remote sensing

MDPI

Article

Accurate Instance Segmentation for Remote Sensing Images via Adaptive and Dynamic Feature Learning

Feng Yang [1,2,*], Xiangyue Yuan [1,2], Jie Ran [1,2], Wenqiang Shu [3], Yue Zhao [1,2], Anyong Qin [1,2] and Chenqiang Gao [1,2]

[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; s200101052@stu.cqupt.edu.cn (X.Y.); s180101118@stu.cqupt.edu.cn (J.R.); zhaoyue@cqupt.edu.cn (Y.Z.); qinay@cqupt.edu.cn (A.Q.); gaocq@cqupt.edu.cn (C.G.)
[2] Chongqing Key Laboratory of Signal and Information Processing, Chongqing 400065, China
[3] Chongqing Geomatics and Remote Sensing Center, Chongqing 401147, China; wqshu@dl023.net
* Correspondence: yangfeng@cqupt.edu.cn

Abstract: Instance segmentation for high-resolution remote sensing images (HRSIs) is a fundamental yet challenging task in earth observation, which aims at achieving instance-level location and pixel-level classification for instances of interest on the earth's surface. The main difficulties come from the huge scale variation, arbitrary instance shapes, and numerous densely packed small objects in HRSIs. In this paper, we design an end-to-end multi-category instance segmentation network for HRSIs, where three new modules based on adaptive and dynamic feature learning are proposed to address the above issues. The cross-scale adaptive fusion (CSAF) module introduces a novel multi-scale feature fusion mechanism to enhance the capability of the model to detect and segment objects with noticeable size variation. To predict precise masks for the complex boundaries of remote sensing instances, we embed a context attention upsampling (CAU) kernel instead of deconvolution in the segmentation branch to aggregate contextual information for refined upsampling. Furthermore, we extend the general fixed positive and negative sample judgment threshold strategy into a dynamic sample selection (DSS) module to select more suitable positive and negative samples flexibly for densely packed instances. These three modules enable a better feature learning of the instance segmentation network. Extensive experiments are conducted on the iSAID and NWU VHR-10 instance segmentation datasets to validate the proposed method. Attributing to the three proposed modules, we have achieved 1.9% and 2.9% segmentation performance improvements on these two datasets compared with the baseline method and achieved the state-of-the-art performance.

Keywords: remote sensing images; instance segmentation; object detection; multi-scale fusion; context attention; sample selection

Citation: Yang, F.; Yuan, X.; Ran, J.; Shu, W.; Zhao, Y.; Qin, A.; Gao, C. Accurate Instance Segmentation for Remote Sensing Images via Adaptive and Dynamic Feature Learning. *Remote Sens.* **2021**, *13*, 4774. https://doi.org/10.3390/rs13234774

Academic Editors: Xiangtao Zheng, Fulin Luo and Qi Wang

Received: 9 October 2021
Accepted: 20 November 2021
Published: 25 November 2021

1. Introduction

With the development of earth observation technologies, a large number of high-resolution remote sensing images (HRSIs) are available, raising the high demands of automatic image analysis by intelligent means [1–10]. The ways to interpret remote sensing images, e.g., object detection, instance segmentation, etc., have recently attracted much attention. Instance segmentation for HRSIs aims at accurately detecting the location of the objects of interest (e.g., vehicles, ships, etc.) in the image, predicting their categories and classifying each instance at the pixel level. It is conducive to the automatic analysis and utilization of the spatial and spectral information of HRSIs, which has a wide range of applications in earth observation, e.g., disaster estimation, urban planning, traffic management, etc.

Recently instance segmentation has been extensively studied based on deep learning both for natural images and aerial images [1,3–5,9,11–24]. The generic instance seg-

mentation methods for natural images can be roughly divided into detection-based and segmentation-based approaches from the point of view of top-down and bottom-up schemes, respectively. The detection-based methods first find out the bounding boxes of the instances of interest through an advanced object detector such as Fast-RCNN [25], Faster-RCNN [26], and R-FCN [27]. Then they perform semantic segmentation in the detection boxes to segment the mask for each instance. Early work [28–32] adopted mask proposals to perform category-agnostic object segments. FCIS [33] uses the position-sensitive inside/outside score maps for instance-aware semantic segmentation with the help of fully convolutional network (FCN) [34]. Mask-RCNN [11] inherits the classification and regression structure from Faster-RCNN [26] and adds a fully convolutional mask branch to predict the category of each pixel. Inspired by Mask-RCNN, [12–15] introduce different designs and structures to refine mask prediction for two-stage instance segmentation. HTC [16] integrates the cascade method into joint multi-stage processing and uses spatial context to further improve the accuracy of segmented masks. YOLACT [35] adds the mask branch to an existing one-stage target detection model, which realizes real-time instance segmentation with competitive results reported by two-stage methods on the MS COCO dataset [36] for the first time. BlendMask [17] combines the instance-level information (such as bounding boxes) and the per-pixel semantic information for mask prediction.

The segmentation-based methods first perform the pixel-level prediction on the image and then use a clustering algorithm to group the pixels into different instances. Bai et al. [37] and Hsu et al. [38] applied the traditional watershed transform and the pairwise relationship between pixels to perform pixel-wise clustering. Liu et al. [39] employed a sequence of neural networks to handle the problems of occlusion and large differences in the number of objects in instance segmentation. Neven et al. [40] proposed a new loss function to achieve real-time instance segmentation with high accuracy. PolarMask [41] models the instance contours in a polar coordinate and transforms the instance segmentation into center point classification and dense distance regression tasks. While [42,43] adopt the location and size information to assign category labels for each pixel within an object, which converts the instance segmentation into a pixel-wise classification problem.

Benefiting from the vigorous success in natural images, the research of instance segmentation for HRSIs has made great progress in recent years. However, Limited by the lack of aerial instance segmentation datasets, previous instance segmentation methods [1,3–5,9] mainly focused on some or other specific categories, such as ship, vehicle, and building. Since Su et al. [19,20] extended the NWPU VHR-10 object detection dataset [44–46] for instance segmentation and the release of the first large-scale instance segmentation dataset named iSAID [47] in the field of aerial imagery, researchers began to turn their attention to multi-category instance segmentation for HRSIs. Su et al. [19] designed the precise RoI pooling incorporated into Mask-RCNN to improve the performance of object detection and instance segmentation in aerial images. HQ-ISNet [20] introduces the HRFPN to reserve high-resolution features of HRSIs and utilizes the ISNetV2 to predict more accurate masks. Ran et al. [21] proposed an adaptive fusion and mask refinement (AFMR) network to learn complementary spatial features and conduct the content-aware mask refinement. CPIS-Net [23] presented a novel adaptive feature extraction network and a proposal consistent cascaded architecture to boost the integral network performance. Zhang et al. [22] proposed a Semantic Attention (SEA) module to alleviate the background interference, and they modified the original mask branch to ameliorate the under-segmentation phenomena.

Remote sensing images are usually taken from a high altitude with the bird's-eye view depending on the perspective of the earth observation platforms. The particular imaging platforms and various kinds of objects lead to some distinctive characteristics of instances presented in HRSIs, such as large scale variation, complex contours, and dense block distribution, as shown in Figure 1. These three issues are the main difficulties complicating the instance segmentation task of HRSIs. This paper specifically focuses on the problems of instance segmentation in HRSIs, which is challenging mainly due to the following three aspects:

(a) (b) (c)

Figure 1. Characteristics of objects in HRSIs. (**a**) There are huge scale variations among different planes. (**b**) Harbors present complex boundaries. (**c**) Densely packed ships appear in the marina. Notice the size gap among objects in the three scenes and the shape differences among the harbors of (**b**,**c**).

Huge scale variation. The intra-class and inter-class scale variation among objects in remote sensing images are much more significant than those in natural images. It will increase the difficulty of feature learning in deep neural networks to a great extent. A general solution to this problem is adopting multi-scale feature extraction and fusion, which has been used in [11,14,15,48–52]. However, these multi-scale feature pyramid based methods such as feature pyramid network (FPN) usually assume that large objects should be represented by high-level features, while small objects should be associated with shallow features. The inconsistency across multi-scale feature maps limits their capability both in object detection and instance segmentation when the scale of objects changes too sharply. Liu et al. [13] proposed a bottom-up path aggregation strategy by using accurate low-level locating information to enhance the feature representation entirely. However, this fixed fusion mechanism does not fully explore the self-learning merit of neural networks. Therefore, it is important to design adaptive feature fusion methods for the combination of useful information at different levels.

Objects with arbitrary shapes. There are numerous objects with arbitrary shapes in HRSIs, such as ships, harbors, helicopters. Since instance segmentation requires to achieve pixel-level accurate classification, these irregular shapes will bring significant challenges to the segmentation task. Zhao et al. [4] attempted to use building boundary regularization to deal with the complex boundary segmentation of buildings. Cheng et al. [5] constructed an energy map in polar coordinates, and then minimized the energy function to evolve the polygon outlines, thereby achieving accurate automatic building segmentation with arbitrary shapes. Limited by the dataset, [4,5] are only for single category segmentation. As for multi-class segmentation, Huang et al. [14] improved their results by using the mask-IoU score to correct the misalignment between the mask quality and mask score. Kirillov et al. [15] converted the instance segmentation into a rendering problem and obtained more nuanced mask predictions on the object contours through an iterative subdivision algorithm. However, current methods ignore the impact of feature upsampling operators when predicting the precise masks. The most widely used bilinear and nearest neighbor interpolation only take the sub-pixel neighborhood into consideration, and the deconvolution operation utilizes the same upsampling kernel across the entire image. These immutable upsampling operators fail to take full advantage of the contextual semantic information required by complex boundary segmentation.

