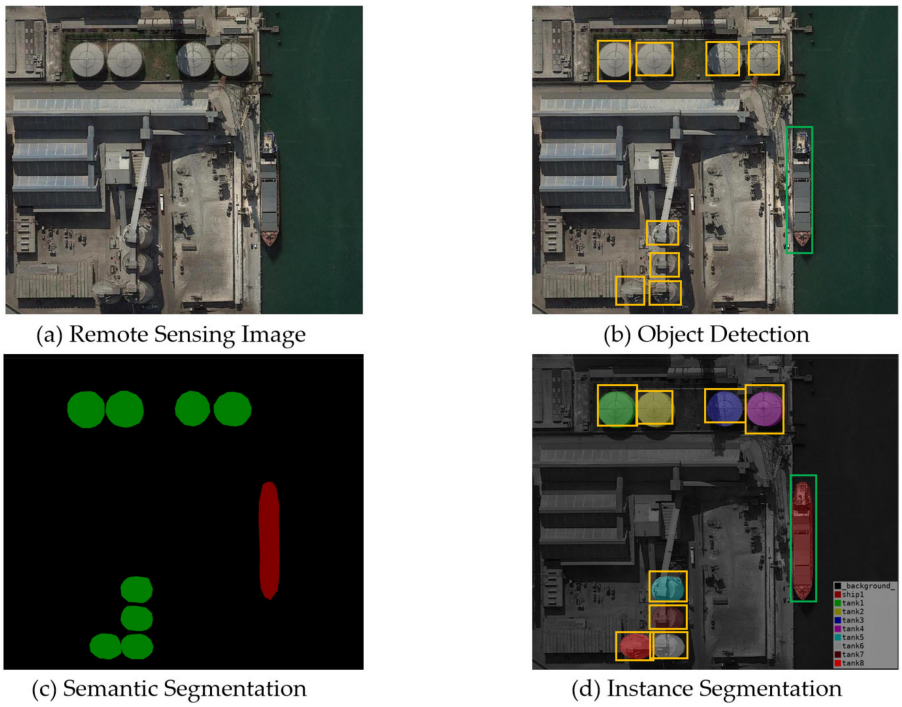

Figure 1. Examples of remote sensing image (**a**), object detection (**b**), semantic segmentation (**c**), and instance segmentation (**d**).

Since the emergence of the two-stage object detection algorithm, various object detection and segmentation algorithms based on convolutional neural networks (CNNs) have emerged, such as the region-based CNN (R-CNN), Faster R-CNN [6], and Mask R-CNN [7]. In recent years, although there are many excellent algorithms, such as the path aggregation network (PANet) [8], Mask Score R-CNN [9], Cascade Mask R-CNN [10] and segmenting objects by locations (SOLO) [11], typical problems remain, such as inaccurate segmentation edges and the establishment of global relations. If the long-range dependencies are captured by dilated convolution or by increasing the number of channels, dimensional disasters will occur due to the expansion of the model.

CNNs are useful for extracting local effective information, but they lack the ability to extract long-range features from global information. Inspired by the use of self-attention in the transformer [12] and in order to mine long-range correlation dependencies in text, many computer vision tasks propose the use of self-attention mechanisms to effectively overcome the limitations of CNNs. Self-attention mechanisms can obtain relationships between long-range elements faster and attend over different regions of the image and integrate information across the entire image. Vision transformer (ViT) [13] is a representative state-of-the-art (SOTA) work in the field of image recognition. It only uses a self-attention mechanism, which makes the image recognition rate far higher than models based on CNNs. End-to-end object detection with transformers (DETR) [14] first involved the use of transformers in high-level vision. This adds positional information to supplement image features and inputs them in the transformer structure to obtain the predicted class label and bounding box. Although transformer-based algorithms have greatly improved the object detection effect, there are still serious problems in the CV field:

1. Low detection performance for small-scale objects, and weak local information acquisition capabilities.

2. The current transformer-based framework is mostly used for image classification, but it is difficult for a single-level transformer to produce good results for the instance segmentation of densely predicted scenes. This has a great impact on object detection and instance segmentation in remote sensing images with a high resolution, a complex background, and small objects.

In order to solve these problems, there are a few works applying ViT models to the dense vision tasks of object detection and semantic segmentation via direct upsampling or deconvolution but with a relatively lower performance [15,16]. Wang et al. [17] proposed a backbone transformer for dense prediction, named "Pyramid Vision Transformer (PVT)", which designed a shrinking pyramid scheme to reduce the traditional transformer's sequence length. However, its calculation complexity is too large, which is quadratic to image size. Therefore, we chose the Swin transformer [18] as the prototype for our design of the backbone network. The Swin transformer builds a hierarchical transformer and performs self-attention calculations in the window area without overlap. The computational complexity is greatly reduced, and it is linearly related to the size of the input image. As a general-purpose visual backbone network, the Swin transformer achieves SOTA performance in tasks such as image classification, object detection, and semantic segmentation. However, the impact of the Swin transformer on context information encoding is limited; it needs to be improved for remote sensing image tasks.

In this paper, we first designed a local perception block and inserted it into each stage. Through the characteristics of dilated convolution, the block extracts a large range of local information from the image, and strengthens the network's learning of local correlation and structural information. We call the improved backbone network the "Local Perception Swin Transformer" (LPSW for short). Secondly, in order to enhance the object detection and instance segmentation of remote sensing images, inspired by the hybrid task cascade (HTC) [19], we designed the spatial attention interleaved execution cascade (SAIEC) network framework. We applied the ideas of the interleaved execution and mask information flow into Cascade Mask R-CNN. Both bounding box regression and mask prediction were combined in a multi-tasking manner. We also added an improved spatial attention module to the mask head, which helps the mask branch to focus on meaningful pixels and suppress meaningless pixels. Finally, we combined the designed LPSW backbone network with the SAIEC framework to form a new network model that achieves a higher accuracy in remote sensing object detection and instance segmentation tasks.

The main contributions of this paper can be summarized as follows:

1. In order to overcome the shortcomings of CNNs' poor ability to extract global information, we chose the Swin transformer as a basic backbone network to build a network model for remote sensing image object detection and instance segmentation.
2. According to the characteristics of remote sensing images, we propose a local perception Swin transformer (LPSW) backbone network. The LPSW combines the advantages of CNNs and transformers to enhance local perception capabilities and improve the detection accuracy of small-scale objects.
3. The spatial attention interleaved execution cascade (SAIEC) network framework is proposed. The mask prediction of the network is enhanced through the multi-tasking manner and the improved spatial attention module. Finally, the LPSW is inserted into the designed network framework as the backbone to establish a new network model that further improves the accuracy of model detection and segmentation.
4. Based on the shortage of existing remote sensing instance segmentation datasets, we selected a total of 1800 multi-object types of images from existing public datasets for annotation and created the MRS-1800 remote sensing mask dataset as the experimental resource for this paper.

2. Related Works

In this section, we introduce some previous works related to object detection and instance segmentation. For comparative analysis, we divide the content into CNN-based and transformer-based object detection and segmentation-related network models.

