applied sciences

MDPI

Article

Query-Based Cascade Instance Segmentation Network for Remote Sensing Image Processing

Enping Chen [1], Maojun Li [1], Qian Zhang [1] and Man Chen [1,2,*]

[1] School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China; 21205050959@stu.csust.edu.cn (E.C.); limj@csust.edu.cn (M.L.); qianqian@stu.csust.edu.cn (Q.Z.)
[2] College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China
* Correspondence: 19205060770@stu.csust.edu.cn

Abstract: Instance segmentation (IS) of remote sensing (RS) images can not only determine object location at the box-level but also provide instance masks at the pixel-level. It plays an important role in many fields, such as ocean monitoring, urban management, and resource planning. Compared with natural images, RS images usually pose many challenges, such as background clutter, significant changes in object size, and complex instance shapes. To this end, we propose a query-based RS image cascade IS network (QCIS-Net). The network mainly includes key components, such as the efficient feature extraction (EFE) module, multistage cascade task (MSCT) head, and joint loss function, which can characterize the location and visual information of instances in RS images through efficient queries. Among them, the EFE module combines global information from the Transformer architecture to solve the problem of long-term dependencies in visual space. The MSCT head uses a dynamic convolution kernel based on the query representation to focus on the region of interest, which facilitates the association between detection and segmentation tasks through a multistage structural design that benefits both tasks. The elaborately designed joint loss function and the use of the transfer-learning technique based on a well-known dataset (MS COCO) can guide the QCIS-Net in training and generating the final instance mask. Experimental results show that the well-designed components of the proposed method have a positive impact on the RS image instance segmentation task. It achieves mask average precision (AP) values of 75.2% and 73.3% on the SAR ship detection dataset (SSDD) and Northwestern Polytechnical University Very-High-Resolution dataset (NWPU-VHR-10 dataset), outperforming the other competitive models. The method proposed in this paper can enhance the practical application efficiency of RS images.

Keywords: remote sensing images; query; transformer; cascade network; instance segmentation

check for updates

Citation: Chen, E.; Li, M.; Zhang, Q.; Chen, M. Query-Based Cascade Instance Segmentation Network for Remote Sensing Image Processing. *Appl. Sci.* **2023**, *13*, 9704. https://doi.org/10.3390/app13179704

Academic Editor: Yutaka Ishibashi

Received: 1 August 2023
Revised: 20 August 2023
Accepted: 25 August 2023
Published: 28 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

All countries are paying great attention to urban land planning, agricultural management, and marine environmental monitoring, which are related to economic and military development. Additionally, they also have significant impacts on the ecological environment of the earth [1–3]. Thus, earth observation technology is of vital importance. Remote sensing (RS) can dramatically increase the scope and quality of earth observation, which provides a certain degree of convenience for researchers to observe the characteristics of objects on the earth with a high resolution. In the field of RS, extracting high-precision objects from optical and synthetic aperture radar (SAR) images has become an enormous challenge and has attracted widespread attention.

Benefiting from the advancement of neural networks, increasing numbers of researchers are applying deep learning methods to RS image processing [4,5]. In general, RS image processing can be divided into several tasks. The first is an image-level task, where different objects are categorized into their respective classes based on their attributes and

features [6,7]. The second task is an object-level task, which locates the specific position of each object in the picture and separates the region of interest (RoI) from the background region using bounding boxes [8–10]. Finally, in the pixel-level task, semantic-segmentation-based methods separate the background from predicted objects on a pixel-by-pixel basis, but they are unable to distinguish different objects within the same category [11,12]. Object-detection and semantic-segmentation methods are also utilized to interpret images in RS image processing [13–15]. By performing instance-level localization and pixel-level classification of objects of interest simultaneously, instance segmentation (IS) methods can combine the benefits of object-level methods and semantic-segmentation methods to interpret more comprehensive information. Due to the dense and compact distribution of objects in SAR and optical images, we adopt IS to interpret objects at the pixel-level and distinguish individual objects within the same category.

IS can be divided into one-stage-segmentation methods [16–18] and two-stage-segmentation methods [19–21]. One-stage IS algorithms directly extract features from the input image and output results in a single forward inference process, which typically results in a faster inference speed. Although the one-stage method has good performance in real-time detection, it is still slightly insufficient for the high-precision segmentation requirements of high-resolution RS images. On the other hand, two-stage IS algorithms are suitable for scenarios that require a high segmentation accuracy because they can locate and segment objects more accurately. However, their processing speed tends to be relatively slow. The mask region-based convolutional neural network (Mask-RCNN) [21] is the most representative algorithm among two-stage-segmentation paradigms. On the basis of the Faster-RCNN [9], a fully convolutional network (FCN) [11] is connected in parallel to the segmentation head (the head usually refers to the network layer that predicts the output in deep learning) obtained by the mask instance. It often serves as a baseline in numerous IS comparative experiments. Owing to its excellent segmentation performance, there are also many IS algorithms in the field of RS that are improved on it [22–24] and have achieved good segmentation results. The choice of the method depends on the specific application scenario and the desired level of accuracy and speed for IS.

Recent years have seen numerous innovations in IS using RS images. The cascade network was first proposed by Cai et al. [25] and is extensively used by researchers in detection and segmentation tasks [19,20]. The structure consists of multiple stages, and the output information of each level is passed to the next level as input information, realizing the level-by-level transfer and processing of information. Inspired by the cascade design pattern in object-level tasks, Su et al. [26] proposed a high-quality IS network (HQ-ISNet). This method combines a high-resolution feature pyramid network (FPN) [27] with a cascade network, making full use of multilevel feature maps effectively to interpret complex backgrounds in RS images. Subsequently, Zeng et al. [28] proposed the consistent proposals of IS network (CPISNet) to improve the feature-extraction method and integrated a cascade network with residual connections into their network to further improve the accuracy of aerial image mask prediction. The development of the attention mechanism [29,30] has had a great impact on RS image IS. The attention mechanism is an important approach in image processing which can automatically select the most relevant or important parts of the input data by learning weights. It allows the model to focus on specific regions or features, emphasizing areas that contain crucial information, thereby improving recognition accuracy. Zhao et al. [31] proposed a SAR ship IS method based on collaborative attention mechanisms which combines the advantages of multiple attention mechanisms and can effectively extract the instance masks of objects in the image. Fang et al. [32] introduced a contour-refinement network that works from coarse to fine. They utilized an attention-based feature pyramid subnetwork to enhance the prediction capability for small objects. The contour extracted from the coarse branch was further refined through the fine branch by learning edge features to achieve fine-grained building contours. Sun et al. [33] proposed a multiscale feature pyramid network which

applied the attention mechanism to multiscale features to improve the performance of multiscale ship detection in high-resolution SAR images. Unfortunately, when some attention mechanisms overemphasize certain local features, other critical global features may be neglected, leading to the loss of important details, and thus negatively affecting the RS image-segmentation outcome.

However, the aforementioned methods are two-stage approaches based on the regional proposal network (RPN) [9]. Although employing this trick yields good results in detection accuracy, it still encounters some limitations, such as the need to predefine different aspect ratios and proportions for the anchor boxes, which requires prior knowledge or experience and lacks flexibility and versatility. To address these constraints, several empirical studies have shown that the algorithm can be improved through a sparse detection paradigm [8,34]. In addition, several anchor boxes designed for objects of different sizes often require manual adjustments, which may not be adaptable to RS images with complex backgrounds and varying sizes.

Following the successive proposals of many effective RS image-segmentation methods, some researchers have found that the quality of the segmentation depends not only on the algorithms themselves but also on the quality of the dataset's annotation. Cheng et al. [35] produced the Northwestern Polytechnical University Very-High-Resolution dataset (NWPU-VHR-10 dataset), which consists of very-high-resolution images. Su et al. [26] expanded the NWPU-VHR-10 dataset by providing pixel-level IS annotations. Building on this work, the SAR ship detection dataset (SSDD) was also extended with closed polygon annotations for pixel-level ship interpretation, overcoming the limitations of existing SAR datasets. Wei et al. [36] created a high-resolution SAR image dataset (HRSID), which is particularly useful for ship detection and the IS of SAR images. Although RS image datasets are already so rich, there are still great difficulties in having a large amount of high-quality data, which often requires significant manual costs. The application of transfer-learning techniques [37,38] substantially alleviates the limitations of training difficulties due to a lack of training data.