Densely packed small objects. Cities and ports commonly contain considerable densely distributed small objects in aerial images. Since the size of these small objects is only a few to a dozen pixels and they always distribute in blocks, the close arrangements will lead to mutual interference among themselves whether in detection or segmentation. Mou et al. [1] proposed a semantic boundary-aware unified multitask learning ResFCN for vehicle instance segmentation. Feng et al. [3] improved the detection and segmentation performance

of small dense objects via implementing multi-level information fusion on the mask prediction branch. Another solution is to improve the quality of the candidate proposals which are important for the subsequent detection and segmentation. Most detection-based instance segmentation methods adopt a hard threshold to divide positive and negative samples. However, simply using a constant IoU threshold as the criterion to separate training samples is not sufficient. As illustrated in [53–55], the dynamic sample selection strategy can solve the incompatibility between the hard threshold settings and the dynamic training process. For that reason, the network can generate candidate boxes of higher quality to facilitate subsequent localization and pixel-wise classification.

In this paper, we design an end-to-end multi-category instance segmentation network for HRSIs, where three key modules based on adaptive and dynamic feature learning are proposed to address the above three problems. Firstly, a cross-scale adaptive fusion (CSAF) module is proposed to integrate the multi-scale information, which can enhance the capability of the model to detect and segment objects with obvious size variation. Secondly, to deal with the complex boundaries of remote sensing instances, we embed a context attention upsampling (CAU) module in the segmentation branch to get more delicate masks. Thirdly, we devise a Dynamic Sample Selection (DSS) module to select more appropriate candidate boxes, especially for densely packed small objects. We evaluate the proposed method on two public remote sensing datasets.

The main contributions of this paper can be summarized as:

- We propose the CSAF module with a novel multi-scale information fusion mechanism that can learn the fusion weights adaptively according to different input feature maps from the FPN.
- We extend the original deconvolution layer into the CAU module in the segmentation branch which can obtain more refined predicted masks by generating different upsampling kernels with contextual information.
- Instead of the traditional fixed threshold for determining positive and negative samples, the DSS module employs a dynamic threshold calculation algorithm to select more representative positive/negative samples.
- The proposed method has achieved state-of-the-art performance on two challenging public datasets for instance segmentation task in HRSIs.

The rest of this paper is organized as follows. Section 2 describes the proposed method in detail. The experimental results and analysis are reported in Section 3. Section 4 gives some discussions about the proposed method. Finally, we draw a conclusion in Section 5.

2. Methodology

For instance, segmentation, we take PANet [13] as a strong baseline for the outstanding performance in aerial images by following the benchmark paper [47]. The overall framework of the proposed method is depicted in Figure 2. The Resnet-FPN backbone is used to extract multi-scale feature maps from an input image. Then we adopt the cross-scale adaptive fusion (CSAF) module to automatically fuse semantic information across multi-level feature maps. The RPN-based dynamic sample selection (DSS) module is employed for candidate proposals generation. We extend the general fixed positive and negative sample judgment threshold in the region proposal network (RPN) into a dynamic threshold selection strategy to choose high-quality training samples. Finally, the proposals are sent to the bounding box branch for object detection and to the mask branch for instance segmentation where a context attention upsampling (CAU) module is embedded for upsampling. In this section, we describe the three proposed modules in detail.

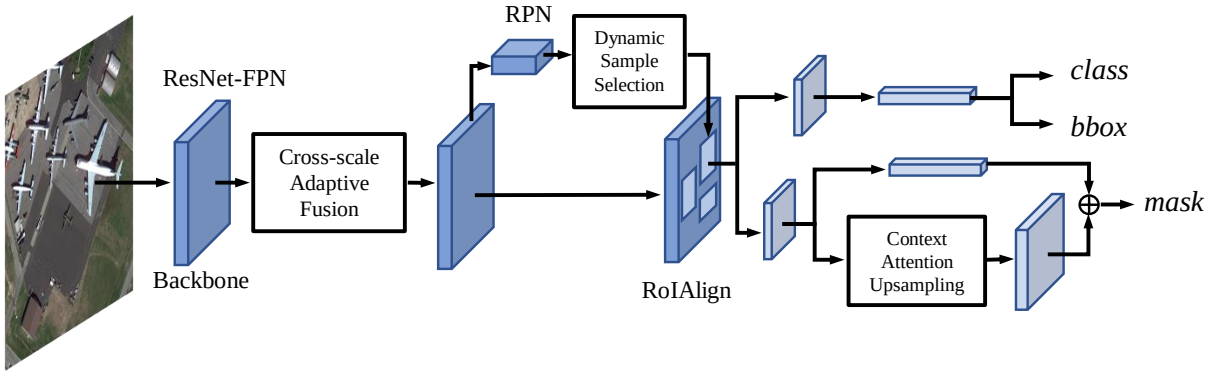


Figure 2. The network structure of the proposed method, which is based on the architecture of PANet and adds the cross-scale adaptive fusion (CSAF) module for multi-scale feature map fusion, context attention upsampling (CAU) module to refine mask prediction and dynamic sample selection (DSS) module to select suitable positive/negative samples.

2.1. Cross-Scale Adaptive Fusion

Most state-of-the-art methods attempt to cope with objects across a wide range of scales via constructing a multi-level feature pyramid [11,13,15,49,56]. The commonly used fusion method is manually setting the fixed fusion weights through a sum or concatenation operator. Different from this way, we design a cross-scale adaptive fusion (CSAF) module to enable the network to learn the fusion weights across feature layers of different levels autonomously. Inspired by [57], we propose a fully connected fusion mechanism between the top-down and bottom-up paths in the PANet. It can adaptively search the optimal fusion operation by adjusting the spatial fusion weights according to the specific pyramidal feature maps. The structure of the cross-scale adaptive fusion module is shown in Figure 3.

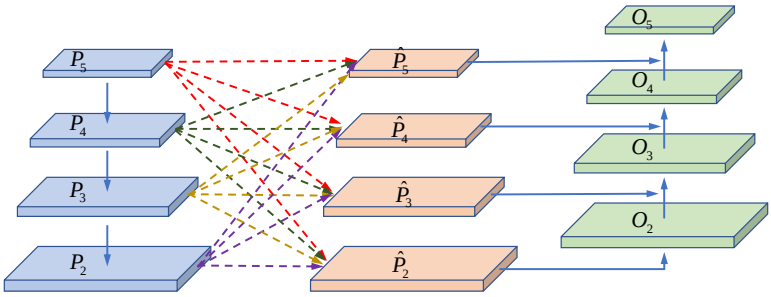


Figure 3. The structure of proposed cross-scale adaptive fusion module. For each pyramidal feature map, the others are rescaled to the same shape and then spatially fused together according to the learned fusion weights.

Given an input image, the ResNet-FPN backbone extracts the multi-level features and that are sent to the cross-scale adaptive fusion module. We denote the feature map at level $l (l \in \{2, 3, 4, 5\}$ in FPN) as P^l. Since the size of these feature maps from FPN is inconsistent, for each level l, we need firstly resize the features maps P^n at all other levels n to the same size as P^l. During the process of resizing, we use the nearest neighbor interpolation to enlarge the feature maps. However, the downsampling method can be a bit more complicated. When the resizing scale is $1/2$, we employ a 2-stride max pooling operation directly; for $1/4$, we adopt the combination of a max pooling layer and a convolution layer whose kernel size is 3×3 with the stride of 2; and when the scale is $1/8$, two consecutive max pooling layers and a two-stride 3×3 convolution are used. After the rescaling step, we further depict the feature fusion mechanism in Figure 4.

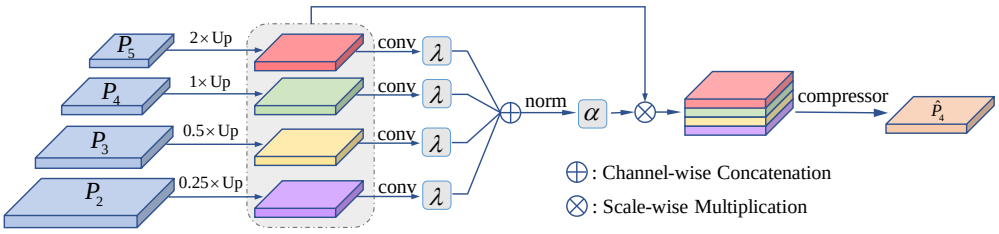


Figure 4. Illustration for the cross-scale adaptive fusion mechanism. Here we take the fusion to the target layer $\hat{P}_4$ for example.

To formulate our method, we denote P^l as the target feature for rescaling and fusion and the other features P^n at level n as the source feature maps. For the feature maps rescaling from source level n to target level l, the feature vector at location (i, j) is $P_{ij}^{n\to l}$. The core of this mechanism is that the network can learn the appropriate fusion weights according to the input multi-scale feature maps adaptively. We firstly use an 1×1 convolution layer to compute the weight scalar maps $\lambda^{n,l}$ from $P^{2\to l}$, $P^{3\to l}$, $P^{4\to l}$ and $P^{5\to l}$, which indicates that λ can be learned through standard back-propagation. Then a softmax function is resorted to normalize these weight maps to obtain the fusion scores. We define the final fusion scores α as follows:

$$\alpha_{ij}^{n\to l} = \frac{e^{\lambda_{ij}^{n\to l}}}{\sum_{k=2}^{5} e^{\lambda_{ij}^{k\to l}}}, \tag{1}$$

where $\alpha_{ij}^{n\to l}$ are the outputs of the softmax function with $\lambda_{ij}^{n\to l}$ as control parameters, respectively. Notice that $\alpha_{ij}^{n\to l}$ are shared by all channels at location (i, j) and they are simple scalar variables and $\sum_{n=2}^{5} \alpha_{ij}^{n\to l} = 1$. Finally, the fused feature map at target level l can be computed as:

$$\hat{P}_{ij}^{l} = \sum_{n=2}^{5} \alpha_{ij}^{n,l} \cdot P_{ij}^{n\to l}, \tag{2}$$

where $\hat{P}_{ij}^{l}$ implies the (i, j)-th vector in the output feature maps $\hat{P}^l$ across channels. $\alpha_{ij}^{n,l}$ refers to the fusion weights for the feature maps between the source level n and the target level l.