2.1. CNN-Based Object Detection and Instance Segmentation

In recent years, CNN-based object detection models have developed rapidly. The current object detection algorithms based on deep learning can be divided into two-stage object detection algorithms and single-stage object detection algorithms. Two-stage object detection is mainly represented by a series of regional convolutional neural network (Region-CNN, R-CNN) algorithms: the spatial pyramid pooling network (SPP-Net) [20] solves the problem of redundant operations; Fast R-CNN [21] based on R-CNN and SPP-Net proposes the concept of an region of interest (ROI) pooling layer, which can map the feature maps of different sizes of candidate regions to fixed-size feature maps; Faster R-CNN [6] uses the CNN-based region proposal network (RPN) to replace the selective search algorithm. The RPN can take an image feature map as an input, and then output a series of candidate regions. The single-stage object detection algorithm directly uses a single network to predict the category and location of the object of interest, mainly represented by the you only look once (YOLO) [22] series of algorithms. The single-shot multibox detector (SSD) [23] uses multiple-scale feature maps to perform detection tasks. On the basis of a feature pyramid network (FPN) [24], Tsung-Yi Lin et al. proposed Retinanet [25], which further improved the performance of the single-stage object detection algorithm.

At present, CNN-based instance segmentation algorithms can be divided into two main types: The top-down method and the bottom-up method. Compared with the top-down instance segmentation algorithm, the bottom-up algorithm usually has lower accuracy and more computation, such as the Proposal-Free [26] network.

The top-down method is based on the object detection algorithm. First, the object detection algorithm is used to find the bounding box of the object, semantic segmentation is then performed within the bounding box of each object, and, finally, each segmentation result is output as an instance. In the single-stage instance segmentation algorithm, inspired by YOLO, SOLO [11] directly decouples the instance segmentation problem into category prediction and instance mask generation problems. There is no need to generate bounding boxes during the prediction process. SOLO V2 [27] makes a further adjustment; Center-Mask [28] adds a head network to predict the mask to the single-order end object detection algorithm, FCOS [29], to complete instance segmentation. Although these methods have a certain speed advantage over the two-step method, they are usually unable to achieve the accuracy of the two-step method. In terms of the two-stage algorithm, He Kaiming et al. proposed Mask R-CNN [7], a simple and effective instance segmentation framework. Mask R-CNN adds a mask branch to the head network of Faster R-CNN. Additionally, the original classification branch and regression branch are juxtaposed with the mask branch. Inspired by Mask R-CNN, Shu Liu [8] et al. proposed PANet, which makes full use of shallow network features for instance segmentation; Mask Scoring R-CNN [9], on the basis of the Mask R-CNN, expands with an additional mask branch in order to obtain a more accurate mask. Cascade Mask R-CNN [10] combines Mask R-CNN with Cascade R-CNN, which slightly improves detection accuracy, but it is still unsatisfactory in mask prediction. The key reason for this is that the ability of the CNN to capture long-range features is relatively weak, and the problem of establishing the global relations in the image has not been solved.

2.2. Transformer-Based Object Detection and Instance Segmentation

Transformers are deep neural networks mainly based on the self-attention mechanism [12], and were originally applied in the field of natural language processing and later extended to computer vision tasks. Compared with the CNN network, the advantage of the transformer lies in the use of self-attention to capture global contextual information to

establish a long-range dependence on the object, thereby extracting more powerful features. The structure of the self-attention mechanism is shown in Figure 2. For each element in the input sequence, it will generate Q (query), K (key), and V (value) through three learning matrices. In order to determine the relevance between an element and other elements in the sequence, the dot product is calculated between the Q vector of this element with the K vectors of other elements. The results determine the relative importance of patches in the sequence. Then, the results of the dot product are then scaled and fed into a softmax. Finally, the value of the vector for each patch embedding is multiplied by the output of the softmax to find the patch with the high attention scores.

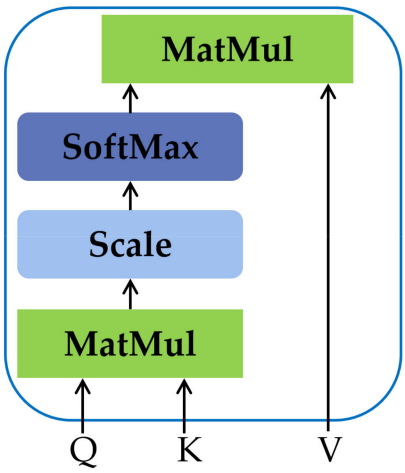

Figure 2. Structure of self-attention mechanism.

In 2020, Carion et al. [14] combined the CNN and the transformer to propose a complete end-to-end DETR object detection framework, applying transformer architecture to object detection for the first time. Zhu [30] et al. proposed the Deformable DETR model that draws on the variable convolutional neural network. Zheng et al. [31] proposed the end-to-end object detection with adaptive clustering transformer (ACT) to reduce the computational complexity of the self-attention module. DETR can naturally extend the panoramic segmentation task by attaching a mask head to the decoder and obtaining competitive results. Wang et al. [32] proposed a transformer-based video instance segmentation (VisTR) model, which takes a series of images as inputs and generates corresponding instance prediction results. Although these models perform well in object detection tasks, they still have many shortcomings. For example, the detection speed of the DETR series models is slow, and the detection performance of small objects is not effective.

For remote sensing images, the image resolution is high, which increases the calculation size of the transformer models. Remote sensing images usually have complex background information and variable object scales, and the training effect of a single-level transformer network is not effective. Based on the above problems, the Swin transformer [18] was proposed to solve the problems of a high amount of computation and the poor detection effect of dense objects, but it still has weak local information acquisition capabilities.

Therefore, for the object detection and instance segmentation of remote sensing images, we need to exploit both the advantages of CNNs to address the underlying vision and those of transformers to address the relationship between visual elements and objects. We need to then design a novel backbone network and detection framework and focus on enhancing the mask prediction ability to improve the detection and segmentation accuracy of remote sensing images.

3. Materials and Methods

This section focuses on the designed network structure. As shown in Figure 3, the model feeds the input image to the local perception Swin transformer (LPSW) backbone network. After the feature map is generated, it is sent to the spatial attention interleaved execution cascade (SAIEC) network model after the FPN structure. The back-end of the model performs feature map classifications, bounding box regression, and instance segmentation tasks. In our model, each bounding box is divided into object and non-object regions. The detailed information of each module is introduced below:

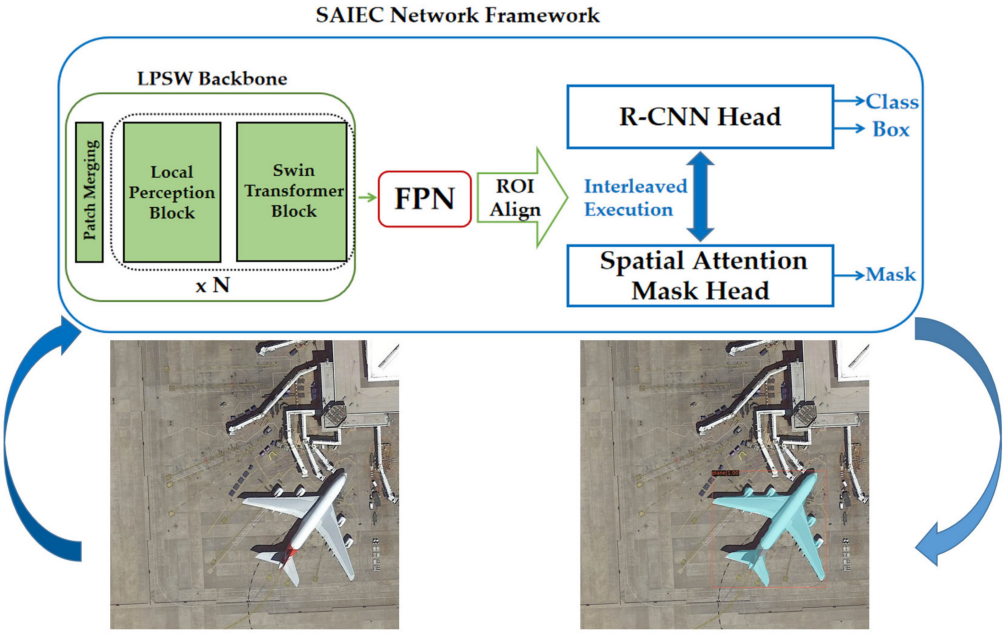

Figure 3. Flow chart of the designed model, which combines the proposed local perception Swin transformer (LPSW) backbone network with the spatial attention interleaved execution cascade (SAIEC) network framework and includes feature pyramid network (FPN) and region of interest (ROI) structures. The new network model can accurately complete remote sensing image object detection and instance segmentation tasks.

