remote sensing

MDPI

Article

CPISNet: Delving into Consistent Proposals of Instance Segmentation Network for High-Resolution Aerial Images

Xiangfeng Zeng [1], Shunjun Wei [1,*], Jinshan Wei [1], Zichen Zhou [1], Jun Shi [1], Xiaoling Zhang [1] and Fan Fan [1,2]

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; zxf@std.uestc.edu.cn (X.Z.); 201952011835@std.uestc.edu.cn (J.W.); 202022010839@std.uestc.edu.cn (Z.Z.); shijun@uestc.edu.cn (J.S.); xlzhang@uestc.edu.cn (X.Z.); ff2019@std.uestc.edu.cn (F.F.)

[2] Science and Technology on Communication Security Laboratory, Institute of Southwestern Communication, Chengdu 610041, China

* Correspondence: weishunjun@uestc.edu.cn

Abstract: Instance segmentation of high-resolution aerial images is challenging when compared to object detection and semantic segmentation in remote sensing applications. It adopts boundary-aware mask predictions, instead of traditional bounding boxes, to locate the objects-of-interest in pixel-wise. Meanwhile, instance segmentation can distinguish the densely distributed objects within a certain category by a different color, which is unavailable in semantic segmentation. Despite the distinct advantages, there are rare methods which are dedicated to the high-quality instance segmentation for high-resolution aerial images. In this paper, a novel instance segmentation method, termed consistent proposals of instance segmentation network (CPISNet), for high-resolution aerial images is proposed. Following top-down instance segmentation formula, it adopts the adaptive feature extraction network (AFEN) to extract the multi-level bottom-up augmented feature maps in design space level. Then, elaborated RoI extractor (ERoIE) is designed to extract the mask RoIs via the refined bounding boxes from proposal consistent cascaded (PCC) architecture and multi-level features from AFEN. Finally, the convolution block with shortcut connection is responsible for generating the binary mask for instance segmentation. Experimental conclusions can be drawn on the iSAID and NWPU VHR-10 instance segmentation dataset: (1) Each individual module in CPISNet acts on the whole instance segmentation utility; (2) CPISNet* exceeds vanilla Mask R-CNN 3.4%/3.8% *AP* on iSAID validation/test set and 9.2% *AP* on NWPU VHR-10 instance segmentation dataset; (3) The aliasing masks, missing segmentations, false alarms, and poorly segmented masks can be avoided to some extent for CPISNet; (4) CPISNet receives high precision of instance segmentation for aerial images and interprets the objects with fitting boundary.

Keywords: instance segmentation; aerial images; region proposals; convolutional neural networks

Citation: 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. https://doi.org/10.3390/rs13142788

Academic Editor: Hanwen Yu

Received: 8 June 2021
Accepted: 10 July 2021
Published: 15 July 2021

1. Introduction

With the rapid development of observation and imaging techniques in the remote sensing field, the quantity and quality of very high-resolution (VHR) optical remote sensing images provided by airborne and spaceborne sensors have significantly increased, which simultaneously puts forward new demands on automatic analysis and understanding of remote sensing images. At present, the VHR images are applied in a wide scope of fields, e.g., urban planning, precision agriculture, and traffic monitoring. Meanwhile, with the strong feature extraction and end-to-end training capabilities, deep convolutional neural network (DCNN)-based algorithms show their superiority in the sub-tasks of computer vision, such as object detection, semantic segmentation, and instance segmentation. Driven by the huge application demands and application prospects, researchers have developed

various methods which are combined with DCNN for intelligent interpretation in remote sensing images.