Through the proposed learning mechanism, we can get the output fused feature maps $\{\hat{P}_2, \hat{P}_3, \hat{P}_4, \hat{P}_5\}$. We denote the final output feature maps as $\{O_2, O_3, O_4, O_5\}$, which will be sent to the following RPN for proposal generation. Each feature map $O_i (i \in \{3, 4, 5\})$ is obtained from the fused feature $\hat{P}_i$ and a higher resolution feature map O_{i-1} through lateral connection and a $0.5\times$ downsampling. The fusion of these two feature maps is a simple addition operation, which is consistent with the bottom-up path in the original PANet. Additionally, O_2 is simply $\hat{P}_2$ without any operation. Through the bottom-up pathway, we integrate the cross-scale fused features into the information flow of PANet. Thus, the semantic information from high-level feature maps is cross linked to the small objects, while the location information from shallow feature pyramids can also be used for the detection and segmentation of large instances.

2.2. Context Attention Upsampling

HRSIs contain many instances with complex boundaries which makes it difficult to achieve pixel-level classification and limits the final segmentation performance. It is mainly because that the network has lost the detailed information of the object boundaries after multiple downsampling operations. However, existing conventional upsampling methods cannot completely recover these lost details. In PANet, the output of mask predictions is obtained from 14×14 RoI features and a deconvolution layer is employed to upsample

them by 2×. Yet this deconvolution layer uses the fixed upsampling kernel across the whole image, paying no attention to the semantic characteristic of specific proposals. In this paper, we replace the deconvolution operation with a context attention upsampling (CAU), where the upsampling kernel can be adaptively generated according to different contextual information.

The pipeline of proposed CAU module is dipicted in Figure 5. We denote the RoI feature map Ψ from RoIAlign operation as the source feature and the upsampled feature map Ψ' as its corresponding target feature in CAU. The width and height of Ψ are W and H, and we denote the number of its channel as C. With an upsampling scale η, the size of the RoI feature map is upsampled from $W \times H \times C$ to $\eta W \times \eta H \times C$. Firstly, we use an 1×1 convolution layer to compress the channel of Ψ, whose output feature shape is referred to $W \times H \times C_f$. The fewer number of channels implies fewer parameters and less computational cost during upsampling. Compressing the channel from C to C_f will not lead to performance decline while being more efficient. It can be inferred that each position $p = (i, j)$ on Ψ corresponds to a size of $\eta \times \eta$ target region $p' = (\eta \times i, \eta \times j)$ on Ψ'. Instead of using the fixed sampling kernel, a context aware kernel is designed for feature reassembly, whose size is denoted as $k_{up} \times k_{up}$. Since the upsampling kernel is dynamically generated by the contextual semantic feature map of each location, the shape of the final predicted kernel map for the whole feature map can be written as $\eta W \times \eta H \times {k_{up}}^2$. We adopt a $3 \times 3 \times C_{up}$ ($C_{up} = \eta \times \eta \times {k_{up}}^2$) convolution layer to encode the compressed feature map and a shuffle operation to generate such upsampling kernel on-the-fly. Afterward, we utilize a softmax function for spatial normalization on the rearranged kernel feature map. The reason for normalization is to ensure the sum of kernel values equal to 1, which will not change the mean values of the feature map. Finally, we integrate the input RoI feature and the upsampling kernel, thus the target feature $\Psi'_{p'}$ on position p' can be denoted as

$$\Psi'_{p'} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} \omega_{p'(n,m)} \cdot \Psi_{(i+n,j+m)}, \tag{3}$$

where $r = k_{up}/2$ and $\omega_{p'(n,m)}$ is the upsampling kernel generated by the above steps. (i, j) denote the original position on RoI feature Ψ.

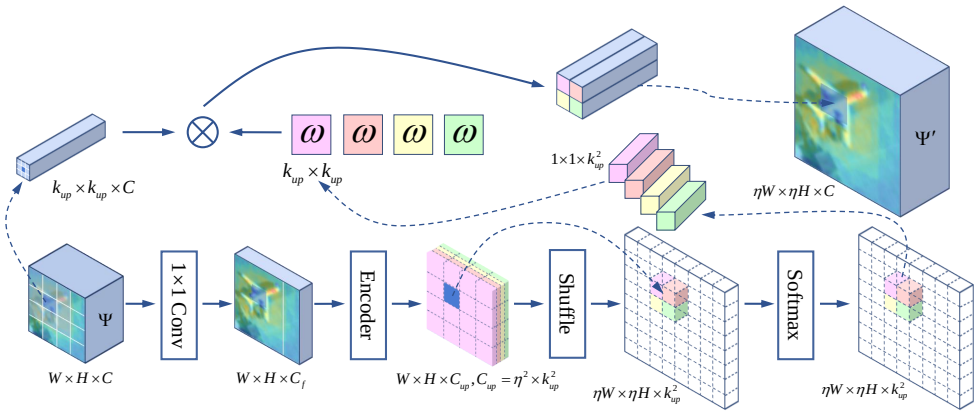


Figure 5. Illustration of the context attention upsampling. A feature map Ψ with size $W \times H \times C$ is upsampled by a factor of η to the output feature map Ψ'. Here we take $\eta = 2$ for example.

With the adaptive sampling kernel, the RoI feature map is upsampled in a context attention manner and retains much more instance-specific semantic information [58]. As shown in Section 3.5, our mask prediction is much more accurate, especially for instances with complex boundaries.

2.3. Dynamic Sample Selection

How to separate positive and negative samples accurately is a noteworthy problem in both of object detection and instance segmentation tasks. The most widely used strategy is to set a hard threshold for the IoU between the anchor bounding box and that of the corresponding one of ground truth. When the IoU is higher than the threshold, the anchor box will be assigned a positive label; otherwise, it will be assigned a negative label. Taking a binary classification loss for simplicity, the paradigm can be formulated as follows:

$$label = \begin{cases} 1, & if\ maxIoU(a, G) \geq t_+, \\ 0, & if\ maxIoU(a, G) < t_-, \\ -1, & otherwise. \end{cases} \tag{4}$$

Here a stands for a anchor bounding box, G represents the set of ground-truths, t_+ and t_- are the positive and negative threshold for IoU. Labels of $1, 0, -1$ stand for positives, negatives and ignored samples, respectively. As for RPN in Faster R-CNN, t_+ and t_- are set to 0.5 by default. So the definition of positives and negatives is essentially pre-set.

In HRSIs, many instances like vehicles and ships present a densely packed distribution. As shown in Figure 6, a region proposal may contain several instances with similar appearances. These adjacent instances can be background interfering noises when extracting accurate features for a certain instance, which increases the difficulty of selecting positive and negative samples. Using manually set and fixed sample judging thresholds adopted by most classic methods is not suitable for remote sensing images because there are a large number of densely packed instances with complex and irregular boundaries. Thereby, we propose a dynamic sample selection (DSS) module with a novel positive and negative samples selection strategy and a constrained IoU calculation method. It works to assign positive and negative samples dynamically, improve the quality of positive samples, reduce the interference between adjacent instances, and alleviate the pressure of coordinate regression. The DSS finally enhances the performance of both object detection and instance segmentation.

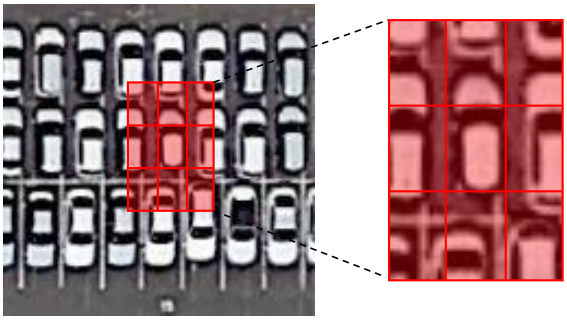


Figure 6. The mutual interference among densely packed instances. There can be multiple objects in a candidate bounding box (as shown in the red box in the figure). These neighboring objects that have similar appearances and structures can be considered as interference noise during locating and classification, which affects the prediction results of the network.

2.3.1. The DSS Strategy

The proposed dynamic scheme is applied in the process of generating positive candidate samples, replacing the original simple IoU threshold filter strategy. Specifically, given an input image, we denote the set of ground-truth boxes as G and the number of feature maps as l. The sets of all anchor boxes and anchor boxes on the i_{th} $(i \in \{1, 2, \cdots, l\})$ feature level are written as A and A_i, respectively. For each ground-truth box $g \in G$, we first compute the distances between the center of each anchor box in A_i and the center of g from feature pyramid level 1 to level l. The top-k closest anchor boxes are labeled as the set of candidate positive samples C_i belonging to g on level i. Then each ground-truth

box will have $k \times l$ candidate positive samples in total. Secondly, we compute the IoU between the candidates set and their corresponding ground-truth box g. The IoU set is denoted as D_g whose mean and standard deviation are m_g and v_g. After that, a dynamic IoU threshold for g is computed as $t_g = m_g + v_g$. Finally, according to the label assignment rule of object detection in Equation (4), the proposed adaptive training sample selection can be formulated as follows:

$$label = \begin{cases} 1, & if\ IoU(c,g) \geq t_g, \\ 0, & if\ IoU(c,g) < t_g. \end{cases} \tag{5}$$

where g represents the ground-truth box, c stands for a candidate positive sample of C_i whose center is in g, and t_g is the dynamic threshold for IoU. 1 and 0 stand for the positive and negative samples. Additionally, if an anchor box is assigned to multiple ground-truth boxes, the one with the greatest IoU will be selected.

2.3.2. Constrained IoU Calculation

The common IoU for comparing similarity between two boxes is computed by:

$$IoU(c,g) = \frac{|c \cap g|}{|c \cup g|}. \tag{6}$$

In order to obtain high-quality positive samples, we add a penalty item on the basis of IoU. It is defined as:

$$IoU_p(c,g) = IoU(c,g) - \Delta, \tag{7}$$

$$\Delta = \frac{|R \backslash (c \cup g)|}{|R|}. \tag{8}$$

Here, c denotes a candidate positive sample, g is the corresponding ground-truth box, R represents the smallest rectangle enclosing them, and Δ stands for the proposed penalty item.