3.1. Local Perception Swin Transformer (LPSW) Backbone

The flow chart of the proposed local perception Swin transformer (LPSW) backbone network is shown in Figure 4. The Swin transformer provides four versions of the model, which, from large to small [18], are Swin-T, Swin-S, Swin-B and Swin-L. Taking into account the particularity and computational complexity of remote sensing images, this paper introduces Swin-T. Each stage has 2, 2, 6, and 2 blocks, respectively.

Similar to ViT, it first splits an input RGB image into non-overlapping patches by patch partition layer. Each patch is treated as a "token" and its feature is set as a concatenation of the raw pixel RGB values. The Swin transformer contains four stages to produce a different number of tokens. Given an image with a size of H × W, a token is a raw pixel concatenation vector of an RGB image patch with the size of 4 × 4. A linear embedding is employed on this token to map it in a vector with the dimension C. Stages 1, 2, 3, and 4 produce $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$ tokens, respectively. Each stage consists of a patch merging block (a combination of a patch partition layer and a linear embedding layer), local perception block, and some Swin transformer blocks.

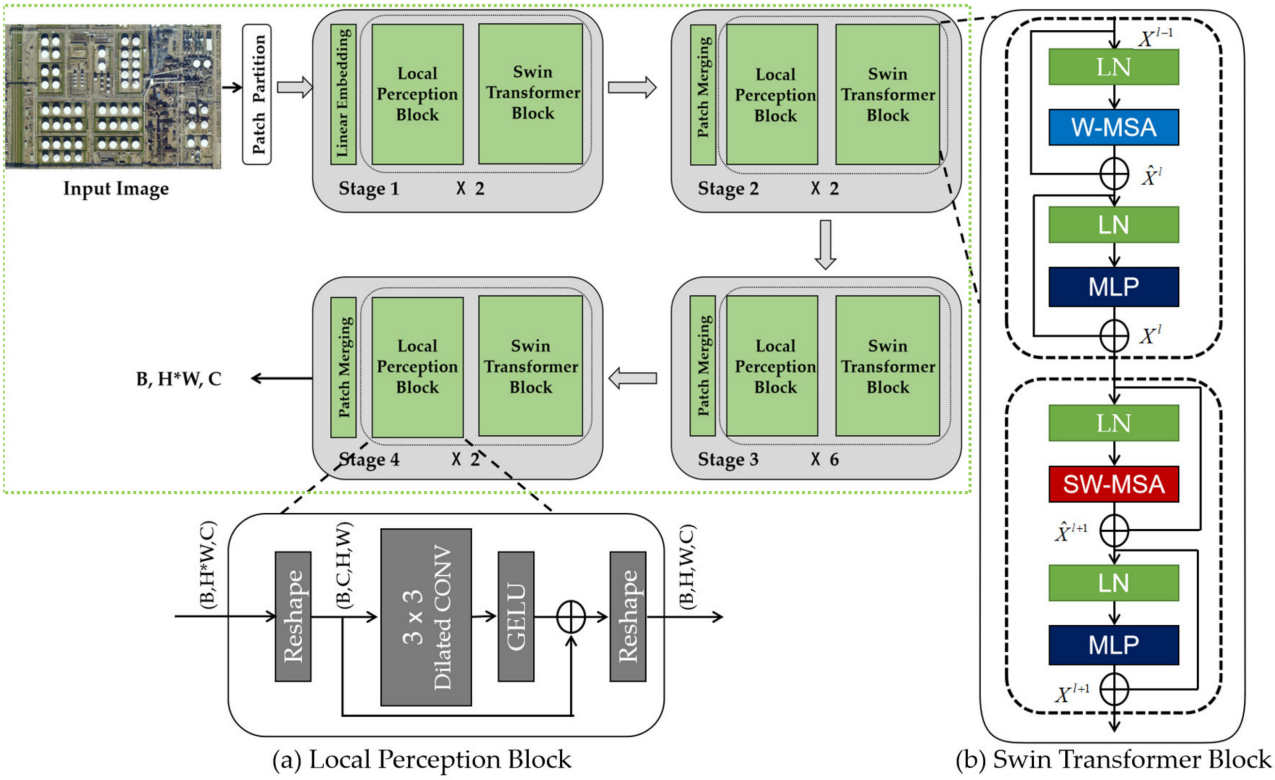

Figure 4. The architecture of the local perception Swin transformer (LPSW). (**a**) The detailed structure of the local perception block; (**b**) the detailed structure of the Swin transformer block.

3.1.1. Swin Transformer Block

The Swin transformer block is the core part of the Swin transformer algorithm. The detailed structure is shown in Figure 4b. The block is composed of window multi-head self-attention (W-MSA), shifted windows multi-head self-attention (SW-MSA) and multilayer perceptron (MLP). Inserting a layernorm (LN) layer in the middle makes the training more stable and uses a residual connection after each module. This part can be expressed as Equation (1):

$$\begin{aligned} \hat{X}^{l} &= W-MSA\left(LN\left(X^{l-1}\right)\right)+X^{l-1} \\ X^{l} &= MLP\left(LN\left(\hat{X}^{l}\right)\right)+\hat{X}^{l} \\ \hat{X}^{l+1} &= SW-MSA\left(LN\left(X^{l}\right)\right)+X^{l} \\ X^{l+1} &= MLP\left(LN\left(\hat{X}^{l+1}\right)\right)+\hat{X}^{l+1} \end{aligned} \tag{1}$$

3.1.2. W-MSA and SW-MSA

Compared with the Multi-Head Self Attention (MSA) [12] in the traditional ViT, the W-MSA in the Swin transformer block controls the calculation area in a window as a unit (window size is set to 7 by default). This reduces the amount of network calculations and reduces the complexity to a linear ratio of the image size, as shown in Figure 5. MSA lacks connections across windows. The position of SW-MSA is connected to the W-MSA layer. Therefore, SW-MSA is required to provide a different window segmentation method after W-MSA to realize cross-window communication.