As mentioned above, compared with natural images, optical and SAR images are characterized by cluttered backgrounds, large differences in the size of objects, and complex shapes of instances caused by factors such as the capturing distance and orientation (as shown in Figure 1), and thus their segmentation effects are limited. To effectively address this, this paper proposes a query-based cascade instance segmentation network (QCIS-Net), a RS image cascade instance segmentation network based on the efficient query mechanism. The network uniformly represents the location and visual information of instances in the RS image through queries. It mainly includes key components such as the efficient feature extraction (EFE) module, the multistage cascade task (MSCT) head, and the joint loss function. The EFE module can combine the global information modeling capability of the Transformer [39] to solve the long-term dependency problem in visual space. The dynamic detection and segmentation head in the MSCT head uses a dynamic convolution kernel based on the query representation to concentrate on the part of the image that is of interest. It facilitates the detection and segmentation of task associations through a multistage structural design, allowing both tasks to benefit from each other and gradually improving the segmentation quality of the mask. The elaborated joint loss function can guide QCIS-Net for training and generate the final instance mask.

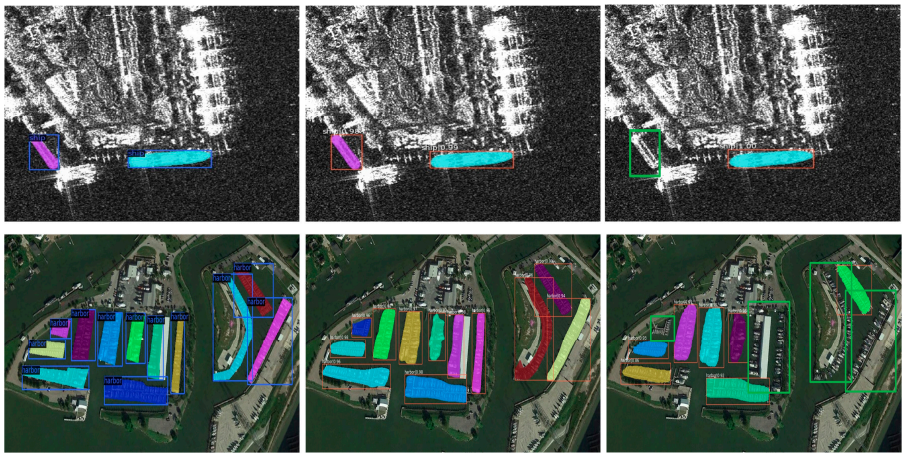

Figure 1. From left to right, the schematic shows the ground-truth, QCIS-Net, and Mask-RCNN results for optical and SAR images, where the images have cluttered backgrounds or complex instance shapes. We use green boxes to represent missed detections. The colors of all instances in the figure are randomly generated to distinguish each instance.

2. Methods

2.1. Overview

Figure 2 provides an overview of the QCIS-Net framework. The EFE module can combine the global information-building module capability of Transformer to solve the long-term dependence problem in visual space. The MSCT head uses the dynamic convolution kernel and cascade structure to allow the detection and segmentation tasks to take advantage of each other. The elaborated joint loss function can guide QCIS-Net for training and generating the final instance mask.

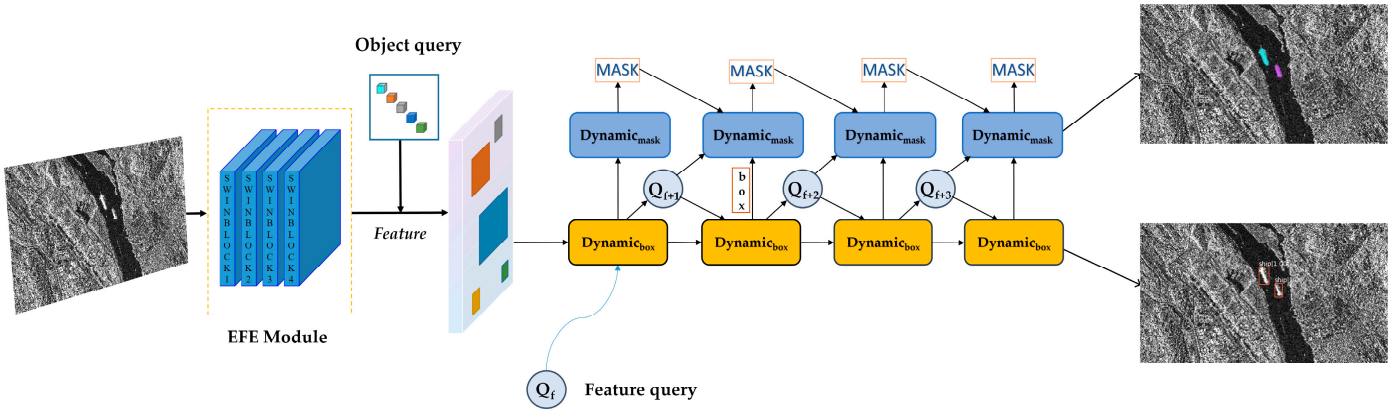

Figure 2. The network architecture of QCIS-Net.

2.2. EFE Module

Most people are accustomed to using CNN-based methods for feature extraction, such as the visual geometry group (VGG) [40] network, which stacks multiple layers of convolutional networks to produce feature images. The residual network (ResNet) [6] introduces residual connections while deepening the network to obtain features with different layers. In view of the small proportion and scattered distribution of off-shore ships in SAR images and of aircraft and storage tanks in optical images, traditional methods may fail to capture global information, connect well with context, and accurately locate objects, leading to large deviations in predicted results. Therefore, this paper introduces the efficient feature extraction (EFE) module. The key part of our EFE module is the Swin (Shift window) Transformer feature extractor, which has demonstrated excellent

performance in many papers. First, compared with the calculation redundancy of large parameters in the Transformer, by replacing multihead self-attention (MSA) [30] with windows multihead self-attention (W-MSA) and shifted windows multihead self-attention (SW-MSA), the experimental calculation cost is greatly reduced [41]. Equations (1) and (2) show the computational complexity of the three types of attention, respectively. Next, the nonoverlapping self-attentive mechanism in the Swin Transformer (ST) can quickly capture long-term dependencies between objects and focus on both local features and global information. In the end, the hierarchical construction of the ST endows it with outstanding feature-representation capabilities at different scales, thereby significantly enhancing the model's performance in IS tasks and demonstrating excellent segmentation results.

$$O(MSA) = 4hwC^2 + 2(hw)^2C \tag{1}$$

$$O(W-MSA, SW-MSA) = 4hwC^2 + 2K^2hwC \tag{2}$$

where h represents the height of the feature map, w represents the width of the feature map, C represents the number of channels, K represents the size of each window, and O represents the computational complexity.

As shown in Figure 3, in order to facilitate the understanding of the transformation process, we use A, B, and C to denote the positional relationships of different patches in the transformation step. Correspondingly, the positional relationships of other patches can also be easily found. Intuitively, SW-MSA will slide all windows as a whole to the right and down by $m(= h/4 = w/4)$ pixels along the horizontal and vertical axes, forming new window combinations. The new windows generated use a cross-window connection strategy to compute their self-attention weights, so that the feature units can interact with each other in different windows, allowing the network to capture more contextual information.