The calculation process of Δ is illustrated in Figure 7. We first find the minimum enclosing rectangle R of the candidate positive sample c and its ground-truth box g. Then we calculate the area occupied by R excluding c and g, which next divide by the total area of R. This represents a normalized measure that focuses on the empty area in the minimum enclosing rectangle of c and g. Finally, the constrained IoU is attained by subtracting this ratio from the IoU value. The value range of IoU_p is within $(-1, 1]$. When the candidate sample coincides with its ground-truth box, IoU_p equals to IoU. When the two are far apart, IoU_p approaches -1. IoU_p holding the property of IoU such as insensitivity for scale variety, and can reflect the relative position between the candidate positive sample and ground-truth box. As shown in Figure 8, the relative positions of the candidate positive sample and the ground-truth box in Figure 8a,b are different while their IoU are the same. We are not able to judge which candidate sample is better according to IoU. However, if we use the IoU_p as the criterion, the candidate box in (a) is superior than that of (b). The coordinate regression difficulty of the candidate positive sample in Figure 8a is lower than that in Figure 8b, and we only need to regress the horizontal axis in (a). From this perspective, the proposed constrained IoU_p is able to distinguish the two candidates and more suitable than the original IoU.

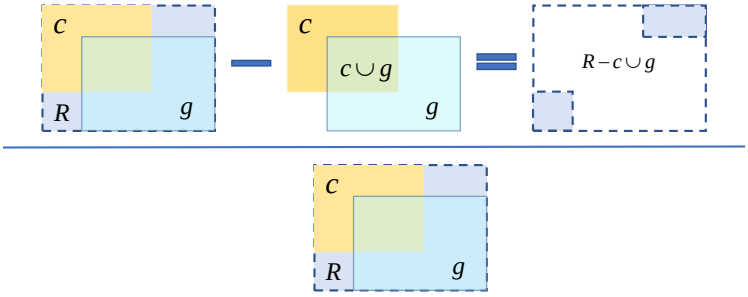


Figure 7. The calculation process of penalty item. The yellow box and the light blue box denote the candidate positive sample and its corresponding ground-truth box. The blue dashed box R represents their minimum enclosing rectangle.

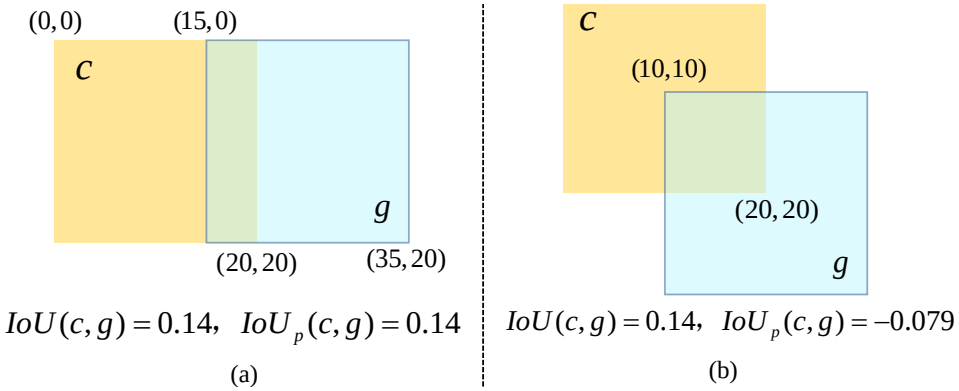


Figure 8. The differences between IoU and IoU_p. The candidate box c of (**a**) is placed parallel to the ground-truth g, while that of (**b**) is placed in a misplaced position. Although the IoU of (**a**,**b**) is the same, their difficulty of coordinate regression is different.

3. Experiments

3.1. Datasets Description

We evaluate the proposed method on two public remote sensing datasets: the iSAID [47] and the NWPU VHR-10 instance segmentation dataset [19,20].

iSAID: The iSAID [47] dataset released in 2019 is the first benchmark dataset for the multi-category instance segmentation in aerial imagery that contains object-level and pixel-level annotations. It consists of 2806 remote sensing images with size ranging from 800 to 13,000 in width. There are totally 655,451 instances annotated in 15 categories including plane (PL), ship, storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), harbor (HB), bridge (BR), large vehicle (LV), small vehicle (SV), helicopter (HC), roundabout (RA), swimming pool (SP) and soccer ball field (SBF). For dataset splitting, half of the images are used for training, 1/6 images for validating, and 1/3 for testing. When the input image size is too large, the computational complexity of the network will rise sharply, which will bring a great burden to the computing device. Therefore, following the iSAID benchmark paper [47], we crop the original images into 800 × 800 patches with the stride of 200 pixels. After the data preprocessing, the number of images for the training set, validation set and test set are 28,029, 9512, and 19,377, respectively. We evaluate our model both on the iSAID validation set and its test set. Since the ground-truth annotations of the test set are not available, the results are tested on the official evaluation server [59]

NWPU VHR-10 Dataset: The NWPU VHR-10 dataset [44–46] is a geospatial object detection dataset and was extended by Su et al. [19,20] for instance segmentation. It contains 650 positive images (the image contains at least one object) and 150 negative images (the image does not contain any object) with the sizes ranging from 533 × 597 to 1728 × 1028 pixels.

In the positive image set, 757 airplanes (PL), 302 ships (SH), 655 storage tanks (ST), 390 baseball diamonds (BD), 524 tennis courts (TC), 159 basketball courts (BC), 163 ground track fields (GTF), 224 harbors (HB), 124 bridges (BR), and 477 vehicles (VC) were manually annotated with bounding boxes and instance masks for ground truths. In consistence with previous studies [19,20,22,23], we randomly divide the original positive images into two parts: 70% for the training set and 30% for the test set which include 454 images and 196 images, respectively.

3.2. Evaluation Metrics

We use the standard COCO metrics [36] for evaluation: AP (averaged over IoU threshold from 0.5 to 0.95 with stride of 0.05), AP_{50} (IoU threshold is 0.5), AP_{75}(IoU threshold is 0.75), AP_S, AP_M and AP_L, where S, M and L represent small $(0 < \text{area} < 32^2$ pixels), medium $(32^2 < \text{area} < 96^2$ pixels) and large objects $(\text{area} > 96^2$ pixels), respectively. Considering the significant scale difference between HRSIs and natural images [22,23,47], we set the area range of different instances for the iSAID dataset by following its benchmark paper [47]. Specifically, the area of small instances ranges from 10^2 to 144^2, medium instances range from 144^2 to 512^2 pixels and large instances range from 512^2 and over. In addition, we use AP^{seg} and AP^{bb} to represent the experimental results of mean average precision of instance segmentation and object detection in Section 3.4.

3.3. Implementation Details

For fair comparisons, all experiments are implemented on PyTorch framework and the hyper-parameters are basically the same as the settings in PANet [13,47]. We use ResNet-FPN as the backbone network and RPN to generate candidate object proposals. Our models are trained on 2 NVIDIA Tesla V100 GPUs with the batch size of 4 for 180 k iterations. The initial learning rate is set as 0.025 and decreased by a factor of 10 at 120 k and 160 k iteration. Besides, we adopt the SGD for training optimization using the weight decay of 0.0001 and the momentum of 0.9. As large scale variation exists among aerial images, we use scale augmentation for the training set at five scales (1200, 1000, 800, 600, 400). For the testing phase, the test images are resized to 1000 pixels on the long side. Considering the number of objects in HRSIs is much larger than in natural images, during evaluation, we set the number of detection boxes as 1000 instead of 100 by default. The source code is available on the website [60].

3.4. Quantitative Results

3.4.1. Results on iSAID

As shown in Tables 1–5, we report the quantitative results achieved by PANet baseline and the proposed method for instance segmentation and object detection tasks on iSAID validation and test sets. We also compared our results with that of those state-of-the-art methods including BlendMask [17], PointRend [15], D2Det [51], etc., for validation.

Table 1. Overall results of instance segmentation and object detection on iSAID validation set.

Method	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_{S}	AP^{seg}_{M}	AP^{seg}_{L}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Mask R-CNN [11]	35.4	57.7	37.8	37.4	50.0	53.0	40.4	62.3	44.3	42.6	48.4	62.4
PANet [13]	38.2	62.4	41.0	40.4	51.9	55.5	43.2	66.0	47.6	45.5	49.4	58.4
BlendMask [17]	36.8	60.5	38.2	20.8	45.5	53.1	43.6	64.1	47.6	28.4	50.1	55.2
PointRend [15]	38.5	62.2	40.8	39.9	56.0	53.5	43.6	65.2	47.9	44.9	58.0	48.8
ours	**40.1**	**64.6**	**43.0**	**42.0**	**56.6**	**77.3**	**46.2**	**68.8**	**51.3**	**48.3**	**58.3**	**80.1**

The overall comparison results on the iSAID validation set are presented in Table 1. The bold numbers represent the highest indicators. It can be seen that our method achieves the highest 40.1% AP^{seg} and 46.2% AP^{bb} which indicates our superiority both in instance segmentation and object detection tasks. Compared with the baseline method PANet,

we yield better performance in all six metrics. The AP^{seg} and AP^{bb} have improved by 1.9% and 3.0%, respectively, which shows the proposed modules effectively improve the overall performance of the model. Furthermore, our method achieves 1.6%, 4.7%, and 11.8% higher values in segmenting small, medium, and large instances in HRSIs while the results of object detection have been also improved to some extents, which indicates that the proposed method has advantages in handling significant scale variation in HRSIs. Compared with other state-of-the-art methods in recent two years, we outperform BlendMask and PointRend by margins of 3.3% and 1.6% respectively in overall AP^{seg} and other indices have been also improved correspondingly.