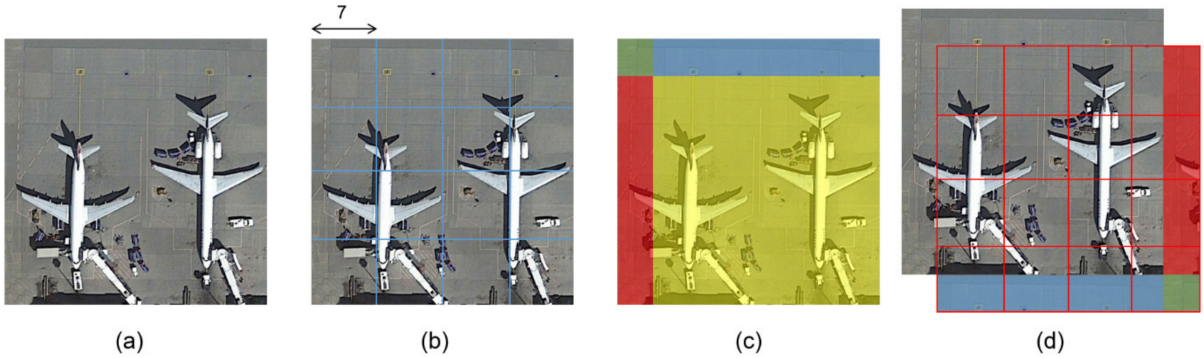

Figure 5. The mechanism of action of the shifted windows. (**a**) The input image; (**b**) Window segmentation (window size is set to 7) of the input image through the window multi-head self-attention (W-MSA); (**c**) Action of the shifted windows; (**d**) A different window segmentation method through the shifted windows multi-head self-attention (SW-MSA).

The result of window segmentation of the input image through W-MSA is shown in Figure 5b. Each cycle of the image is moved up and left by half the size of the window, and the blue and red areas in Figure 5c are then moved to the lower and right sides of the image, respectively, as shown in Figure 5d. On the basis of these shifts, the window is divided according to W-MSA, and SW-MSA has a window segmentation method different from W-MSA.

3.1.3. Local Perception Block (LPB)

Position encoding in a transformer can easily fail to detect the local correlation and structural information of the image. Although the Swin transformer has a shift window scheme of sequential layers in a hierarchical structure, a large range of spatial context information is still not well encoded. In order to alleviate this problem, we proposed the local perception block (LPB), which is inserted in front of the Swin transformer block. The composition of the local perception block is shown in Figure 4a.

Considering that the data flow in the Swin transformer consists of vectors instead of feature maps in traditional CNNs, in the LPB, it firstly reshapes a group of vector features into a spatial feature map. For example, a token (B, H * W, C) is reshaped as a feature map (B, C, H, W). A layer of 3×3 dilated convolution (dilation = 2) and a GELU activation function is then added, and a residual connection is used to increase the extraction of spatial local features while keeping the receptive field sufficiently large. Finally, the feature map is reshaped to (B, H, W, C) and sent to the Swin transformer block.

Through the characteristics of dilated convolution, the receptive field of the spatial image is increased, such that a large range of contextual information can be coded well at different scales. Dilated convolution was proposed by Yu and Koltun [33] in 2015. Compared with the traditional convolution operation, dilated convolution supports the expansion of the receptive field. It is worth noting that the traditional 3×3 convolutions each have a 3×3 field. If it is a dilated convolution (dilation = 2) with the same kernel size, the receptive field is 7×7. Therefore, dilated convolution can extend the corresponding field without a loss of feature resolution.

3.2. Spatial Attention Interleaved Execution Cascade (SAIEC)

The proposal of Cascade R-CNN mainly defines the input intersection over union (IoU) threshold of positive and negative samples at different stages. The detector pays more attention to the positive samples within the threshold because of the difference in IoU input at each stage. The output IoU threshold is better than the input IoU threshold, which provides better positive samples for the next stage. Each stage is in a progressive relationship, such that the detector effect can gradually improve. Cascade Mask R-CNN is a product that directly combines Mask R-CNN and Cascade R-CNN. Although it improves in box AP, it does not improve significantly in mask AP. Therefore, inspired by the HTC

algorithm, we improve Cascade Mask R-CNN and propose the spatial attention interleaved execution cascade (SAIEC), a new framework of instance segmentation. The specific improvement methods are as follows.

3.2.1. Interleaved Execution and Mask Information Flow

We improved the network head of Cascade Mask R-CNN, as shown in Figure 6. Although Cascade R-CNN forces two branches into each stage, there is no interaction between the two branches during the training process, and they are executed in parallel. Therefore, we propose the interleaved execution; that is, in each stage, the box branch is executed first, and the updated bounding box predictions are then passed to the mask branch to predict the mask, as shown in Figure 6b. In the figure, F represents the features of the backbone network, P is the ROI Align or ROI pooling, and Bi and Mi denote the box and mask head at the i-th stage. This not only increases the interaction between different branches in each stage, but also eliminates the gap between training and testing processes.

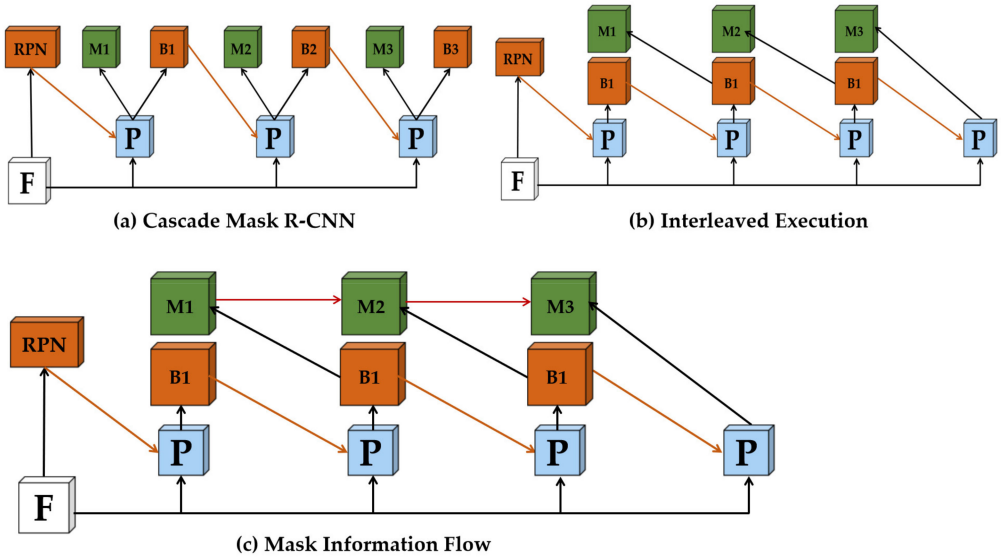

Figure 6. The Cascade Mask R-CNN network head improvement process. (**a**) The Cascade Mask R-CNN network head; (**b**) The addition of the interleaved execution in the network head; (**c**) The final network head structure after adding Mask Information Flow.

At the same time, in the Cascade Mask R-CNN, only the current stage in the box branch has an impact on the next stage, and the mask branch between different stages does not have any direct information flow. In order to solve this problem, we added a connection between adjacent mask branches, as shown in Figure 6c. We provided mask information flow for the mask branch so that M_{i+1} could obtain the features of M_i. The specific implementation is shown above in the red part of Figure 7. We used the feature of Mi to perform feature embedding through a 1×1 convolution, and then entered it into M_{i+1}. In this way, M_{i+1} could obtain the characteristics of not only the backbone, but also the previous stage.