So far, object detection in remote sensing images can be divided into traditional methods, machine learning-based methods, and deep learning-based methods. Traditional methods include template matching-based, knowledge-based, and OBIA-based methods. Some machine learning-based methods regard object detection as a classification problem. A classifier, for instance, Adaboost, support vector machine (SVM), and *k*-nearest neighbors, captures the object appearance variation and generates the predicted labels. Other machine learning-based methods consider object detection in feature space. The most popular bag-of-words (BoW) model treats the image region as an unordered local descriptor for quantizing and computing corresponding histogram representation. The histogram of oriented gradients (HOG) feature represents objects through the distribution of gradient intensities and orientations in spatial, which shows impressive performance. Compared with previous methods, deep learning-based methods unload the traditional human-engineering-based features designed by human ingenuity and replace them with network construction. Moreover, the deep architecture of CNN can extract the semantic abstracting properties of remote sensing images. Both can boost the performance of object detection. Fundamentally, deep learning-based object detection methods for remote sensing images adopt bounding boxes, e.g., horizontal bounding box and oriented bounding box, as the criterion to locate the objects. The coordinates of bounding box can be regressed by related loss function and the backpropagation of DCNN to generate a precise rectangular area that fits the object. Its characteristics of simple but effective make it widely applied. Ref. [1] proposed the rotation-invariant CNN (RI-CNN) which is based on R-CNN framework to detect multi-class geospatial objects in VHR remote sensing images. Ref. [2] proposed rotation-invariant and fisher discriminative CNN (RIFD-CNN) to impose CNN features with fisher discriminative regularizer and rotation-invariant regularizer. Ref. [3] presented a hyper-light deep learning network (HyperLi-Net) which realizes high-accurate and high-speed ship detection. Ref. [4] puts forward a high-resolution ship detection network (HR-SDNet) to realize precise and robust synthetic aperture radar (SAR) ship detection. Meanwhile, researchers adopt the rotatable bounding box (RBox) to detect objects with arbitrarily changed orientation angles. Ref. [5] designed the detector with RBox (DRBox) to learn the correct orientation angle of objects, which can better deal with interferes from the background and locate objects for geospatial object detection in remote sensing images. Ref. [6] came up with an improved RBox-based SAR target detection framework to improve the precision and recall rate when detecting objects.

Distinguished from object detection, semantic segmentation performs pix-wise classification in an image. Such distinguishing characteristics make it widely implemented into vegetation classification, land-cover classification, and infrastructure management, etc., for remote sensing images. Among them, CNN-based semantic segmentation methods generally use the fully conventional network (FCN) to exert end-to-end segmentation for input images. Ref. [7] presents the recurrent network in fully convolutional network (RiFCN) to better fuse the multi-level feature maps. Ref. [8] proposed ResUNet-a with performant results in monotemporal VHR aerial images. Ref. [9] used patch attention module (PAM) and attention embedding module (AEM) to embed local focus from high-level features for semantic segmentation in remote sensing images. Ref. [10] presented the DCNN with built-in awareness of semantical boundaries to realize semantic segmentation.

As instance segmentation simultaneously possesses the characteristic of instance-wise localization in object detection and pixel-wise category classification in semantic segmentation, it has been a hotspot for high-resolution aerial and satellite image analysis recently. In essence, it endows each object under a certain category the attribute of representational color. In the sequel, the instance mask, which is generated by the segmentation sub-net, is responsible for generating the contour information of the objects. Ref. [11] proposed the high-quality instance segmentation network (HQ-ISNet) to implement instance segmentation in high-resolution (HR) remote sensing images. Ref. [12] constructed a high-resolution

SAR images dataset (HRSID) for instance segmentation and ship detection. Ref. [13] introduced the precise region of interests (RoI) pooling for Mask R-CNN [14] to segment multi-category instances in VHR remote sensing images. Ref. [15] came up with a sequence local context (SLC) module to avoid confusion in dense-distributed ships. Ref. [16] introduced the semantic boundary-aware multitask learning network for vehicle instance segmentation. Ref. [17] presented a large-scale instance segmentation dataset for aerial images that contains 655,451 instances across 2806 HR images. Ref. [18] proposed a marine oil spill instance segmentation network to identify the similarity of the oil slick and other elements. Despite the above-mentioned works that predecessors have done, it still lacks algorithms of instance segmentation for high-resolution aerial images.

In this paper, we proposed a novel instance segmentation network for high-resolution aerial images, termed consistent proposals of instance segmentation network (CPISNet), which maintains consistent proposals between object detection and instance segmentation with cascaded architecture. CPISNet consists of three procedures. First, the adaptive feature extraction network (AFEN) is responsible for extracting the multi-level feature maps. Second, the single RoI extractor (SRoIE) and bounding box regression branch are adopted to construct the cascaded architecture, and the refined proposals from the last cascaded stage are transmitted to the elaborated RoI extractor (ERoIE) for mask RoI pooling while maintaining consistent proposals. Third, a consequence of fully convolutional blocks with shortcut connection replaces the interspersed FCN in the cascaded architecture of Cascade Mask R-CNN or HTC.

The main contributions of this paper are summarized as below:

- CPISNet is proposed for multi-category instance segmentation of aerial images;
- Effects of AFEN, ERoIE, and proposal consistent cascaded (PCC) architecture to the CPISNet are individually verified, which boost the integral network performance;
- CPISNet achieves the best AP of instance segmentation in high-resolution aerial images compared to the other state-of-the-art methods.