To further verify the capability of our method, we report the class-wise results in Tables 2 and 3. From Table 2, our method has higher AP than PANet in most categories except tennis court (TC) and roundabout (RA). The performance of our model and PANet in these two categories is almost the same, with the difference of only 0.1% and 0.3% respectively. For categories with irregular boundaries and rich detailed information, such as harbor (HB), we have achieved significant performance improvements by 1.1% and 2.5% in AP^{seg} and AP^{bb}. It confirms that our method can handle the instances with arbitrary shapes in aerial images. Furthermore, for the categories containing a great amount of small and densely distributed objects such as small vehicle (SV) and ship, segmentation accuracy has improved by about 2% while 4% for object detection. These quantitative results show that our model can effectively address the problems of insufficient and inaccurate segmentation for hard samples in multi-class instance segmentation of HRSIs. Considering the results of comparison with other advanced methods, we achieve better performance in many categories while lower AP within 1% decrease for plane (PL), tennis court (TC), and roundabout (RA). It is mainly because the number of objects for these categories is too small (300 vs. 232,457). Therefore, AP can be easily affected by a tiny deviation in prediction.

Table 2. Class-wise results of instance segmentation on iSAID validation set.

Method	AP	SV	LV	PL	ST	ship	SP	HB	TC	GTF	SBF	BD	BR	BC	RA	HC
Mask R-CNN	35.4	15.5	37.3	51.5	38.5	44.3	33.5	28.3	76.8	26.0	38.7	49.8	20.3	34.2	31.7	4.8
PANet	38.2	18.0	41.0	54.0	39.6	54.4	35.4	30.3	78.1	29.9	39.3	51.7	20.7	38.5	35.5	7.0
BlendMask	36.8	14.7	37.8	**54.7**	40.3	41.0	34.2	30.4	78.4	23.6	41.4	51.9	20.9	38.5	35.6	8.1
PointRend	38.5	16.4	40.7	54.4	37.4	49.2	32.6	31.3	**79.2**	**35.3**	43.0	52.7	22.1	39.4	**35.6**	**8.3**
ours	**40.1**	**20.0**	**41.9**	54.1	**41.7**	**56.6**	**36.0**	**31.4**	78.0	31.5	**45.4**	**55.2**	**22.4**	**43.9**	35.2	8.1

Table 3. Class-wise results of object detection on iSAID validation set.

Method	AP	SV	LV	PL	ST	ship	SP	HB	TC	GTF	SBF	BD	BR	BC	RA	HC
Mask R-CNN	40.4	19.7	42.8	69.3	38.7	46.3	37.7	47.3	76.0	40.4	37.4	49.1	24.5	32.6	30.5	14.3
PANet	43.2	22.4	45.8	70.6	39.5	57.0	40.2	50.1	77.6	41.8	37.1	51.4	24.6	36.6	35.3	18.5
BlendMask	43.6	19.9	44.7	**74.3**	42.5	47.7	39.5	49.4	**80.0**	34.9	43.6	52.6	26.0	39.1	**37.6**	21.3
PointRend	43.6	20.3	45.0	70.7	37.0	52.8	36.1	50.9	79.9	**45.8**	42.7	53.3	25.9	39.1	36.3	17.8
ours	**46.2**	**25.8**	**47.5**	71.0	**42.3**	**60.6**	**41.6**	**53.6**	78.3	44.2	**45.0**	**55.8**	**27.8**	**42.9**	34.8	**21.5**

Tables 4 and 5 show the comparison results on iSAID test set. The models we use here are trained with the ResNet-101-FPN and all results are tested on the official iSAID evaluation server. Although using deeper networks such as ResNet-152 can improve the performance for remote sensing images to some extent, which is mentioned in [47], it also indicates greatly increasing parameter quantities and heavier computation, and the training and inference time also increase multiply. Considering the computing resources and the model size, we take PANet as our baseline instead of PANet++ in [47] for comparison under the backbone of ResNet-101-FPN. As reported in Table 4, with the designed three modules, we achieve the best performance in every evaluation metric. The average AP^{seg} and AP^{bb} rises by 1.3% and 2.3% compared with PANet. Besides, our method is superior to the other multi-category instance segmentation methods. From the class-wise results in Table 5, we

also display higher AP in all categories of iSAID than the original PANet. Notice that the AP increment of large vehicle (LV), plane (PL), harbor (HB) is 2.3%, 2.2%, 2.4% and that of ship is up to 4%. Instances in these categories have obvious characteristics like scale variation, complex contours, and dense block distribution, which are the problems we focus on most and want to solve. The improvement shows that our model can effectively address these issues.

Table 4. Overall results of instance segmentation and object detection on iSAID test set.

Method	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_{S}	AP^{seg}_{M}	AP^{seg}_{L}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Mask R-CNN	33.4	56.8	34.7	35.8	46.5	23.9	37.2	60.8	40.7	39.8	43.7	16.0
PANet	38.4	61.8	41.4	41.2	47.2	11.8	43.8	65.3	49.3	46.2	52.6	17.5
blendmask	36.7	59.5	39.0	39.5	44.7	10.6	43.1	62.3	48.4	45.8	49.0	19.1
PointRend	38.1	61.3	41.0	40.7	48.6	16.9	43.5	64.6	48.5	45.8	53.0	22.8
D2det	37.5	61.0	39.8	-	-	-	-	-	-	-	-	-
ours	**39.7**	**62.7**	**42.8**	**41.9**	**54.6**	**25.5**	**46.1**	**66.4**	**52.2**	**48.5**	**53.8**	**26.2**

Table 5. Class-wise results of instance segmentation on iSAID test set.

Method	AP	SV	LV	PL	ST	ship	SP	HB	TC	GTF	SBF	BD	BR	BC	RA	HC
Mask R-CNN	33.4	16.9	30.4	41.7	32.0	48.8	36.7	29.6	72.9	25.9	26.7	39.6	15.2	43.1	36.0	5.6
PANet	38.4	17.6	32.1	45.0	37.3	50.7	38.3	33.3	76.2	28.8	38.9	53.4	18.9	48.5	41.3	15.2
BlendMask	36.7	15.8	31.1	45.4	35.8	47.0	38.5	34.5	74.8	22.7	33.9	52.9	16.4	49.1	40.4	11.8
PointRend	38.1	16.3	32.3	46.2	36.3	47.7	37.8	**36.5**	75.0	28.6	36.9	51.8	17.5	48.2	41.5	13.2
D2det	37.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ours	**39.7**	**18.9**	**34.4**	**47.2**	**37.5**	**54.7**	**40.8**	35.7	**77.6**	**31.2**	**40.0**	**54.5**	**19.3**	**49.4**	**41.9**	**15.5**

Considering the comparison results with other methods in these two tables, similar to the class-wise performance on the validation set, our method has better capability in the main categories such as vehicle (SV and LV) and ship which make up the vast majority of the dataset. The instances in these categories are mostly small-scale and densely distributed. The improvement of the results indicates that our method is superior in detecting and segmenting densely packed small objects. As for the subtle decrease in the result of harbor (HB) compared with PointRend [15], we attribute this to the iterative subdivision algorithm adopted by PointRend, which helps to better segment the edges of these instances with extremely complex contours. However, PointRend is much more complicated than our method for the mask branch of the upsampling convolution, which achieves the comparable segmentation performance but the implementation approach is much simpler.

3.4.2. Results on NWPU VHR-10 Instance Segmentation Dataset

We report the overall performance of the proposed method on the NWPU VHR-10 test set in Table 6. It can be seen that our method presents the best capability in both instance segmentation and object detection among the state-of-the-art with the highest AP^{seg} and AP^{bb} of 67.7% and 69.4%. Comparing with the results of PANet, we achieve higher performance on NWPU VHR-10 dataset where the improvements are up to 2.9% in AP^{seg} and 3.1% in AP^{bb}. Our improvements in AP_S and AP_L indicate that our method can detect and segment objects of different sizes better. Notice that the results of instance segmentation are better than that of object detection for Mask R-CNN [11] and PointRend [15], which can be attributed to the characteristics of the dataset. Most instances in the NWPU VHR-10 dataset are with regular boundaries and easy to segment. However, similar object appearances can lead to mutual interference that means they are difficult to detect accurately.

Table 6. Overall results of instance segmentation and object detection on NWPU VHR-10 test set.

Method	AP^{seg}	AP^{seg}_{50}	AP^{seg}_{75}	AP^{seg}_{S}	AP^{seg}_{M}	AP^{seg}_{L}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Mask R-CNN	62.8	90.2	69.0	42.3	61.3	69.5	60.6	90.8	69.2	52.2	60.9	54.8
PANet	64.8	92.3	72.7	45.1	63.7	72.3	66.3	91.3	77.3	57.9	57.5	60.2
PointRend	65.4	88.1	73.0	42.8	63.4	77.5	64.9	88.3	76.3	54.0	65.1	60.7
BlendMask	65.7	91.3	73.7	41.2	64.5	69.8	68.0	91.1	79.0	56.7	68.1	57.7
ours	**67.7**	**93.3**	**76.7**	**48.2**	**65.0**	**78.3**	**69.4**	**93.1**	**81.5**	**58.4**	**69.5**	**65.6**

Tables 7 and 8 depict the class-wise segmentation and detection results. Compared with the baseline method, our model achieves higher performance in 9 categories except a comparable AP^{bb} for tennis court (TC). The segmentation AP of vehicle (VC), basketball court (BC) and ground track field (GTF) have gained more than 5% improvement, which proves the effectiveness of the cross-scale adaptively fusion in dealing with the scale variation problem. It can make full use of multi-scale information and is conducive to better detection and segmentation of objects with different sizes appearing in HRSIs. Besides, with the help of the proposed context attention upsamping, the ability of our model to cope with complex contours has been greatly enhanced, which can be varified from the improvement of our method in the two categories of harbor (HB) and airplane (PL). Furthermore, for the vehicle (VC) class, we obtain 5.9% and 4.3% performance gains in segmentation and detection results. That is because the proposed dynamic sample selection can greatly improve the quality of positive samples, reduce the interference between adjacent targets and enhance the segmentation ability of densely distributed objects.