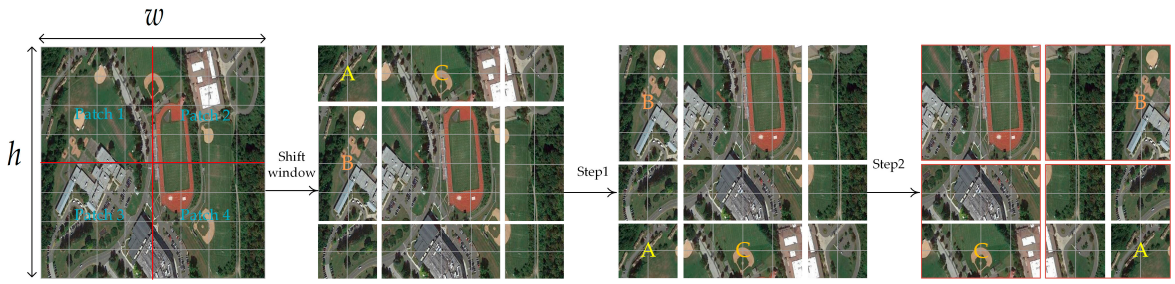

Figure 3. The SW-MSA shifting window process.

As shown in Figure 4, the Swin Transformer has four stages, and the outputs of stages 1 to 4 are $H/4 \times H/4 \times C$, $H/8 \times W/8 \times 2C$, $H/16 \times W/16 \times 4C$, and $H/32 \times W/32 \times 8C$, respectively, where H and W are the height and width of the input image. Some connected Swin Transformer blocks are used in each stage to establish feature correlation. Firstly, the optical or SAR images are fed into the network, partitioned into nonoverlapping patches, and then mapped into a C-dimensional vector using embedding. Then, after two consecutive Swin blocks are used to establish feature correlation, the output dimension is $H/4 \times W/4 \times C$. Next, three consecutive identical stages consisting of patch merging and some consecutive Swin blocks are passed. Finally, the output dimension is the feature map of $H/32 \times W/32 \times 8\ C$.

2.3. Query

The RPN-based prediction model has become the most frequently used effective method for extracting anchor boxes in various detection algorithms. However, the large number of redundant boxes it generates imposes a huge burden on the subsequent post-processing tasks. To alleviate this burden, query, an information factor that can circulate in the network, is introduced to make the detection process dense to sparse.

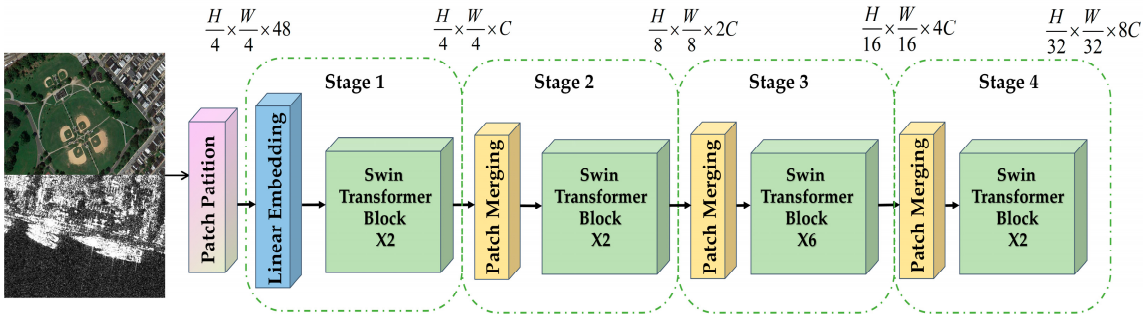

Figure 4. The Swin Transformer architecture.

In this paper, we categorized the queries into two parts based on their different functions, namely the object query and the feature query. The object query can locate the coordinate information, while the feature query represents the visual features, such as texture and shape. The design concept of the object query is derived from [8,42], which are used to represent the positional information. Specifically, the object query is set to a set of learnable proposal boxes ($G \times 4$), where G is the number of preset objects and 4 represents the coordinate information of the object bounding box. Namely, the center point coordinates (x and y coordinates) of the box, and the height and width of the box. During the training period, the coordinate information of the bounding box will be updated using back propagation to finally obtain the prediction result. Considering the significant role of visual features such as texture and shape in representing the object, we also introduce the feature query to encode the visual feature information of the object. The shape of the feature query is $[G, v]$, where G is the predetermined number of objects, which is the same as G in the object query, and v represents a high-dimensional vector (usually 256-dimensional) used to encode the visual feature information of objects. Therefore, the QCIS-Net not only uses object queries to obtain rough location information of the object, but also combines feature queries to further represent the visual feature information of the object. This helps to accurately detect objects in images. In addition, this query-based approach generates less redundancy and does not require complex postprocessing work.

2.4. Multistage Cascade Task Head

In SAR images, the edges of inshore ships often appear blurry and are susceptible to background factors, such as coastlines and town buildings, which can affect the segmentation process. Similarly, in optical images, distinguishing objects such as aircraft and runways, baseball stadiums, and surrounding vegetation from the background can be difficult due to their high similarity, leading to lower quality edge segmentation.

To overcome these limitations, in our work, we have adopted the cascade structure of the detection heads mentioned in Section 1. This cascading transmission enables the network to gradually extract and learn more abstract and high-level features, excelling in handling challenging tasks. But, unlike that, we replace the original detection head with a dynamic detection head and a dynamic segmentation head. This dynamic head design is motivated by dynamic convolution [43].

2.4.1. Dynamic Detection Head

As shown in Figure 5, we initially randomly generate a predefined number of object queries on the multilevel feature map F extracted by the feature extractor, and then the object queries are cropped to an RoI feature map of 7×7 size at a uniform scale by the region-of-interest pooling (RoI pooling) [9] operation, which is used as the input. Subsequently, we dynamically interact the feature query with the input RoIs after dimensional reorganization. The purpose is to make the feature query efficiently learn the feature information in each RoI, and the parameter update is performed one-to-one for each instance, which demonstrates excellent flexibility.

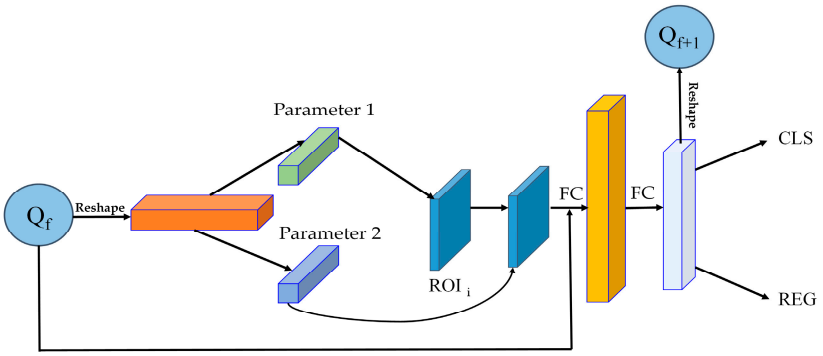

Figure 5. The internal structure of the dynamic box head.

In this process, the feature query updates the perceptual ability of the corresponding instance, including spatial location, color, edges, and other information through self-learning. We apply rectified linear unit (ReLU) activation [44] with layer normalization regularization [45] to the feature map after completing the interaction to enhance its representational ability and generalization performance. At the same time, we fuse its elemental features with the initial feature query through residual concatenation. The reason is that this preserves the original information from the previous feature query and avoids the loss of original features. On the one hand, to minimize the computational complexity, we perform the mapping of output features from low-level features to high-level features with more abstract features using only two consecutive fully-connected layers, and then classify and regress the high-level features to obtain the final detection box. On the other hand, newer feature queries are converted to original dimensions by reshaping operations and output to the next stage for a new round of learning.

2.4.2. Dynamic Mask Head

As shown in Figure 6, similar to the dynamic box detection head, we simply changed to a more fine-grained mask segmentation after completing the instance interaction. Notably, compared to the detection task, the segmentation task requires more accurate pixel-location information. In the segmentation phase, the RoI is obtained through the RoI Align [21] operation and cropped into a feature map with a size of 14 × 14. Specifically, the output feature after the instance interaction is completed is subjected to four 3 × 3 convolution blocks to obtain richer semantic features, which are returned to the original image size by an upsampling operation (Transposed convolution [46]). In addition, we perform a 1 × 1 convolution on the output of the feature map before upsampling. This step ensures its alignment with the next stage and connects it to the next mask prediction through elementwise summation. The aim is to enhance the flow of mask information between the different stages. Eventually, the predicted mask information is matched with the real label for loss calculation.