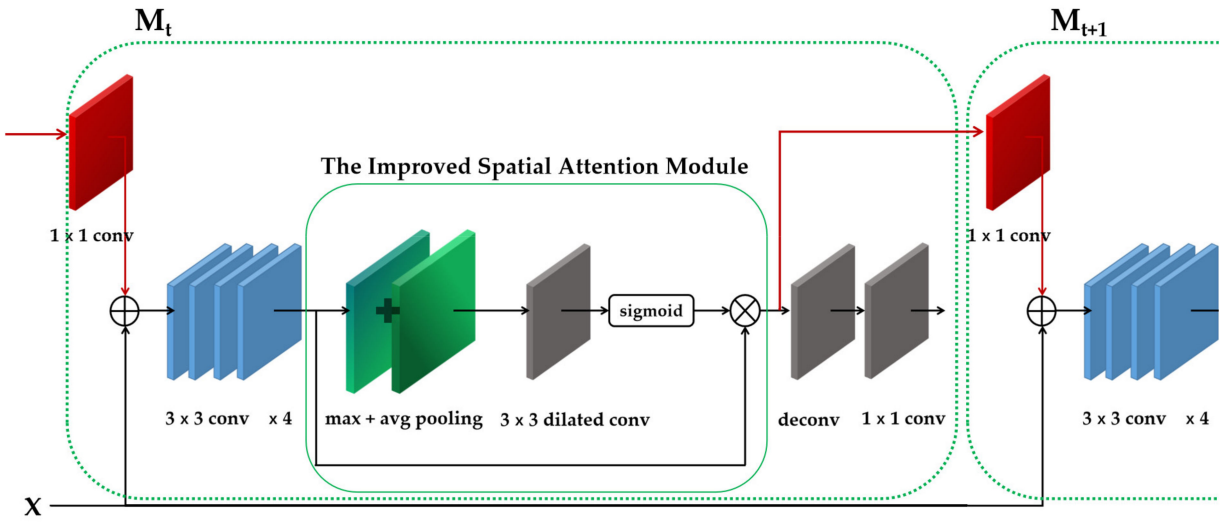

Figure 7. Structure of the spatial attention mask head. It includes the improved spatial attention module, which helping to focus on objects and suppressing noise.

3.2.2. Spatial Attention Mask Head

The attention method [34] helps one to focus on important features and suppress unnecessary noise. Inspired by the spatial attention mechanism [35], we designed the spatial attention mask head, using the spatial attention module to guide the mask head, in order to highlight meaningful pixels and suppress useless pixels. As shown in Figure 7, we improved on the original mask head. We designed an improved spatial attention module and inserted it before transposed convolution. In the spatial attention mask head, the resized local features need to pass through four 3×3 convolution layers with 256 channels, and then pass through the improved spatial attention module. The improved spatial attention module first generates pooled features P_{max} and P_{avg} by both average and max pooling operations, respectively, along the channel axis, and then aggregates them via concatenation. This is followed by a 3×3 dilated convolution layer and is normalized by the sigmoid function. The computation process is summarized as follows:

$$X_{sa} = X_i \otimes sigmoid(D_{3\times3}(P_{\max} \circ P_{avg})) \quad (2)$$

where $\otimes$ denotes element-wise multiplication, X_{sa} is the attention-guided feature map, $D_{3\times3}$ is the 3×3 conv layer, and $\circ$ represents the concatenate operation. Afterwards, 2×2 deconv is used for upsampling and 1×1 conv is used to predict the category of the specific mask. By combining the above structures, we completed the design of the mask branch in the SAIEC framework. The spatial attention mask head not only effectively improves the cross-stage information communication in the network, but also adds a spatial attention mechanism to help with focusing on objects and suppressing noise.

4. Results

4.1. Dataset

There are many conventional object detection datasets. Models that are trained based on conventional datasets do not perform well on remote sensing images. The main reason is the particularity of remote sensing images, and few datasets are related to remote sensing image object detection and instance segmentation. Therefore, we selected images from three public datasets (Object Detection in Optical Remote Sensing Images (DIOR) [36], High Resolution Remote Sensing Detection (HRRSD) [37], and convolutional neural networks for object detection in VHR optical remote sensing images (NWPU VHR-10) [38]) to produce new remote sensing image object detection and instance segmentation datasets. The research group of the Western University of Technology proposed a large-scale benchmark

dataset "DIOR" for object detection in optical remote sensing images, which consists of 23,463 images and 190,288 object examples and is based on deep learning. The image size is 800 × 800 pixel, and the resolution ranges from 0.5 m to 30 m. The aerospace remote sensing object detection dataset "NWPU VHR-10," annotated by Northwestern Polytechnical University, has a total of 800 images, including 650 of the objects and 150 background images. Objects include: airplanes, ships, oil tanks, baseball fields, and nets. There are 10 categories of courts, basketball courts, track and field arenas, ports, bridges, and vehicles. HRRSD is a dataset produced by the Optical Image Analysis and Learning Center of the Xi'an Institute of Optics and Fine Mechanics, Chinese Academy of Sciences for research on object detection in high-resolution remote sensing images. The image resolution ranges from 500 × 500 pixels to 1400 × 1000 pixels.

We selected high-resolution images from these three public datasets for manual annotation, and performed data enhancement on the labeled dataset by vertically flipping, horizontally flipping, rotating, and cutting to create the MRS-1800 remote sensing mask dataset. We merged these three classic remote sensing datasets together, which can be regarded as a means of data enhancement and expansion. This approach allowed our dataset to contain more styles and sizes of remote sensing images, making the dataset more challenging. Training our model in this way can help overcome the overfitting problem, thereby improving the robustness and generalization ability of the model.

The MRS-1800 dataset has a total of 1800 remote sensing images. The size of the images varies and the dataset contains a variety of detection objects. The detection objects are divided into three categories: planes, ships, and storage tanks. The specific information of the dataset is shown in Table 1.

Table 1. Number distribution of datasets and class.

Dataset	Dior	Hrrsd	Nwpu Vhr-10	Statistics
Number	403	1093	304	1800
Class	Plane	Ship	Storage tank	
Number	674	687	557	

Figure 8 shows part of the images and mask information of the MRS-1800 dataset. Different sizes of high-resolution images contain different types of objects. We used LabelMe 4.5.9 (Boston, MA, USA) to mark the image with mask information and generate the corresponding "json" files. The dataset contains planes, ships, and storage tanks of different sizes. A total of 16,318 objects were collected, and the object sizes include three types: large, medium and small (ranging from 32 × 32 pixels to 500 × 500 pixels), and the numbers of these types are evenly distributed. We used 1440 images as the training set, 180 images as the validation set, and 180 images as the test set, according to the 8:1:1 allocation ratio.

4.2. Experiments and Analysis

Throughout the experiment, we used a computer equipped with a Geforce RTX 3060 GPU (12 G) as the hardware platform for the experiment. We used pytorch as the DL framework, and the compilation environment was python 3.8 and pytorch 1.8.1. We used multiple classic frameworks such as Mask R-CNN [7], Sparse R-CNN [39], Cascade Mask R-CNN [10], DETR [14], and so on. Additionally, we used Resnet-50 (R-50), the Swin transformer and LPST backbone networks. Suitable pre-training models were chosen to train the self-made dataset, MRS-1800.

We used the same settings in training for the proposed models: multi-scale training (the input size was adjusted so that the short side was between 480 and 800, and the long side was, at most, 1333), the AdamW [40] optimizer (the initial learning rate was 0.00001, the weight decay was 0.05, and the batch size was 1), and 3× scheduling (50 epochs with the learning rate decayed by 10× at 27 epochs). We chose some deep learning indicators as our experimental evaluation criteria, such as frames per second (FPS), AR_S (the average

recall measurement value of object frames smaller than 32 × 32 pixels), average precision (AP), AP_{50} (AP measurement value when the IoU threshold is 0.5), AP_{75} (AP measurement value when the IoU threshold is 0.75), AP_S (the AP measurement value of object frames smaller than 32 × 32 pixel), and their mask counterparts: mask AP, mask AP_{50}, mask AP_{75}, and mask AP_S. AP and AR are averaged over multiple intersection over union (IoU) values, where the IoU threshold value ranges from 0.5 to 0.95, with a stride of 0.05. Mask AP is used to comprehensively evaluate the effectiveness of the instance segmentation model. The difference from box AP is only that the objects of the IoU threshold are different. The box AP functions in the standard ordinary ground truth and the IoU value of the prediction box, while the mask AP functions in the ground truth mask and the mask IoU of the prediction mask.