Table 7. Class-wise results of instance segmentation on NWPU VHR-10 test set.

Method	AP	PL	SH	ST	BD	TC	BC	GTF	HB	BR	VC
Mask R-CNN	62.8	43.6	50.4	79.0	82.9	73.0	76.4	83.1	53.9	35.1	50.8
PANet	64.8	50.6	53.5	78.4	83.5	73.0	78.1	87.2	58.6	33.8	51.6
PointRend	65.4	**54.5**	53.2	75.7	84.3	72.4	74.4	90.1	58.8	35.9	54.7
BlendMask	65.7	48.1	51.1	79.8	84.0	72.4	76.7	91.5	58.9	39.6	54.6
ours	**67.7**	51.7	**55.6**	**79.8**	**84.4**	**73.9**	**83.0**	**93.3**	**60.7**	**42.3**	**57.5**

Table 8. Class-wise results of object detection on NWPU VHR-10 test set.

Method	AP	PL	SH	ST	BD	TC	BC	GTF	HB	BR	VC
Mask R-CNN	60.6	64.8	49.4	76.6	82.8	66.3	69.1	67.9	43.4	29.3	55.9
PANet	66.2	79.0	61.5	75.4	80.7	**76.9**	74.9	74.1	51.5	32.3	56.3
PointRend	63.8	75.6	60.1	76.3	81.5	71.4	68.4	72.8	43.4	31.9	56.2
BlendMask	68.0	78.5	60.3	76.6	83.7	75.3	79.1	77.2	54.3	35.4	59.7
ours	**69.4**	**82.9**	**63.2**	**77.0**	**83.9**	76.2	**81.4**	**77.4**	**54.8**	**36.2**	**60.6**

Compared with other state-of-the-art methods, the proposed method shows better performance in both instance segmentation and object detection tasks, which achieves the highest accuracy in most categories, except airplane (PL). Although the boundaries of many instances in the NWPU VHR-10 dataset are relatively regular, accurate segmentation of airplanes is still challenging due to the existence of wings and propellers with overly complex boundaries. The proposed method alleviates this problem to a certain extent through a simple upsampling operation, but there is still a slight gap compared with PointRend [15]. We attribute it to the iterative subdivision algorithm proposed by PointRend, which is more conducive to segmenting the elaborate part of instances such as the wings of an airplane. However, the implementation of PointRend is much more complicated than ours because it requires pointwise optimization for the hard-to-segment points in object boundaries. Moreover, our method has more advantages in handling the huge scale variations among different instances and dense packed small objects, which PointRend cannot realize.

3.5. Qualitative Results

The class-wise visualization instance segmentation results are shown in Figures 9 and 10. It can be seen that the proposed method has achieved great visual performance in each category and shows satisfactory capabilities in detecting and segmenting instances with different sizes (e.g., ship and ground track field). For densely packed small instances (e.g., small vehicle and large vehicle), there are few cases of false and missing segmentation. In addition, the outlines of the predicted masks are clear and accurate, which is shown in the visualization results on categories such as helicopter, airplane, and harbor.

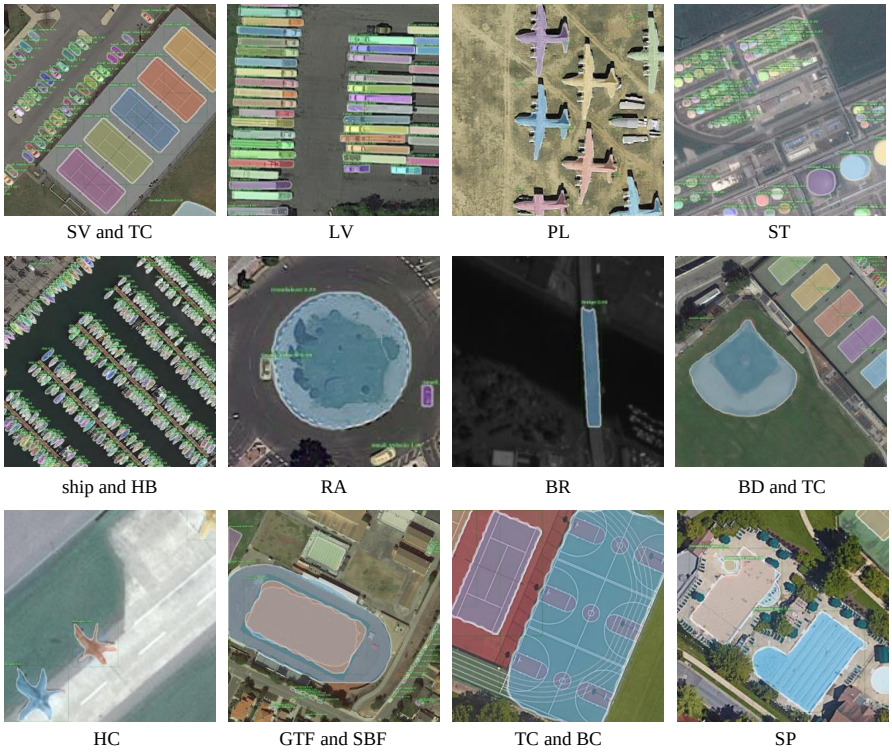


Figure 9. Class-wise instance segmentation of proposed approach on iSAID validation set.

To further validate the instance segmentation performance, the qualitative results of the proposed approach in comparison with the baseline method PANet on iSAID and the NWPU VHR-10 dataset are depicted in Figures 11 and 12. The first column (a) are the original images in the dataset and the second column (b) are their corresponding ground-truth masks. Columns (c) and (d) denote predicted instance results of PANet and the proposed method. These three images correspond to three types of situations that are difficult to deal with in HRSIs, respectively. The first row is images with huge scale variation among different objects. There are small vehicles in the right part of the image whose sizes are extremely small while the size of the ships and harbor is hundreds of times their size. As we can see, our model eliminates the missing detection and under-segmentation of PANet. The reason is that the cross-scale adaptive fusion strategy can take full advantage of cross-scale feature information and optimize the feature extraction procedure through the network. Therefore, our method can effectively deal with the huge scale differences among objects and remove the false and missing detection phenomenon. The second problem is the existence of a large number of objects with irregular contours and complex boundaries in remote sensing images. As shown in the second row of Figure 11, the shape of objects such as harbors is too complex to segment. Our method can segment the harbors more accurately than the baseline, which suggests that the proposed strategy can indeed mitigate the under-segmentation problem of objects with arbitrary shape. The third row of images shows that the remote sensing images contain large quantity dense

and small objects which are formidable to detect and segment. The visualization results demonstrate that our method addresses this issue as mentioned above.

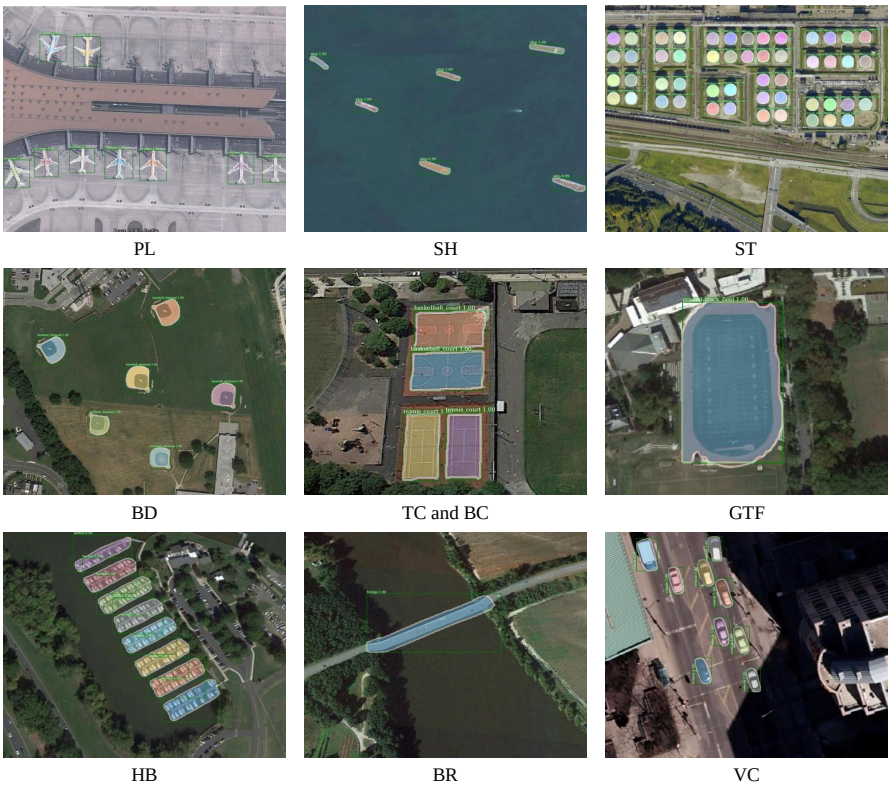


Figure 10. Class-wise instance segmentation of proposed approach on NWPU VHR-10 test set.