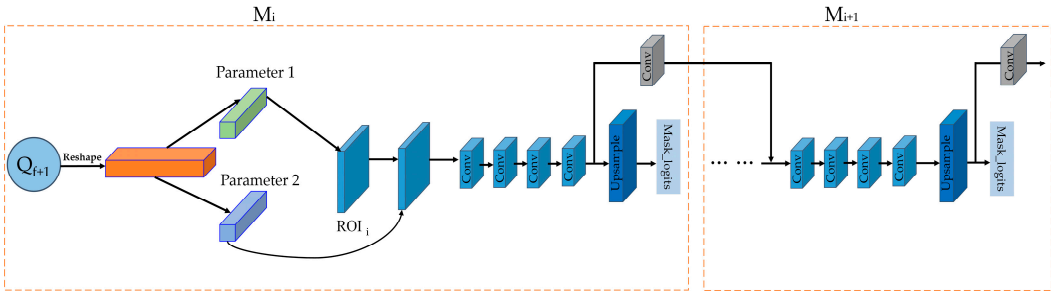

Figure 6. The internal structure of the dynamic mask head.

Figure 7 shows the diagram of the multistage cascade task head. Assuming that the EFE module outputs a multilevel feature map of F_i, i = 1, 2, 3, 4, all object queries are

pooled within the input feature map F_i by RoI pooling, and the detection process can be represented as:

$$R_{box} = P_{box}(F_i, Q_o) \tag{3}$$

$$Q_{f+1}, T_{box} = Dynamic_{box}(R_{box}, Q_f) \tag{4}$$

$$R_{mask} = A_{box}(F_i, T_{box}) \tag{5}$$

$$T_{mask} = Dynamic_{mask}(R_{mask}, Q_{f+1}) + M_i \tag{6}$$

where P_{box} and A_{box} denote RoI pooling and RoI Align, respectively. R_{box} and R_{mask} denote the RoI after RoI Pooling and RoI Align, respectively. T_{mask} and T_{box} denote the prediction mask and box prediction, respectively. Q_f denotes the feature query and Q_o denotes the object query. M_i denotes that the mask information of the current stage is passed to the next layer.

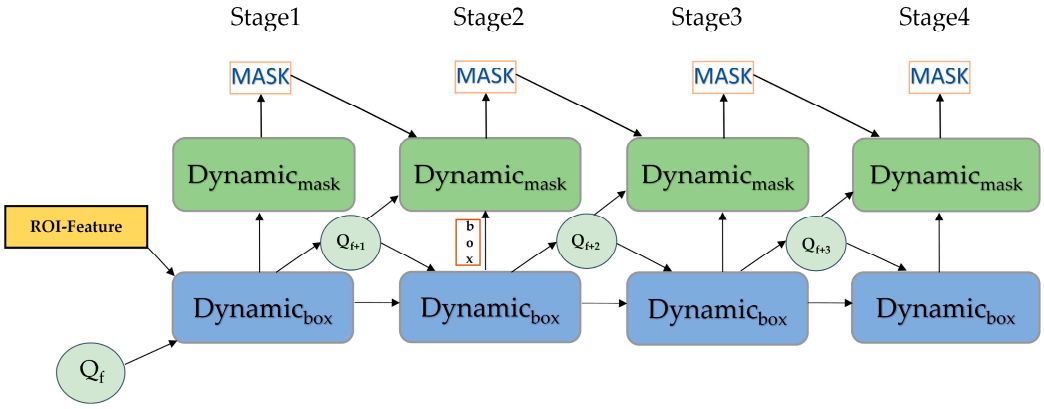

Figure 7. Multistage cascade task head schematic.

This design mode aims to ensure that the query feature of the next stage is closely related to the current stage. In this way, the model is able to gradually integrate the information from different layers to improve the understanding and expression of the goal. Unlike an ordinary cascade head, our task head can dynamically improve the quality of the mask with the flow of information and generate the final mask without any postprocessing.

2.5. Joint Loss Function and Transfer Learning

To enable the network to be trained efficiently on the dataset, we designed a joint loss function and introduced the transfer-learning technique. Transfer learning is mainly applied to initialize networks via fully pretrained networks on the famous MS COCO [47] dataset, and then the initialized network is directly applied to train the optical and SAR image datasets to enhance the network's learning efficiency and effectiveness. The loss function consists of four aspects during the training phase:

$$L_{total} = L_{mask} + L_{reg} + L_{cls} + L_{giou} \tag{7}$$

where L_{mask}, L_{reg}, L_{cls}, and L_{giou} represent the segmentation loss, the regression loss, the classification loss, and the intersection-over-union (IoU) loss, respectively.

To overcome the problem of unbalanced samples, the gradient-harmonizing mechanism [48] was used for both classification and regression losses. It reduces the weights of easy samples (the images that are easily distinguished by the model) and outliers, and the

dynamic properties of its loss make training more effective and robust. The L_{reg} and L_{cls} are defined as (8) and (9):

$$\begin{aligned} L_{\text{reg}} &= \frac{1}{M}\sum_{i=1}^{M}\lambda_i ASL_1(d_i) \\ &= \sum_{i=1}^{M}\frac{ASL(d_i)}{GD(gr_i)} \end{aligned} \tag{8}$$

where GD is the designed gradient density function and ASL_1 is the modified form of smooth-L_1.

$$\begin{aligned} L_{cls} &= \frac{1}{M}\sum_{i=1}^{M}\lambda_i * L_{CE} \\ &= \sum_{i=1}^{M}\frac{L_{CE}}{GD(g_i)} \end{aligned} \tag{9}$$

in which

$$L_{CE} = \begin{cases} -\log(p), & if\ p* = 1 \\ -\log(1-p), & if\ p* = 0 \end{cases} \tag{10}$$

where M represents the overall sample count and λ_i can be regarded as the loss weight of the i-th sample; p represents the probability distribution predicted by the model, and $p*$ is the true label. The IoU loss L_{giou} is defined as [49]:

$$L_{giou} = 1 - IoU + \frac{A^c - U}{A^c} \tag{11}$$

where U represents the union between the ground truth and the predicted bounding box, and A^c represents the area of the smallest enclosing box. The L_{mask} can be represented as Equation (12), which is the dice loss form in the V-Net [50].

$$L_{mask} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{12}$$

Similarly, the operator $\cap$ represents the elementwise dot product, and the operator $||$ represents the numerical square. Therefore, Equation (12) can be formulated as:

$$L_{mask} = 1 - \frac{2\sum_{pix} Y_T \cdot Y_F}{\sum_{pix} {Y_T}^2 + \sum_{pix} {Y_F}^2} \tag{13}$$

where $Y_T \cdot Y_F$ is the dot product between the pixels of the ground-truth label image and the pixels of the predicted result.

3. Experiment and Analysis

3.1. Implementation Details

The QCIS-Net algorithm and other experiments were trained and tested on the PyTorch framework using an NVIDIA Tesla V100 GPU. The QCIS-Net was trained for 50 epochs with an initial learning rate of 1×10^{-4} based on the pretrained weights of the MS COCO dataset, which is reduced by a factor of 10 at the 40th epoch. The Adam optimizer with a momentum of 0.9 was used to train the network. Data augmentation, such as left–right flip, was used during the training phase, while no data augmentation was applied during the inference phase. It is worth noting that, during the experiment, the long and short edges of the image were randomly adjusted to be between 800 and 1333 for both training and testing.

3.2. Datasets

NWPU-VHR-10 dataset: This dataset [26] consists of 800 labeled high-resolution optical RS images, including 650 images with objects and 150 images with pure backgrounds.

The dataset comprises a total of 10 categories. We randomly selected 70% of the images with objects to be the training set and 30% of the images to be the test set.

SSDD dataset: The dataset [51] consists of SAR images that contain 1160 images. The dataset includes many small- and medium-sized objects, but relatively few large objects. In the experiment, we randomly allocated the training set and test set according to a 7:3 ratio. The training set contains 812 images, while the test set contains 348 images.