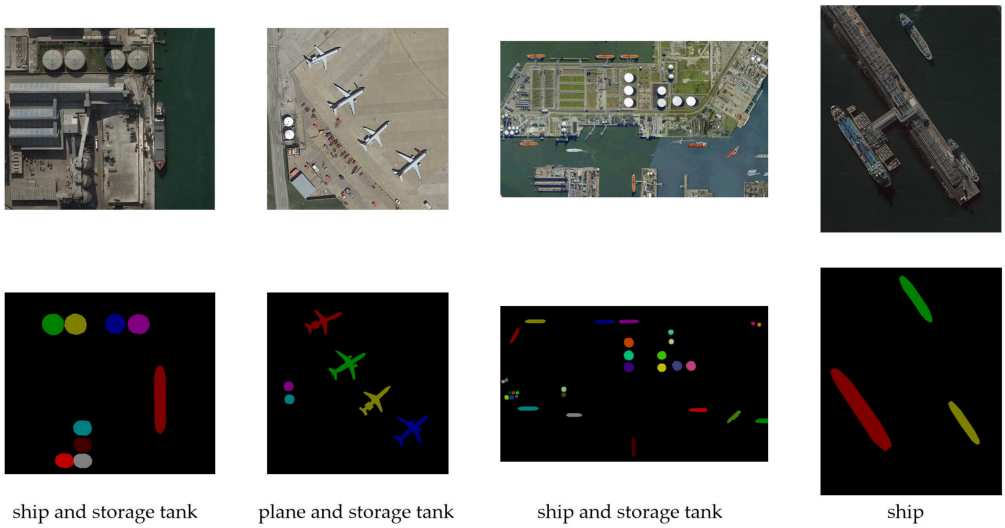

Figure 8. MRS-1800 dataset display. The top row is the remote sensing images of different sizes randomly selected in the dataset, and the next row contains corresponding mask images produced with LabelMe.

Figure 9 shows the mask loss function graph during the training of the network model we designed. It can be seen that the network model is still under-fitting during the first 38 k steps (27 epochs), and the loss function fluctuates greatly. We adjusted the learning rate in time after 38 k steps to avoid overfitting. The training loss value after the final step was 0.03479.

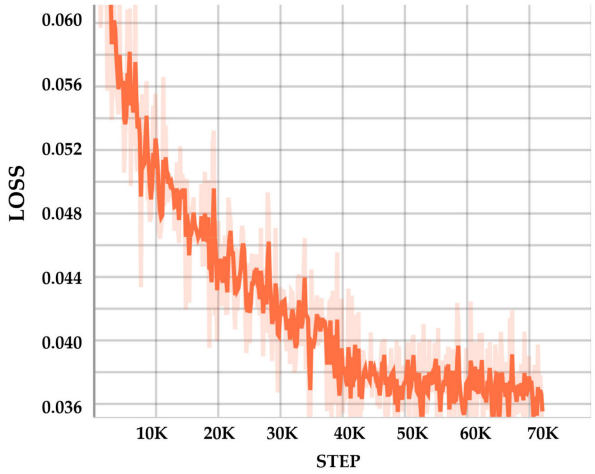

Figure 9. The training mask loss function diagram of the LPSW backbone using the SAIEC framework on the dataset.

4.3. Ablation Experiment

We performed a number of ablation experiments to gradually verify each component in the proposed method in this section. We analyzed and compared the data trained on the MRS-1800 dataset. The specific experiments are as follows:

4.3.1. Study for Optimizer and Initial Learning Rate

The optimizer plays an important role in deep learning. We first conducted ablation experiments on the selection of the optimizer and the corresponding parameter values. Commonly used optimizers for object detection are the SGD [41] and the AdamW [40]. We chose Cascade Mask R-CNN as the network framework and the Swin transformer as the backbone network, using the SGD and the AdamW optimizers for experiments. At the same time, in order to explore the influence of the optimizer's initial learning rate parameters on the experiment, we set the initial learning rate to 1×10^{-4}, 1×10^{-5}, and 1×10^{-6} for comparison experiments.

It can be seen from Table 2 that the overall performance of the AdamW is better than that of the SGD, and AP can increase by more than 8% by replacing the optimizer. In addition, it can be drawn from the table that when the initial learning rate is 1×10^{-5}, the model can achieve the highest detection accuracy. Therefore, we can conclude that the Swin transformer can achieve a better performance when the AdamW optimizer is used for model training and the initial learning rate is 1×10^{-5}.

Table 2. The results of optimizers and learning rate ablation study.

Method	Optimizer	Learning Rate	AP^{box}	AP^{mask}
Swin-T	SGD	1×10^{-4}	60.1	33.9
		1×10^{-5}	69.2	52.1
		1×10^{-6}	53.6	41.5
	AdamW	1×10^{-4}	73.9	58.0
		1×10^{-5}	77.2	60.7
		1×10^{-6}	75.0	58.4

4.3.2. Experiment for the Swin Transformer and LPST Backbone

We inserted the Swin transformer (Swin-T) and LPST as a new backbone network into typical object detection frameworks: Mask R-CNN and Cascade Mask R-CNN, for object detection and instance segmentation experiments. We compared them with traditional convolutional networks (Sparse R-CNN, PANet, and Mask Scoring R-CNN) and previous transformer networks (DETR). The experimental results are shown in Table 3.

Table 3. Detection and segmentation performance of different methods.

Various Frameworks											
Method	**Backbone**	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{s}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{s}	AR_S	**FPS**
Mask R-CNN	R-50	69.0	91.5	83.3	31.6	57.2	90.5	58.9	25.0	44.1	11.5
	Swin-T	75.5	92.8	88.1	44.6	60.9	91.7	66.6	34.1	47.2	8.6
	LPST	75.8	93.1	88.0	46.6	60.4	92.1	65.8	36.2	49.2	8.1
Cascade Mask R-CNN	R-50	72.1	91.0	83.3	31.3	56.6	90.3	57.7	32.9	38.5	8.4
	Swin-T	77.2	92.7	87.6	41.5	60.7	91.4	66.3	31.7	45.5	5.4
	LPST	77.4	93.0	88.0	46.7	61.3	91.7	68.3	36.8	50.0	5.1
Mask Scoring	R-50	71.9	91.5	84.5	40.3	60.7	90.4	67.4	32.4	43.5	11.4
Sparse R-CNN	R-50	73.9	91.0	83.8	35.4					39.4	13.4
PANet	R-50	71.6	91.8	84.5	35.3					38.3	12.1
DETR	R-50	65.3	86.7	74.3	21.4					29.7	15.1

Table 3 shows that, compared with the traditional CNN models, in each framework, the use of the Swin transformer and the LPSW as the backbone network has a greater improvement in the various indicators of the experimental results. Compared with the previous transformer network, the experimental result of Swin-T based on Cascade Mask R-CNN is 11.9% AP and 20.1% APs higher than DETR, which is sufficient to prove the superiority of the Swin transformer. It overcomes the shortcoming of the transformer's poor small-scale objects detection and slow convergence.