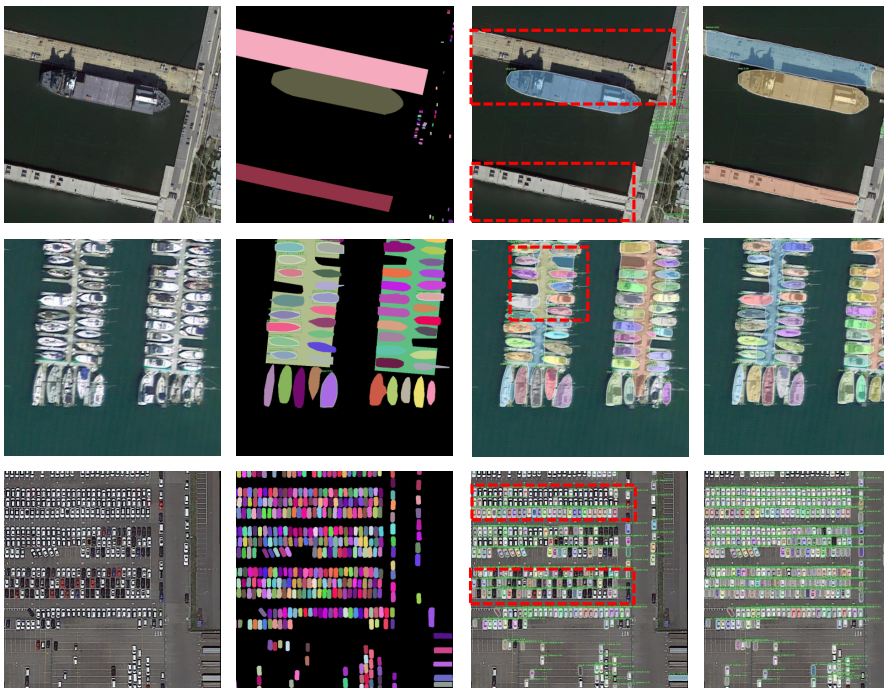
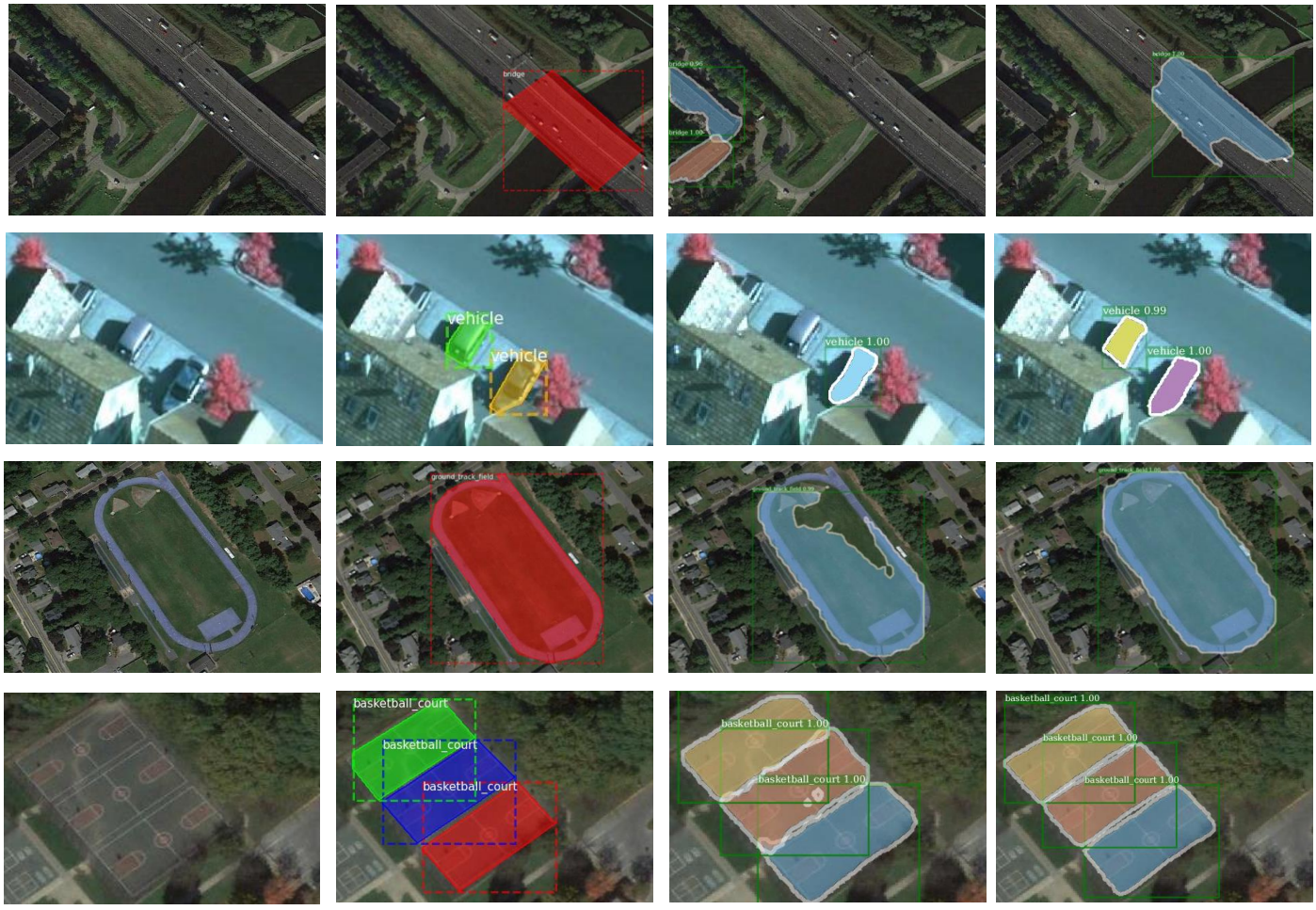


Figure 11. Visual instance segmentation results of the proposed method on iSAID validation set. (**a**) input images; (**b**) ground-truth mask; (**c**,**d**) predicted results of PANet and our method. The red rectangles indicate the missing prediction and under-segmentation problems of PANet.

The visualization results of the proposed method on NWPU VHR-10 instance segmentation dataset are shown in Figure 12. Our overall performance is better than that of the baseline method PANet. It can be observed from the first and second rows that PANet has false and missing detection of objects (e.g., the bridge and the vehicle) with different sizes, while our method can handle this situation. It is because adaptive multi-scale feature fusion can make fuller use of high-level and low-level object-wise semantic information and enhance the ability to detect and segment both of large and small objects. As shown in the third row, our model compensates for under segmentation problem of PANet, which comes from interference caused by similar backgrounds. We attribute this refinement to high-quality proposals selected by DSS. It strengthens the ability of feature representation and improves the quality of the candidate boxes sent into the regression and mask branches. Similarly, when objects distribute in dense blocks, our segmentation masks of objects are much smoother than the baseline. It is mainly because the proposed context attention upsampling has advantages over other general strategies when segmenting elaborate boundaries. The module generates convolution kernels associated with the semantic content of the feature map, which can guide the subsequent segmentation operation.



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

Figure 12. Visual instance segmentation results of the proposed method on NWPU VHR-10 test set. (**a**) input images; (**b**) ground-truth mask; (**c**,**d**) predicted results of PANet and our method.

3.6. Ablation Study

To verify the effectiveness of each proposed strategy, this section tests their performance on iSAID and NWPU VHR-10 datasets. The results are shown in Tables 9 and 10. AUG denotes the random multi-scale scaling strategy used in the training phase; CSAF is

the proposed cross-scale adaptive fusion module; CAU represents the Context Attention Upsampling module; and DSS denotes the dynamic sample selection module. Our method is to deploy the above three modules based on PANet and the results without any proposed strategy in Tables 9 and 10 are tested with the baseline PANet.

Table 9. Validations of different strategies on iSAID validation set.

AUG	CSAF	CAU	DSS	Instance Segmentation						Object Detection					
				AP	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**
				34.2	56.6	35.8	19.6	42.3	46.6	41.7	60.9	46.6	26.9	47.8	51.0
✓				38.2	62.4	41.0	40.4	51.9	55.5	43.2	66.0	47.6	45.5	49.4	58.4
✓	✓			39.4	63.1	42.6	41.2	56.2	71.5	45.7	67.0	50.8	47.5	57.2	77.4
✓	✓	✓		39.7	63.4	42.9	41.4	56.4	71.9	45.8	67.2	51.1	47.5	57.3	77.4
✓	✓	✓	✓	**40.1**	**64.6**	**43.0**	**42.0**	**56.6**	**77.3**	**46.2**	**68.8**	**51.3**	**48.3**	**58.3**	**80.1**

Table 10. Validations of different strategies on NWPU VHR-10 instance segmentation test set.

AUG	CSAF	CAU	DSS	Instance Segmentation						Object Detection					
				AP	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**	**AP**	**AP_{50}**	**AP_{75}**	**AP_S**	**AP_M**	**AP_L**
				63.1	90.5	68.6	42.0	61.9	69.6	62.3	90.3	73.6	54.2	63.5	52.2
✓				64.8	91.3	72.7	45.1	63.7	72.3	66.3	91.3	77.3	57.9	57.5	60.2
✓	✓			65.4	92.4	72.6	45.8	64.1	73.4	68.0	92.5	79.0	58.3	68.6	61.8
✓	✓	✓		66.6	91.3	75.7	46.9	64.8	74.8	68.2	92.8	79.6	58.4	68.6	62.2
✓	✓	✓	✓	**67.7**	**93.3**	**76.7**	**48.2**	**65.0**	**78.3**	**69.4**	**93.1**	**81.5**	**58.5**	**69.5**	**65.6**

It can be observed from Table 9 that after adding multi-scale training to the benchmark method PANet, the performance of instance segmentation task and the object detection task has been improved by 4% and 1.5%, respectively. Since the random multi-scale scaling strategy introduces the prior knowledge of artificial visual invariance into the network, the generalization performance of the model is greatly enhanced. Besides, with the CASF module, our results have been further improved by 1.2% for segmentation and 2.5% for detection. Thanks to the efficient use of information brought by the adaptive fusion strategy, the detection and segmentation accuracy of both small and large objects have been promoted in varying degrees. Furthermore, with the help of CAU, our AP^{seg} result has improved by 0.3% while 0.1% for AP^{bb}. Because this strategy is only embedded in the branch of instance segmentation, and it is relatively independent of the object detection branch, so that the improvement in the target detection task is not as obvious as the former. As shown in the last line, through the DSS module, we achieved the best results with the highest segmentation AP 40.1% and detection AP 46.2%. These improvements can be attributed to the fact that our sample selection strategy can dynamically adjust the threshold for selecting positive and negative samples changing with the training process. DSS makes full use of the statistical characteristics of different input images and generates samples of higher quality. We also add a penalty item when calculating IoU in DSS, which reflects the relative position of the candidate bounding boxes and the ground truth to some extent compared with the original IoU. Therefore, our model is more discernible to the quality of candidate boxes.