2. Related Work

2.1. Object Detection

The primary task of object detection is locating each object in the rectangular area with bounding box. Generally, existing object detection methods can be mainly divided into two formats: one-stage and two-stage methods. One-stage method omits the time consuming process of preparing region proposals and generates bounding boxes directly, e.g., You Only Look Once (YOLO) v1-v4 [19–22], Single Shot MultiBox Detector (SSD) [23] and RetinaNet [24]. Ref. [25] proposed the Fully Convolutional One-Stage Object Detection (FCOS) to eliminate the predefined anchors and detect objects in the per-pixel prediction formula. Ref. [26] adopted keypoint triplets into object detection to suppress the number of incorrect object bounding box and presented CenterNet for one-stage object detection. Ref. [27] came up with R3Det to progressively regress rotated bounding boxes from coarse to fine granularity. Relatively, two-stage methods first generate region proposals by a preliminary screening network such as Region Proposal Network (RPN) [28,29] then perform classification and localization via related network branch. The methods derived from Region with Convolutional Neural Network (R-CNN) [28], e.g., Fast R-CNN [30], Faster R-CNN [31], constitute the main-stream two-stage methods. Generally, a feature pyramid network (FPN) [29] is attached to the feature extraction network to generate high-level semantic feature maps. Based on the basic architecture of Faster R-CNN, Cascade R-CNN [32] integrate a sequence of detection branches and train them with increasing Intersection over Union (IoU) thresholds to improve the accuracy. To sum up, one-stage methods are superior in detection speed but attenuated in detection precision, while two-stage methods are the opposite.

2.2. Instance Segmentation

Instance segmentation aims at predicting instance-level mask and pixel-level category of the objects. Mainstream instance segmentation methods can be roughly divided into top-down methods and bottom-up methods. Top-down methods follow the paradigm of detect-then-segment. Fully Convolutional Instance-aware Semantic Segmentation (FCIS [33]) jointly inherits the region proposals generated by RPN to integrate the position-sensitive score maps and the FCN for semantic segmentation. On the basis of Faster R-CNN, Mask R-CNN adds the mask branch to predict the instance-aware mask on each RoI. Path aggregation network (PANet [34]) proposed bottom-up path aggregation to boost the information flow which propagates in top-down instance segmentation methods. Mask Scoring R-CNN [35] presents the mask IoU head to improve the quality of the predicted mask. Hybrid Task Cascade (HTC [36]) proposed the joint multi-stage processing of the mask branch and detection branch. Bottom-up methods aim at grouping the pixels of each instance in an image and predict the corresponding semantic category. Polarmask [37] uses polar coordinate to classify the instance center and regress dense distance. Segmenting objects by locations (SOLO [38]) uses the location and size of the instance to assign the pixel-category, which transfers instance segmentation as a pixel-wise classification problem. SOLOv2 [39] extends SOLO with mask kernel prediction, mask feature learning, and matrix non-maximum suppression (Matrix NMS). BlendMask [40] presents the blender module which is inspired by both top-down and bottom-up methods. Analogously, top-down methods perform well in segmentation precision while bottom-up methods are superior in segmentation speed.

3. The Proposed Method

Our CPISNet follows the formula of top-down instance segmentation that detecting the object first and followed by performing instance-wise segmentation on each RoI. The detailed architecture of CPISNet is shown in Figure 1. First, AFEN is responsible for extracting the multi-level bottom-up augmented feature maps in the design space level. Second, the SRoIE and ERoIE are adopted for extracting the RoIs within the region proposals from RPN and multi-level feature maps from AFEN. Finally, the cascaded bounding box detection architecture and shortcut connection reconstructed mask branch are used for refining the bounding box detection result and generating the high-quality segmentation mask, respectively. The outputs from detection branch and mask branch constitute the instance segmentation result of CPISNet.