At the same time, we compared the LPSW with Swin-T using the same basic framework. The experimental results show that, after using the LPSW, the experimental indicators are improved: when using the Cascade Mask R-CNN framework, APs increased by 5.2%, mask AP_S increased by 5.1%, ARs increased by 4.5%, and mask AP and AP increased by 0.6% and 0.2%, respectively. The data show that, for the Swin transformer, the LPSW significantly improved the detection and segmentation of small-scale objects without a significant reduction in the inference speed. Due to the large number of small objects in remote sensing images, this improvement was exactly what was necessary.

The result generated by the traditional Cascade Mask R-CNN, the Swin-T, and LPSW are shown in Figures 10–12. Compared with the traditional CNN network, the Swin transformer pays more attention to the learning of global features; particularly, the detection ability of image edge objects was greatly improved. As shown in the enlarged images on the right side of Figures 10 and 11, Cascade Mask R-CNN has a low confidence in terms of the detection of ships in the upper right of the image, and false detection objects appeared. The Swin transformer does not detect false objects for the same edge detection area, and the confidence of object detection increases.

Compared with the Swin transformer, the LPSW pays more attention to local features. As shown in Figures 11 and 12, the most obvious difference between the two images is that the LPSW eliminates the false detection of white buildings in the lower part of the image. In addition, the number of real objects detected by the LPSW increases, and the confidence of object detection also improves.

Figure 10. The results of Cascade Mask R-CNN using the Resnet-50 backbone.

Figure 11. The results of Cascade Mask R-CNN using the Swin transformer backbone.

Figure 12. The results of Cascade Mask R-CNN using the LPSW backbone.

4.3.3. Experience for SAIEC and the New Network Model

We used the newly designed SAIEC network framework to perform object detection and instance segmentation on remote sensing images. The MRS-1800 dataset was used, and the backbone network used the LPSW and Swin-T. In order to verify the effectiveness of the improved model designed, we compared the experimental results with data in Section 4.3.1. At the same time, we compared and analyzed the designed model with the current SOTA object detection model on the COCO dataset (the Swin transformer using an HTC framework) [18].

Since this paper improves Cascade Mask R-CNN and the Swin transformer, respectively, we considered Swin-T using the Cascade Mask R-CNN framework as the baseline. It can be concluded from Table 4 that, compared with the baseline, the object detection and instance segmentation model we designed (the SAIEC network framework using the LPSW backbone) saw an improvement in all indicators. Among them, mask AP increased by 1.7%, mask AP_{75} increased by 4.0%, mask AP_S increased by 3.6%, AP increased by 1.1%, AP_S increased by 4.6%, and AR_S increased by 7.7%.

Table 4. Performance comparison of each part of the improved model.

Method	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{s}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{s}	AR_S	FPS
Cascade Mask R-CNN (Swin-T) *baseline*	77.2	92.7	87.6	41.5	60.7	91.4	66.3	31.7	45.5	5.4
Cascade Mask R-CNN (LPSW)	77.4	93.0	88.0	46.7 (+5.2)	61.3	91.7	68.3	36.8 (+5.1)	50.0	5.1
HTC (Swin-S [18])	77.8	93.3	88.1	46.6	61.9	92.4	68.8	35.9	51.8	4.6
HTC (Swin-T)	77.4	92.7	88.2	41.7	61.6	91.9	69.7	31.4	49.6	5.4
SAIEC (Swin-T)	77.8	93.2	88.7	43.4	62.3	92.0	69.4	33.7	50.0	5.5
SAIEC (LPSW) (ours)	78.3 (+1.1)	93.0	88.7	46.1 (+4.6)	62.4 (+1.7)	92.3	70.3 (+4.0)	35.3 (+3.6)	53.2 (+7.7)	5.1

The data show that the network model we designed greatly improved the detection and segmentation of small-scale objects in remote sensing images. The increase in the detection rate of small-scale objects affects the improvement of AP_{75} and mask AP_{75}. Compared with the current SOTA network (the Swin transformer using an HTC framework), the indicators of the model designed in this article are similar or even surpassed, and the inference speed is higher (5.1 FPS vs. 4.6 FPS). The above experimental data demonstrate the advantages of the model proposed in this paper in remote sensing image object detection and instance segmentation.

Figure 13 shows the remote sensing image segmentation results of traditional Cascade Mask R-CNN, the Swin transformer using Cascade Mask R-CNN and the network proposed in this paper. It can be seen from the figure that Cascade R-CNN is not ideal in terms of overall segmentation effect or edge detail processing. Although the Swin transformer is optimized for the overall segmentation effect, it does not accurately present the details of the edge. In contrast, it can be seen from the figure that the network model proposed in this paper shows good results in remote sensing images, and the details at the edges are well segmented.

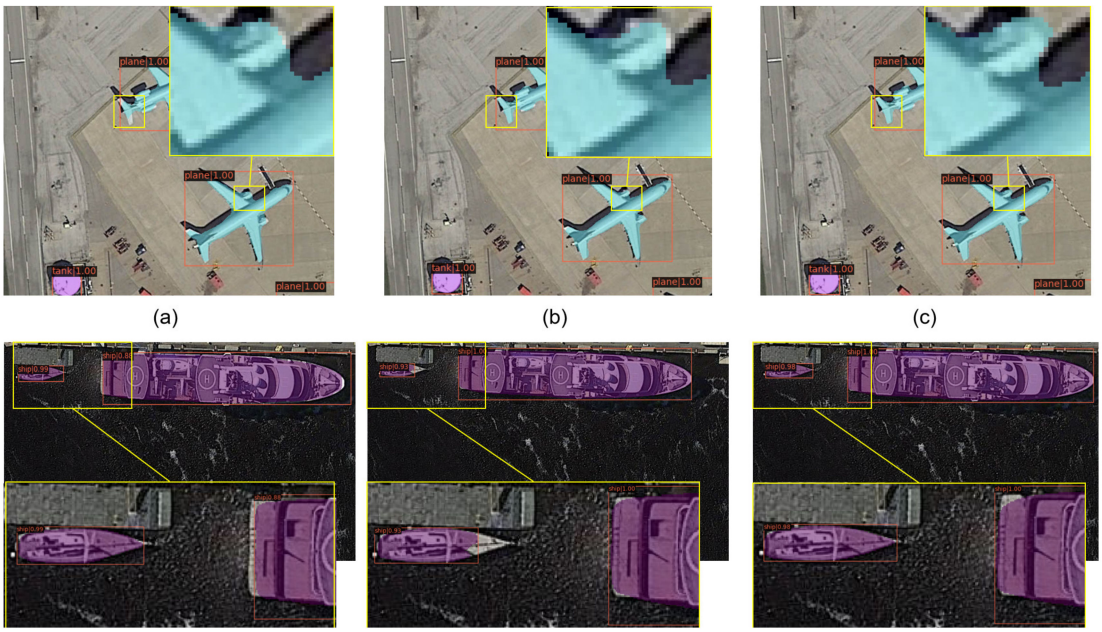

Figure 13. Segmentation results of remote sensing images by various networks. (**a**–**c**) Detection results of the traditional Cascade Mask R-CNN, the Swin transformer using Cascade Mask R-CNN and the LPSW using SAIEC.

5. Discussion

Because convolutional neural networks (CNN) have shown defects in the object detection of remote sensing images. We innovatively introduced the Swin transformer as the basic detection network, and designed the LPSW backbone network and SAIEC network framework for improvement. Experimental results show that the new network model we designed can greatly improve the detection effect of small-scale objects in remote sensing images and can strengthen the segmentation accuracy of multi-scale objects. However, it is worth noting that our experiment was only conducted on the MRS-1800 dataset due to the lack of mature and open remote sensing mask datasets, which may be limited in number and type. Moreover, our research on the improvement and promotion of the model inference speed is not sufficient. Generally, the processed images will be affected by uncertain factors [42]; however, it is also necessary to use fuzzy preprocessing techniques on images. In future research, we will focus on solving the above problems. First, we will search for and create more remote sensing mask datasets containing more object types, and use more realistic and representative datasets to validate our new models. Secondly, designing a lightweight network model to improve the inference speed without the loss of detection accuracy will be our next research direction.