3.3. Evaluation Metrics

The standard MS COCO [47] evaluation metric is used to evaluate the quantitative IS results. The IoU of each prediction result is defined as $IoU = \frac{P_{mask} \cap G_{mask}}{P_{mask} \cup G_{mask}}$, which is calculated based on the intersection over union of the prediction result and the ground-truth result. Setting an a priori IoU threshold criteria, precision and recall values are then calculated accordingly:

$$Precision = \frac{TP}{TP + FP}, \tag{14}$$

$$Recall = \frac{TP}{TP + FN}. \tag{15}$$

The experiments use the standard MS COCO metric to quantitatively and comprehensively evaluate the performance of the IS methods in the RS images. These metrics include the average precision AP, AP_{50}, AP_{75}, AP_s, AP_m, and AP_L. The average precision (AP) for a given IoU threshold can be gained by:

$$AP_{IoU} = \int_0^1 P(r)dr, \tag{16}$$

where r represents the recall and $P(r)$ represents the precision corresponding to the recall. Thus, AP_{50} and AP_{75} denote the results of the above equations for the IoU thresholds of 0.5 and 0.75, respectively. The AP value represents the average of the IoU for 10 thresholds, 0.50–0.95 (steps of 0.05), and can be written as:

$$AP = \frac{1}{10} \sum_{IoU=0.5}^{0.95} AP_{IoU}. \tag{17}$$

In addition, AP_S, AP_M, and AP_L mean to evaluate the model's mask prediction performance on small (<32^2 pixels), medium (>32^2 pixels), and large (<96^2 pixels) objects. These metrics provide a comprehensive assessment of the mask-prediction quality of the model.

3.4. Ablation Experiments

We performed ablation experiments where certain components of the algorithm were removed or replaced to measure the impact of these components on the overall performance. To verify the performance of the Swin Transformer for the network, we replaced the backbone of the network with the mainstream ResNet101 for comparison and designed the following ablation experiments. In addition, we also conducted experiments on the use of transfer learning. The results are shown in Tables 1 and 2.

Table 1. Effect of the EEF module.

Dataset	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
NWPU-VHR-10 [26]	ResNet101	67.3	88.5	77.1	53.8	56.1	73.9
	Swin Transformer	73.3	93.0	75.7	56.6	67.2	79.8
SSDD [51]	ResNet101	74.2	97.1	92.9	72.2	80.5	85.0
	Swin Transformer	75.2	98.7	92.6	73.1	81.8	75.5

Table 2. Effect of transfer learning on MS COCO.

Dataset	Backbone	MS COCO	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
NWPU-VHR-10 [26]	ResNet101		67.3	88.5	77.1	53.8	56.1	73.9
		√	53.4	75.0	58.1	36.4	52.3	66.7
	SwinT-Large		73.3	93.0	75.7	56.6	67.2	79.8
		√	54.6	76.5	59.5	35.4	52.9	70.0
SSDD [51]	ResNet101		74.2	97.1	92.9	72.2	80.5	85.0
		√	59.1	85.7	74.6	56.7	65.8	75.3
	SwinT-Large		75.2	98.7	92.6	73.1	81.8	75.5
		√	60.2	86.3	75.9	58.1	65.9	70.6

As shown in Table 1, compared to the stacked convolutional layers, the Swin Transformer as the backbone achieved a 6.0% AP improvement over ResNet-101 on the NWPU-VHR-10 dataset and a 1.0% AP improvement on the SSDD. This indicates that this attention-based backbone network has good adaptability to RS images and can effectively extract the features of the objects in the input images. Table 2 shows that transfer learning on the MS COCO dataset significantly improved the IS performance of QCIS-Net, which is due to the difficulty of training the cascade structure of QCIS-Net itself. The use of transfer learning based on large datasets allows better initialization of the network, which is beneficial for fine-tuning the optical and SAR image datasets.

From Table 3, we conducted an ablation study on the number of cascade stages in the MSCT head in datasets. Without any cascading stages, the network's segmentation and detection performance unsurprisingly achieved the worst results. As the number of cascade stages in the MSCT head increased, both the segmentation and detection performance of the network improved. However, when the number of cascade stages became too large, the performance improvement was not significant. When we chose 4 or 6 as the number of cascade stages, the performance was good, and the difference was insignificant. Therefore, we chose 6 as the value of cascade number for optical images, and for SAR images, we chose 4 as the value.

Table 3. The influence of stage numbers on the MSTC head.

	NWPU-VHR-10 [26]		SSDD [51]	
Stage	AP_{Seg}	AP_{box}	AP_{Seg}	AP_{box}
No cascade	35.6	22.3	46.1	31.8
2	63.2	64.1	65.6	68.0
4	72.7	76.6	75.2	79.1
6	73.3	77.1	75.1	79.1
8	71.9	76.5	70.8	74.1

From Table 4, we experimented with different numbers of object queries. Increasing the number of object queries achieved higher AP values in both datasets. As for the results in the SSDD, the number of 300 achieves a salient AP performance compared to the rest of the situations. In the counterpart results for the NWPU-VHR-10 dataset, the relative increase in AP values was not quite so dramatic. However, increasing the number of object queries can make the model training extremely time-consuming.

3.5. Ship Segmentation Results of the SSDD

To validate the performance of QCIS-Net, we compared it with several popular IS algorithms, including the Mask R-CNN [21], Mask Scoring R-CNN [52], Cascade Mask R-CNN [19], SC-Net [53], HTC [20], and Insta-Boost [54]. These algorithms include both

one-step and multistage detection methods. For both training and testing, we followed the default hyperparameters in MMDetection 2.0 [55], using ResNet-101 and FPN [27] as the backbone. Additionally, we use Giga floating point operations (GFLOPs) to estimate the model's complexity and computational requirements. The experimental results are presented in Tables 5 and 6.

Table 4. The influence of the number of object queries.

	NWPU-VHR-10 [26]		SSDD [51]	
Num	**AP_{Seg}**	**AP_{box}**	**AP_{Seg}**	**AP_{box}**
100	72.0	75.6	71.5	74.3
200	72.6	76.5	71.7	74.5
300	73.3	77.1	75.1	79.1

Table 5. IS results on the SSDD test set.

Method	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**	**GFLOPs**
Mask-RCNN [21]	65.8	94.8	82.1	66.4	64.8	45.9	134.43
MS-RCNN [52]	66.2	94.6	83.2	66.4	65.5	50.2	134.43
CM-RCNN [19]	66.1	93.7	82.7	66.4	66.0	45.3	265.54
HTC [20]	66.4	94.6	83.3	66.7	65.9	60.7	268.13
SC-Net [53]	67.0	95.6	83.9	67.1	67.5	46.1	348.80
Insta-Boost [54]	64.4	94.3	80.6	64.6	64.1	40.2	134.43
QCIS-Net	75.2	98.7	92.6	73.1	81.8	75.5	221.13

Table 6. Object-detection results on the SSDD test set.

Method	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**	**Time**
Mask-RCNN [21]	69.4	96.5	83.6	69.9	69.1	43.2	50.4 ms
MS-RCNN [52]	69.9	96.5	85.6	70.3	69.8	36.8	51.5 ms
CM-RCNN [19]	70.9	95.7	86.5	70.9	72.6	50.4	66.6 ms
HTC [20]	71.5	96.6	85.2	71.4	72.9	53.4	66.9 ms
SC-Net [53]	71.1	96.6	85.6	71.4	71.3	45.1	73.8 ms
Insta-Boost [54]	68.1	96.3	81.0	68.3	69.0	35.6	62.7 ms
QCIS-Net	79.1	98.7	93.6	78.1	85.1	80.1	55.3 ms

Tables 5 and 6 present the performance of QCIS-Net in IS and object detection. Compared to the classical Mask-RCNN algorithm, QCIS-Net achieves a 9.4% and 9.7% improvement in segmentation and detection AP, respectively, demonstrating the outstanding performance of the proposed method in SAR image-segmentation and detection tasks. Compared to other IS methods, such as MS-RCNN, HTC, and SC-Net, QCIS-Net also demonstrates certain advantages in segmentation and detection performance. This further highlights the powerful capabilities of our method in handling SAR image-segmentation and detection tasks. In summary, the proposed QCIS-Net exhibits excellent performance in SAR image IS and object-detection tasks due to the elaborated EFE module, MSCT head, joint loss function, and transfer-learning training scheme. It outperforms various methods used for comparison with an absolute advantage.