Table 10 displays the results of our ablation experiments on the NWPU VHR-10 instance segmentation dataset. Similar to the results on the iSAID validation set, each of the proposed strategies has achieved further performance improvement. By using scale augmentation (AUG) during the training procedure, the overall performance of PANet improves a lot. The cross-scale adaptive fusion (CSAF) makes the network more invariant to the size change of various objects. After using CSAF, the segmentation results of the model for large and small instances has been obviously improved where AP_S and AP_L has

been boost by 0.7% and 1.1%. As for the performance of CAU, similar to the results on iSAID, our segmentation result has dramatically improved by 1.2%, while only 0.2% for object detection, which is because the proposed module mainly works on the mask branch and has little influence on the detection branch. Finally, based on the above strategies, our model utilizes the DSS module and gets the highest AP^{seg} and AP^{bb} of 67.7% and 69.4%, which achieves the state-of-the-art performance on the dataset.

4. Discussion

In this paper, we study the instance segmentation task for remote sensing images, and the proposed instance segmentation method has achieved impressive results. However, there are some limitations in this work, which need to be further improved and optimized. The first limitation is that it is difficult to well segment instances with large aspect ratios such as bridges. Our detection and segmentation AP results for bridges on both iSAID and NWPU VHR-10 instance segmentation datasets are not very satisfactory. The default anchor ratio settings commonly adopted by most state-of-the-art methods (including our baseline PANet [13]) are 1:1, 1:2, and 2:1. These default anchors can indeed well represent most of the instances in the two datasets whose aspect ratios do not vary significantly. However, they are not appropriate for the instances with large aspect ratios in remote sensing images, which brings difficulties for subsequent location regression and instance-wise segmentation. Secondly, since the HRSIs are usually taken from the bird's-eye view, instances in HRSIs always present arbitrary orientations [47,61]. However, current detection-based instance segmentation methods still use horizontal candidate boxes to represent the rotated instances, paying no attention to their orientation information. The horizontal candidate boxes typically lead to misalignment between the bounding boxes and the oriented instances, which brings challenges for extracting object features and identifying the instance's accurate localization [7,8], thus also affects the segmentation results. Therefore, introducing the orientation information into the instance segmentation network for aerial images is conducive to predicting accurate oriented candidate boxes with rotation angles, which can further improve the performance of the detect-then-segment models.

5. Conclusions

In this paper, we design an end-to-end multi-category instance segmentation network for high-resolution remote sensing images aiming at address the problems of the huge scale variation, arbitrary instance shapes, and numerous densely packed small objects in HRSIs. Firstly, we propose a cross-scale adaptive fusion module with a novel multi-scale information fusion mechanism to enhance the capability of the model to detect and segment objects with obvious size differences. Then, we use a context attention upsampling module to substitute the original deconvolution layer in the segmentation branch which can obtain more refined predicted masks. Finally, We extend the traditional fixed positive and negative sample judgment threshold into a dynamic sample selection module to select more suitable positive and negative samples flexibly. It has obvious advantages in the segmentation of densely packed objects. The proposed method has achieved the state-of-the-art performance on two challenging public datasets for instance segmentation task in HRSIs.

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

Funding: This work is supported by the National Natural Science Foundation of China (No. 61906025, 62176035), the Natural Science Foundation of Chongqing, China (No. cstc2020jcyj-msxmX0835, cstc2021jcyj-bsh0155), the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant (No.KJZD-K202100606, KJQN201900607, KJQN202000647, KJQN202100646), the China Postdoctoral Science Foundation (No.2021MD703940).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Mou, L.; Zhu, X.X. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6699–6711. [CrossRef]
2. Zheng, X.; Gong, T.; Li, X.; Lu, X. Generalized Scene Classification from Small-Scale Datasets with Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, in press. [CrossRef]
3. Feng, Y.; Diao, W.; Zhang, Y.; Li, H.; Chang, Z.; Yan, M.; Sun, X.; Gao, X. Ship Instance Segmentation from Remote Sensing Images Using Sequence Local Context Module. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1025–1028.
4. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building Extraction From Satellite Images Using Mask R-CNN With Building Boundary Regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 247–251.
5. Cheng, D.; Liao, R.; Fidler, S.; Urtasun, R. Darnet: Deep active ray network for building segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7431–7439.
6. Luo, F.; Zou, Z.; Liu, J.; Lin, Z. Dimensionality reduction and classification of hyperspectral image via multi-structure unified discriminative embedding. *IEEE Trans. Geosci. Remote Sens.* **2021**, in press. [CrossRef]
7. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2849–2858.
8. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3520–3529.
9. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1480–1484.
10. Zheng, X.; Chen, X.; Lu, X.; Sun, B. Unsupervised Change Detection by Cross-Resolution Difference Learning. *IEEE Trans. Geosci. Remote. Sens.* **2021**, in press. [CrossRef]
11. 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.
12. Chen, L.C.; Hermans, A.; Papandreou, G.; Schroff, F.; Wang, P.; Adam, H. Masklab: Instance segmentation by refining object detection with semantic and direction features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4013–4022.
13. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
14. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring r-cnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6409–6418.
15. Kirillov, A.; Wu, Y.; He, K.; Girshick, R. Pointrend: Image segmentation as rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9799–9808.
16. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid task cascade for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
17. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. BlendMask: Top-down meets bottom-up for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8573–8581.
18. Xie, S.; Chen, Z.; Xu, C.; Lu, C. Environment upgrade reinforcement learning for non-differentiable multi-stage pipelines. *J. Chongqing Univ. Posts Telecommun.* **2020**, *32*, 857–858.
19. Su, H.; Wei, S.; Yan, M.; Wang, C.; Shi, J.; Zhang, X. Object detection and instance segmentation in remote sensing imagery based on precise mask R-CNN. In Proceedings of the IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1454–1457.
20. 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]
21. Ran, J.; Yang, F.; Gao, C.; Zhao, Y.; Qin, A. Adaptive Fusion and Mask Refinement Instance Segmentation Network for High Resolution Remote Sensing Images. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 2843–2846.
22. Zhang, T.; Zhang, X.; Zhu, P.; Tang, X.; Li, C.; Jiao, L.; Zhou, H. Semantic Attention and Scale Complementary Network for Instance Segmentation in Remote Sensing Images. *IEEE Trans. Cybern.* **2021**, in press. [CrossRef] [PubMed]

23. 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]
24. Luo, F.; Zhang, L.; Zhou, X.; Guo, T.; Cheng, Y.; Yin, T. Sparse-adaptive hypergraph discriminant analysis for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 1082–1086. [CrossRef]
25. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 November 2015; pp. 1440–1448.
26. 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.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
27. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
28. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1312–1328. [CrossRef] [PubMed]
29. Arbeláez, P.; Pont-Tuset, J.; Barron, J.T.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 328–335.
30. Pinheiro, P.O.; Collobert, R.; Dollár, P. Learning to segment object candidates. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 7–12 December 2015; pp. 1990–1998.
31. Pinheiro, P.O.; Lin, T.Y.; Collobert, R.; Dollár, P. Learning to refine object segments. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 75–91.
32. Dai, J.; He, K.; Li, Y.; Ren, S.; Sun, J. Instance-sensitive fully convolutional networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 534–549.
33. Arnab, A.; Jayasumana, S.; Zheng, S.; Torr, P.H. Higher order conditional random fields in deep neural networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 524–540.
34. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2359–2367.
35. 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, South Korea, 27 October–2 November 2019; pp. 9157–9166.
36. 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 European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
37. Bai, M.; Urtasun, R. Deep watershed transform for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5221–5229.
38. Hsu, Y.C.; Xu, Z.; Kira, Z.; Huang, J. Learning to Cluster for Proposal-Free Instance Segmentation. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
39. Liu, S.; Jia, J.; Fidler, S.; Urtasun, R. Sgn: Sequential grouping networks for instance segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3496–3504.
40. Neven, D.; Brabandere, B.D.; Proesmans, M.; Gool, L.V. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8837–8845.
41. Xie, E.; Sun, P.; Song, X.; Wang, W.; Liu, X.; Liang, D.; Shen, C.; Luo, P. Polarmask: Single shot instance segmentation with polar representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12193–12202.
42. 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.
43. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. In Proceedings of the Advances in Neural Information Processing Systems, virtual, 6–12 December 2020; pp. 17721–17732.
44. 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]
45. Cheng, G.; 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**, *54*, 7405–7415. [CrossRef]
46. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
47. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 28–37.
48. Luo, F.; Huang, H.; Ma, Z.; Liu, J. Semisupervised sparse manifold discriminative analysis for feature extraction of hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6197–6211. [CrossRef]
49. 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.
50. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2021**, in press. [CrossRef]

51. Cao, J.; Cholakkal, H.; Anwer, R.M.; Khan, F.S.; Pang, Y.; Shao, L. D2det: Towards high quality object detection and instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11485–11494.
52. Gao, C.; Chen, X. Deep learning based action detection: A survey. *J. Chongqing Univ. Posts Telecommun.* **2020**, *32*, 991–1002.
53. 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.
54. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 260–275.
55. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9759–9768.
56. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
57. Wang, G.; Wang, K.; Lin, L. Adaptively connected neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1781–1790.
58. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Carafe: Content-aware reassembly of features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 3007–3016.
59. iSAID: A Large-Scale Dataset for Instance Segmentation in Aerial Images. Available online: https://captain-whu.github.io/iSAID/evaluation.html (accessed on 18 November 2021).
60. Source Code for Accurate Instance Segmentation for Remote Sensing Images via Adaptive and Dynamic Feature Learning. Available online: https://github.com/yuanxiangyuee/ins_seg_HRSIs (accessed on 10 November 2021).
61. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.