6. Conclusions

Remote sensing image object detection and instance segmentation tasks have important research significance for the development of aviation and remote sensing fields, and have broad application prospects in many practical scenarios. First, we created the MRS-1800 remote sensing mask dataset, which contains multiple types of objects. Second, we introduced the Swin transformer into remote sensing image object detection and instance segmentation. This paper improved the Swin transformer based on the advantages and disadvantages of transformers and CNNs, and we designed the local perception Swin transformer (LPSW) backbone network. Finally, in order to increase the mask prediction accuracy of remote sensing image instance segmentation tasks, we designed the spatial attention interleaved execution cascade (SAIEC) network framework. Experimental conclusions can be drawn for the MRS-1800 remote sensing mask dataset: (1) According to experiments, the SAIEC model using the LPSW as the backbone can improve mask AP by 1.7%, mask AP_S by 3.6%, AP by 1.1%, and AP_S by 4.6%. (2) The innovative combination of CNNs and transformers' advantages in capturing local information and global information can significantly improve the detection and segmentation accuracy of small-scale objects. Inserting the interleaved execution structure and the improved spatial attention module into the mask head can help to suppress noise and enhance the mask prediction of the network. (3) Compared with the current SOTA model in the COCO dataset, the model proposed in this paper also demonstrates important advantages.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: The authors thank the team of optical sensing and measurement of Xidian University for their help. This research was supported by the National Natural Science Foundation of Shaanxi Province (Grant No.2020 JM-206), the National Defense Basic Research Foundation (Grant No.61428060201) and the 111 project (B17035).

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

References

1. Cao, C.; Wang, B.; Zhang, W.; Zeng, X.; Yan, X.; Feng, Z.; Liu, Y.; Wu, Z. An Improved Faster R-CNN for Small Object Detection. *IEEE Access* **2019**, *7*, 1. [CrossRef]
2. Zhu, W.T.; Xie, B.R.; Wang, Y.; Shen, J.; Zhu, H.W. Survey on Aircraft Detection in Optical Remote Sensing Images. *Comput. Sci.* **2020**, *47*, 1–8.
3. Wu, J.; Cao, C.; Zhou, Y.; Zeng, X.; Feng, Z.; Wu, Q.; Huang, Z. Multiple Ship Tracking in Remote Sensing Images Using Deep Learning. *Remote Sens.* **2021**, *13*, 3601. [CrossRef]
4. Li, X.Y. Object Detection in Remote Sensing Images Based on Deep Learning. Master's Thesis, Department Computer Application Technology, University of Science and Technology of China, Hefei, China, 2019.
5. Hermosilla, T.; Palomar, J.; Balaguer, Á.; Balsa, J.; Ruiz, L.A. Using street based metrics to characterize urban typologies. *Comput. Environ. Urban Syst.* **2014**, *44*, 68–79. [CrossRef]
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE ICCV, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
8. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2018; pp. 8759–8768. [CrossRef]
9. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask Scoring R-CNN. In Proceedings of the IEEE/CVF CVPR, Long Beach, CA, USA, 16–20 June 2019; pp. 6409–6418.
10. Dai, J.F.; He, K.M.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
11. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. Solo: Segmenting objects by locations. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 649–665.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA; 2017; pp. 5998–6008.
13. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), Virtual Event, Austria, 3–7 May 2021.
14. Nicolas, C.; Francisco, M.; Gabriel, S.; Nicolas, U.; Alexander, K.; Sergey, Z. End-to-End Object Detection with Transformers. In Proceedings of the 16th ECCV, Glasgow, UK, 23–28 August 2020; pp. 213–229.
15. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th ICML, Virtual Event, 18–24 July 2021; pp. 10347–10357.
16. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 6881–6890.
17. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.Q.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030. Available online: https://arxiv.org/abs/2103.14030 (accessed on 19 October 2021).
19. Chen, K.; Pang, J.M.; Wang, J.Q.; Xiong, Y.; Li, X.X.; Sun, S.Y.; Feng, W.F.; Liu, Z.W.; Shi, J.P.; Wangli, O.Y.; et al. Hybrid Task Cascade for Instance Segmentation. In Proceedings of the IEEE CVPR, Long Beach, CA, USA, 15–21 June 2019; pp. 4974–4983.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
21. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the IEEE ECCV, Amsterdam, Netherlands, 11–14 October 2016; pp. 21–37.
24. Lin, T.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HA, USA, 21–26 July 2017; pp. 2117–2125.
25. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
26. Liang, X.; Lin, L.; Wei, Y.C.; Shen, X.H.; Yang, J.C.; Yan, S.C. Proposal-Free Network for Instance-Level Object Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2978–2991. [CrossRef] [PubMed]

27. Wang, X.L.; Zhang, R.F.; Kong, T.; Li, L.; Shen, C.H. SOLOv2: Dynamic and Fast Instance Segmentation. *arXiv* **2020**, arXiv:2003.10152. Available online: https://arxiv.org/abs/2003.10152v3 (accessed on 19 October 2021).
28. Lee, Y.; Park, J. Centermaslc: Real-Time Anchor-Free Instance Segmentation. In Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13906–13915.
29. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully Convolutional One-Stage Object Detection. In Proceedings of the the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
30. Zhou, X.Z.; Su, W.J.; Lu, L.W.; Li, B.; Wang, X.G.; Dai, J.F. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2020.
31. Zheng, M.H.; Gao, P.; Wang, X.G.; Li, H.S.; Dong, H. End-to-End Object Detection with Adaptive Clustering Transformer. *arXiv* **2020**, arXiv:2011.09315. Available online: https://arxiv.org/abs/2011.09315 (accessed on 19 October 2021).
32. Wang, Y.Q.; Xu, Z.L.; Wang, X.L.; Shen, C.H.; Cheng, B.S.; Shen, H.; Xia, H.X. End-to-End Video Instance Segmentation with Transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 8741–8750.
33. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. In Proceedings of the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
34. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 3–19.
35. Zhu, X.Z.; Cheng, D.Z.; Zhang, Z.; Lin, S.; Dai, J.F. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the ICCV, Seoul, Korea, 27 October–2 November 2019; pp. 6687–6696.
36. Li, K.; Wang, G.; Cheng, G.; Meng, L.Q.; Han, J.W. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *arXiv* **2019**, arXiv:1909.00133. Available online: https://arxiv.org/abs/1909.00133v2 (accessed on 19 October 2021). [CrossRef]
37. Zhang, Y.L.; Yuan, Y.; Feng, Y.C.; Lu, X.Q. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
38. Gong, C.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *12*, 7405–7415.
39. Sun, P.Z.; Zhang, R.F.; Jiang, Y.; Kong, T.; Xu, C.F.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.H.; Wang, C.H.; et al. Sparse R-CNN: End-to-End Object Detection with Learnable Proposals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 14454–14463.
40. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
41. Robbins, H.; Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [CrossRef]
42. Versaci, M.; Calcagno, S.; Morabito, F.C. Fuzzy Geometrical Approach Based on Unit Hyper-Cubes for Image Contrast Enhancement. In Proceedings of 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, Malaysia, 19–21 October 2015; pp. 488–493.