Figure 8 visualizes the result of QCIS-Net compared to the other IS methods on the SSDD. The first to third columns are the SAR images of the inshore, and the fourth column

is SAR images of the offshore. Specifically, in the scene in the first column, there are unfavorable factors, such as the dense arrangement of ships and mutual occlusion, which leads to the common phenomenon of missed detections and false detections in the IS comparison method. However, our proposed QCIS-Net demonstrates the ability to detect and segment ships in this scene to a significant extent, with segmented masks closely resembling the ground truth. This demonstrates the powerful detection and segmentation performance of QCIS-Net in SAR images. In the second column, the background contains objects that resemble ship appearances, making them prone to be easily and incorrectly detected as ships. In this scene, the comparison methods have exhibited varying degrees of missed detections and false detections. However, QCIS-Net effectively discriminates ships from the background, avoiding false detections. In the third column, although all comparison methods are capable of segmenting ships in the images, the obtained instance masks differ significantly from the ground truth. QCIS-Net overcomes the drawback of low mask quality and accurately detects object ships, verifying the superiority of our network. The fourth column reflects the performance of various algorithms on offshore SAR images, where the proportion of pixels occupied by the object is small, and there are many speckle noises, posing challenges for detection and segmentation. The proposed QCIS-Net can accurately detect and segment ships in the image, further proving its performance in SAR image IS. Overall, the proposed QCIS-Net exhibits good performance in SAR image IS and object-detection tasks, capable of handling various challenging scenarios, such as dense ship arrangements, occlusions, and a small proportion of object-occupied pixels.

3.6. Instance Segmentation Results on NWPU VHR-10

The experiments follow the same several most popular IS algorithms as in Section 3.5, and the training hyperparameters are configured with the same default values as in the experiments. Different from the SSDD, NWPU-VHR-10 is a multicategory dataset that includes 10 categories, including airplane (AI), baseball diamond (BD), ground track field (GTF), vehicle (VC), ship (SH), tennis court (TC), harbor (HB), storage tank (ST), basketball court (BC), and bridge (BR). Therefore, we conducted experimental evaluations for each category. The results are shown in Tables 7–9.

Table 7. IS on the NWPU-VHR-10 test set.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask-RCNN [21]	66.4	89.2	73.8	50.9	64.2	75.3
MS-RCNN [52]	67.5	89.7	74.9	51.0	65.3	76.0
CM-RCNN [19]	66.6	89.7	65.0	37.3	58.5	69.2
HTC [20]	67.7	90.5	74.2	53.7	65.7	68.9
SC-Net [53]	68.4	92.4	74.6	53.5	66.7	76.6
Insta-Boost [54]	68.3	92.0	81.4	67.3	69.2	68.3
QCIS-Net	73.3	93.0	75.7	56.6	67.2	79.8

Table 8. Classwise IS results on the NWPU-VHR-10 test set.

Method	AI	BD	GTF	VC	SH	TC	HB	ST	BC	BR
Mask-RCNN [21]	48.5	80.5	94.2	54.4	54.2	79.2	53.2	81.8	77.4	37.7
MS-RCNN [52]	49.5	83.3	94.9	54.4	57.5	79.7	54.6	82.6	79.5	38.7
CM-RCNN [19]	48.9	80.5	95.8	54.7	55.9	78.7	53.2	82.6	81.5	34.0
HTC [20]	50.2	83.6	95.4	54.7	54.5	82.3	54.4	82.7	82.2	36.6
SC-Net [53]	52.5	84.0	95.0	55.5	59.4	82.6	54.7	82.6	82.5	36.0
Insta-Boost [54]	50.8	84.0	94.2	56.5	56.2	83.3	53.7	84.0	81.5	37.0
QCIS-Net	56.5	86.5	96.0	63.2	63.2	83.8	62.0	83.6	84.6	47.8

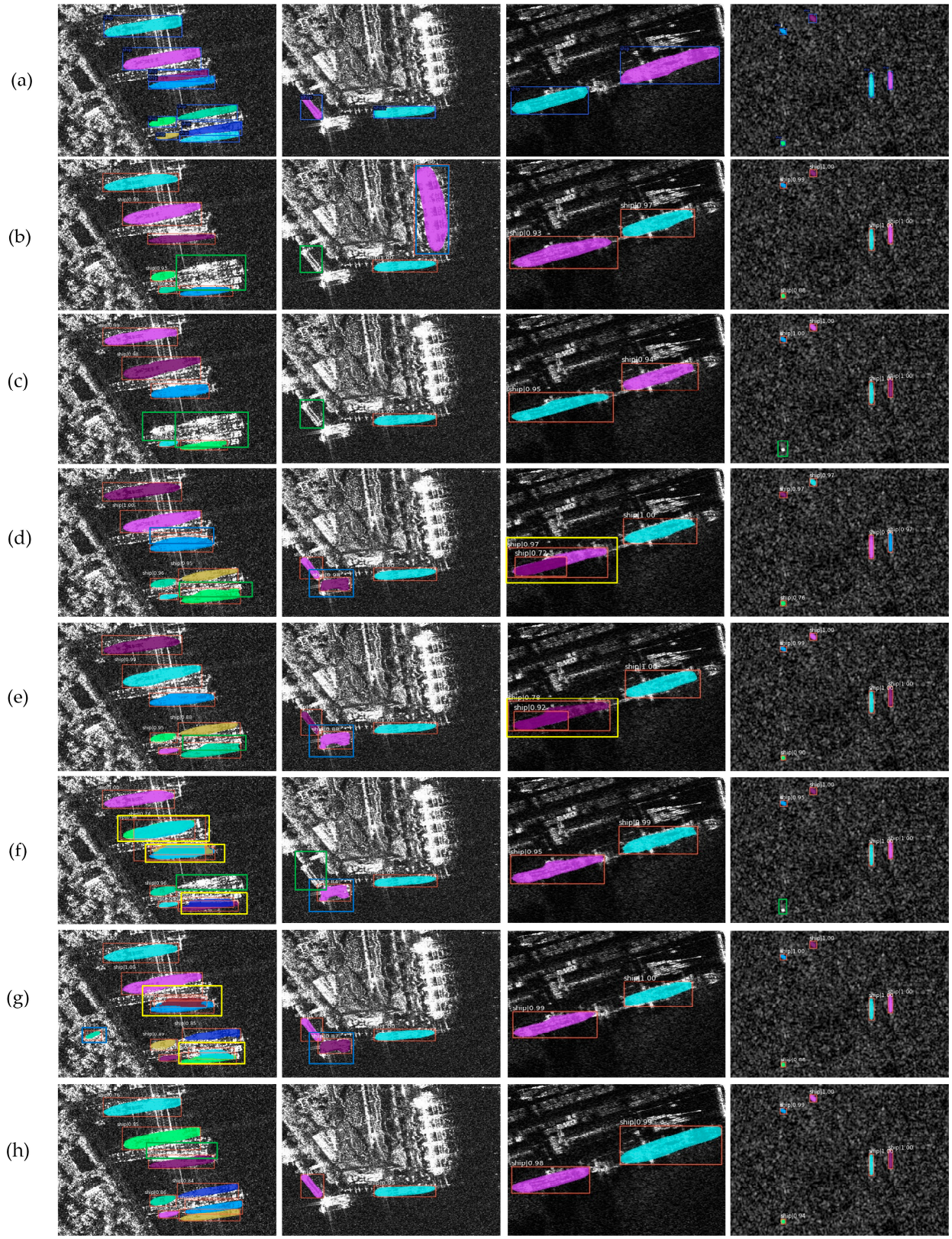

Figure 8. Qualitative experimental results on the SSDD. (**a**) Shows the ground-truth annotations of the objects; (**b**–**g**) Shows the results of the other methods, Cascade Mask-RCNN, Mask-RCNN, MS-RCNN, HTC, SC-Net, and Insta-Boost, respectively; (**h**) Shows the results of our QCIS-Net. We use green boxes to represent missed detections, yellow boxes to represent mixed detections, and false detections with blue boxes.

Table 9. Classwise detection results on the NWPU-VHR-10 test set.

Method	AP	AI	BD	GTF	VC	SH	TC	HB	ST	BC	BR
Mask-RCNN [21]	68.5	78.4	77.6	87.3	56.1	60.0	77.7	51.1	80.6	77.4	37.4
MS-RCNN [52]	68.8	79.0	79.9	87.7	55.7	61.1	78.0	50.7	81.8	79.8	34.4
CM-RCNN [19]	70.0	80.7	78.0	91.2	58.0	64.1	79.4	52.9	81.9	80.3	31.9
HTC [20]	71.4	80.0	80.5	92.8	57.6	62.3	81.6	53.3	83.0	82.0	35.7
SC-Net [53]	70.8	79.0	80.9	89.1	58.3	64.8	82.6	56.2	82.3	81.0	32.0
Insta-Boost [54]	70.6	79.6	80.8	86.9	57.7	58.7	81.9	53.3	81.1	80.5	36.3
QCIS-Net	77.1	82.7	84.2	95.3	66.6	71.3	84.1	65.5	84.6	86.3	45.9

From Table 7, QCIS-Net achieves an AP score of 73.3% on the NWPU-VHR-10 test dataset for IS tasks. This score is 6.9% higher than the classical Mask-RCNN, indicating the network's effectiveness in IS of optical images. Compared to more advanced IS methods, such as HTC, SC-Net, and Insta-Boost, QCIS-Net also demonstrates superior performance, further highlighting its advantages in optical image IS tasks. Overall, our QCIS-Net achieves excellent results in IS tasks for optical images.

Tables 8 and 9 report the IS and object-detection results of QCIS-Net on the optical dataset for each category. In terms of IS, our method achieves the best results in all categories except for storage tanks (which is only 0.4% lower than the optimal value), indicating QCIS-Net's effectiveness in extracting accurate masks for objects in optical images. In terms of object detection, QCIS-Net achieves the best results in all categories, showcasing its outstanding performance. Based on the experimental results from Tables 7–9, it is evident that our QCIS-Net excels in both IS and object-detection tasks for optical images.

Figure 9 shows the results of QCIS-Net and the compared methods on the NWPU-VHR-10 dataset. QCIS-Net accurately segments the instance masks of the objects without apparent over- or under-segmentation phenomena. In the first column, QCIS-Net accurately detects and segments even the less prominent aircraft in the middle of the image, while compared methods such as Mask-RCNN and SC-Net exhibit omissions in detecting and segmenting this object, highlighting the strong performance of QCIS-Net. Furthermore, our QCIS-Net also achieves the accurate IS of the GTF and BD categories in the second column. For the bridges with larger aspect ratios in the third column, some compared methods such as MS-RCNN and SC-Net demonstrate suboptimal performance, while QCIS-Net effectively detects and segments them. In the fifth column, the harbor scene experiences missed detections and segmentations in almost all compared algorithms, while our method is able to detect and segment the objects in the image, demonstrating the robust capability of our approach in handling challenging scenes. Overall, the proposed QCIS-Net exhibits excellent performance in optical image IS and object-detection tasks.

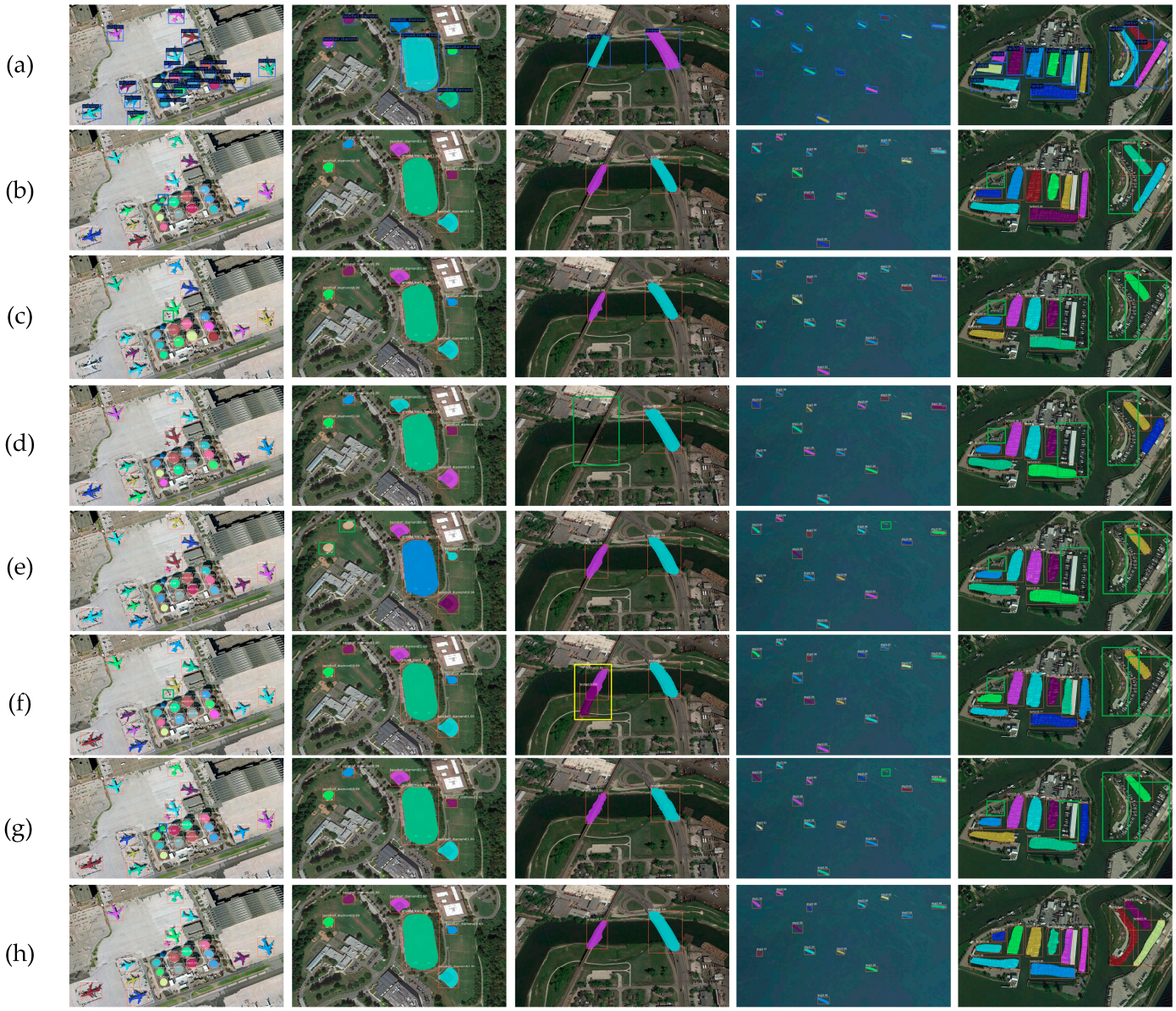

Figure 9. Qualitative experimental results on the NWPU-VHR-10 dataset. (**a**) Shows the ground truth annotations of the objects; (**b**–**g**) Shows the results of the other methods, Cascade Mask-RCNN, Mask-RCNN, MS-RCNN, HTC, SC-Net, and Insta-Boost, respectively; (**h**) Shows the results of our QCIS-Net. We use green boxes to represent missed detections, yellow boxes to represent mixed detections, and false detections with blue boxes.

4. Discussion

Since the QCIS-Net model is larger, it converges slowly in the two datasets without transfer learning, which leads to poor results. However, the accuracy of the model after transfer learning improves significantly, and the performance of the proposed method is better than the comparison experiments. However, there are large fluctuations in the objects of AP_L in the experiments, which may be due to the fact that the proportion of small objects in the SAR and optical image experimental datasets is significantly higher than large objects. This may be one of the reasons for the fluctuations.

Although our approach has achieved good results in the IS of SAR images, the dataset for the SAR images is limited, and the annotations in the original images are not detailed enough. As a result, the final training results may not always perfectly fit the instance objects. Therefore, we are considering expanding the SAR image dataset in the future

and planning to annotate the dataset with more precision. In addition, initializing the object query as zero during training can lead to the relatively slow convergence of the network. Hence, we plan to use the pyramid layer to output feature maps at different levels. By utilizing instance activations from different levels of feature maps, we can guide the initialization process of the query. This allows the query to possess semantic information and instance cues from the image during initialization, thereby accelerating the model's convergence speed. The cascade architecture can give decent results for the network, but it still affects the inference speed, which is not suitable for use on mobile devices. Therefore, building real-time inference models through lightweight model techniques, such as knowledge distillation, becomes one of the priority tasks in our future work.

5. Conclusions

We proposed a query-based multicascade instance segmentation method called QCIS-Net. The method is based on the Swin Transformer backbone network in the EFE module to extract high-quality feature maps for segmentation. Compared with traditional transformer architectures, our method improves detection efficiency and reduces parameter capacity. It can integrate and optimize detection and segmentation, which significantly improves segmentation quality. Compared with traditional detection and segmentation methods, our end-to-end detector does not require the tedious RPN an a priori box and complicated postprocessing process. With the learnable and adjustable feature of the query, we can obtain good performance through its self-learning ability. Experimental results show that, when using QCIS-Net for instance segmentation and object detection, the box detection AP reaches 77.1% and the IS detection AP reaches 73.3% in the NWPU-VHR-10 dataset. In the SAR dataset, the box detection AP reaches 79.1% and the IS detection AP reaches 75.2%. The method proposed in this paper can enhance the practical application efficiency of RS images.

Author Contributions: Conceptualization, E.C. and M.L.; methodology, M.C.; resources, Q.Z.; data curation, Q.Z.; writing—original draft, E.C.; writing—review and editing, M.L. and M.C. 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: Not applicable.

Acknowledgments: Not applicable.

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

References

1. Amitrano, D.; Di Martino, G.; Guida, R.; Iervolino, P.; Iodice, A.; Papa, M.N.; Riccio, D.; Ruello, G. Earth environmental monitoring using multi-temporal synthetic aperture radar: A critical review of selected applications. *Remote Sens.* **2021**, *13*, 604. [CrossRef]
2. Liu, C.; Xing, C.; Hu, Q.; Wang, S.; Zhao, S.; Gao, M. Stereoscopic hyperspectral remote sensing of the atmospheric environment: Innovation and prospects. *Earth-Sci. Rev.* **2022**, *226*, 103958. [CrossRef]
3. Chen, D.; Ma, A.; Zheng, Z.; Zhong, Y. Large-scale agricultural greenhouse extraction for remote sensing imagery based on layout attention network: A case study of China. *ISPRS J. Photogramm. Remote Sens.* **2023**, *200*, 73–88. [CrossRef]
4. Liu, Y.; Chen, D.; Ma, A.; Zhong, Y.; Fang, F.; Xu, K. Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6106–6120. [CrossRef]
5. Fang, L.; Jiang, Y.; Yan, Y.; Yue, J.; Deng, Y. Hyperspectral image instance segmentation using spectral–spatial feature pyramid network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5502613. [CrossRef]
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
7. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.

8. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16. pp. 213–229.
9. Faster, R. Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *9199*, 2969239–2969250.
10. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
11. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. pp. 234–241.
13. Wan, D.; Lu, R.; Wang, S.; Shen, S.; Xu, T.; Lang, X. YOLO-HR: Improved YOLOv5 for Object Detection in High-Resolution Optical Remote Sensing Images. *Remote Sens.* **2023**, *15*, 614. [CrossRef]
14. Chen, Z.; Liu, C.; Filaretov, V.; Yukhimets, D. Multi-Scale Ship Detection Algorithm Based on YOLOv7 for Complex Scene SAR Images. *Remote Sens.* **2023**, *15*, 2071. [CrossRef]
15. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408715. [CrossRef]
16. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9157–9166.
17. Tian, Z.; Shen, C.; Chen, H. Conditional convolutions for instance segmentation. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16. pp. 282–298.
18. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
19. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef]
20. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4974–4983.
21. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
22. Guan, Z.; Miao, X.; Mu, Y.; Sun, Q.; Ye, Q.; Gao, D. Forest fire segmentation from Aerial Imagery data Using an improved instance segmentation model. *Remote Sens.* **2022**, *14*, 3159. [CrossRef]
23. Zhang, T.; Zhang, X.; Li, J.; Shi, J. Contextual Squeeze-and-Excitation Mask R-CNN for SAR Ship Instance Segmentation. In Proceedings of the 2022 IEEE Radar Conference (RadarConf22), New York, NY, USA, 21–25 March 2022; pp. 1–6.
24. Hu, A.; Wu, L.; Chen, S.; Xu, Y.; Wang, H.; Xie, Z. Boundary shape-preserving model for building mapping from high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**. [CrossRef]
25. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
26. Su, H.; Wei, S.; Liu, S.; Liang, J.; Wang, C.; Shi, J.; Zhang, X. HQ-ISNet: High-quality instance segmentation for remote sensing imagery. *Remote Sens.* **2020**, *12*, 989. [CrossRef]
27. Lin, T.-Y.; Dollár, 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, HI, USA, 21–26 July 2017; pp. 2117–2125.
28. Zeng, X.; Wei, S.; Wei, J.; Zhou, Z.; Shi, J.; Zhang, X.; Fan, F. CPISNet: Delving into consistent proposals of instance segmentation network for high-resolution aerial images. *Remote Sens.* **2021**, *13*, 2788. [CrossRef]
29. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
31. Zhao, D.; Zhu, C.; Qi, J.; Qi, X.; Su, Z.; Shi, Z. Synergistic attention for ship instance segmentation in SAR images. *Remote Sens.* **2021**, *13*, 4384. [CrossRef]
32. Fang, F.; Wu, K.; Liu, Y.; Li, S.; Wan, B.; Chen, Y.; Zheng, D. A Coarse-to-Fine Contour Optimization Network for Extracting Building Instances from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2021**, *13*, 3814. [CrossRef]
33. Sun, Z.; Meng, C.; Cheng, J.; Zhang, Z.; Chang, S. A multi-scale feature pyramid network for detection and instance segmentation of marine ships in SAR images. *Remote Sens.* **2022**, *14*, 6312. [CrossRef]
34. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
35. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]

36. Wei, S.; Zeng, X.; Qu, Q.; Wang, M.; Su, H.; Shi, J. HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation. *Ieee Access* **2020**, *8*, 120234–120254. [CrossRef]
37. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
38. Zheng, Z.; Qi, H.; Zhuang, L.; Zhang, Z. Automated rail surface crack analytics using deep data-driven models and transfer learning. *Sustain. Cities Soc.* **2021**, *70*, 102898. [CrossRef]
39. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
40. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
41. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
42. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C. Sparse r-cnn: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14454–14463.
43. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11030–11039.
44. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 807–814.
45. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
46. Dumoulin, V.; Visin, F. A guide to convolution arithmetic for deep learning. *arXiv* **2016**, arXiv:1603.07285.
47. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13. pp. 740–755.
48. Li, B.; Liu, Y.; Wang, X. Gradient harmonized single-stage detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; pp. 8577–8584.
49. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
50. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
51. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
52. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6409–6418.
53. Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Wang, C.; Feng, J. Improving convolutional networks with self-calibrated convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10096–10105.
54. Fang, H.-S.; Sun, J.; Wang, R.; Gou, M.; Li, Y.-L.; Lu, C. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 682–691.
55. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.

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