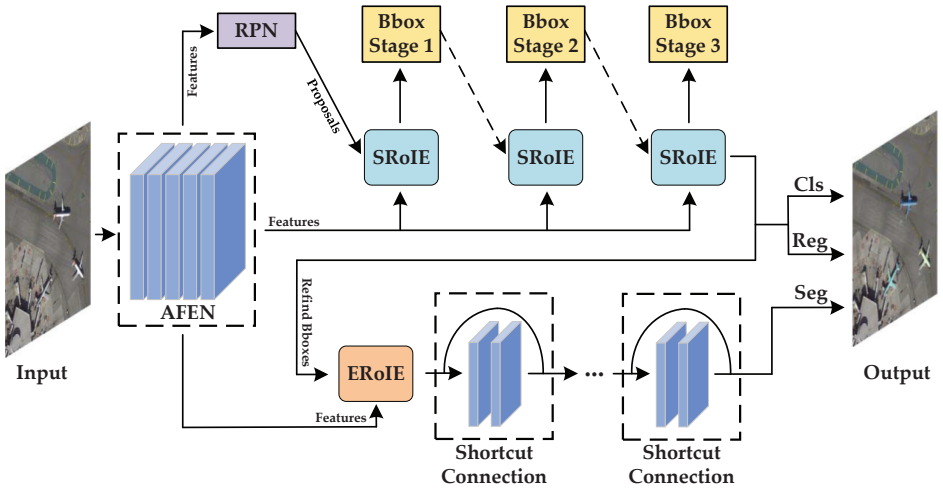

Figure 1. The network architecture of CPISNet, which follows the top-down instance segmentation formula. It constitutes AFEN for multi-level feature map extraction, PCC architecture containing cascaded bounding box stages and mask branch with shortcut connection, and ERoIE for extracting the mask RoI.

3.1. The Adaptive Feature Extraction Network

Our adaptive feature extraction network (AFEN) is separately introduced in two parts: the backbone network and multi-level feature extraction network.

3.1.1. Backbone Network

Instead of inheriting individual designed feature extraction network instances, we introduce RegNetx [41] as the backbone network which processes high-resolution aerial images in the design space level.

As illustrated in Figure 2, RegNetx consists of the stem (3×3 convolution with the stride of 2), stage (consecutive network blocks), and head (average pooling followed by fully connected layer), which is the same as classic backbone networks such as ResNet. Elevated to the structural details, classic backbone networks regard the combination of 1×1 convolution, 3×3 convolution, and 1×1 convolution followed by batch normalization and ReLU as a block. On this basis, RegNetx replaces the standard 3×3 convolution to 3×3 group convolution [42] with the hyperparameter g_i to optimize the rudimentary residual bottleneck structure in the block. Meanwhile, classic backbones, e.g., ResNet, keep the same expansion ratio of block width (number of feature layers) among stages, and manually set the depth of network blocks, e.g., 3, 4, 6, and 3 depths of network blocks for stage 1 to 4 in ResNet-50, respectively. Relatively, RegNetx interpretably parametrizes the width and depth of network blocks among stages with a quantized linear function. First, the width v_i of the $i-th$ network block is computed via a linear parameterization:

$$v_i = w_0 + w_a \cdot i \quad r.t.\ 0 \leq i < d, \tag{1}$$

where the default parameters $w_0 > 0$, $w_a > 0$, and d represent initial width, slope, and network depth, respectively. However, as v_i should be an integer, we supplement the default constraint w_m to compute s_i via the following formulation:

$$v_i = w_0 \cdot w_m^{s_i} \quad r.t.\ 0 \leq i < d. \tag{2}$$

Then, s_i is rounded to compute the quantized width u_i of the $i-th$ network block as follows:

$$u_i = w_0 \cdot w_m^{\lfloor s_i \rceil} \quad r.t.\ 0 \leq i < d. \tag{3}$$

Considering the width of each network block is restricted by the hyperparameter g_i of group convolution, u_i is further normalized to the integer multiple of g_i via:

$$\tilde{u}_i = \lfloor u_i / g_i \rceil \cdot g_i \quad r.t.\ 0 \leq i < d, \tag{4}$$

where $\lfloor * \rceil$ represents the rounding operation. Finally, the network blocks with the same width $\tilde{u}_i$ constitute a certain stage of RegNetx. From a quantitative point of view, give the hyperparameters w_0, w_a, w_m, g_i, and d, the width $\tilde{u}_i$ of the $i-th$ residual block is obtained, which simultaneously defines the universal RegNet. Meanwhile, by employing the flop regime [43], hyperparameters w_0, w_a, w_m, g_i, and d of top model performance define the design space of RegNetx.

Compared to classic backbone such as ResNet, RegNetx inherits its merit of the shortcut connection and further explores the designing space from block and stage to the whole backbone network structure. Based on the general describable network architecture in Figure 2 but with distinct hyperparameter settings of RegNetx, the output width of each stage, the number of blocks, and group ratio for ResNet and RegNetx are summarized in Table 1. Obviously, RegNetx has reduced output width and flexible expansion ratio of output width between consecutive stages. Moreover, by implementing group convolution for each network block, the model size of RegNetx is more lightweight compared to ResNet.

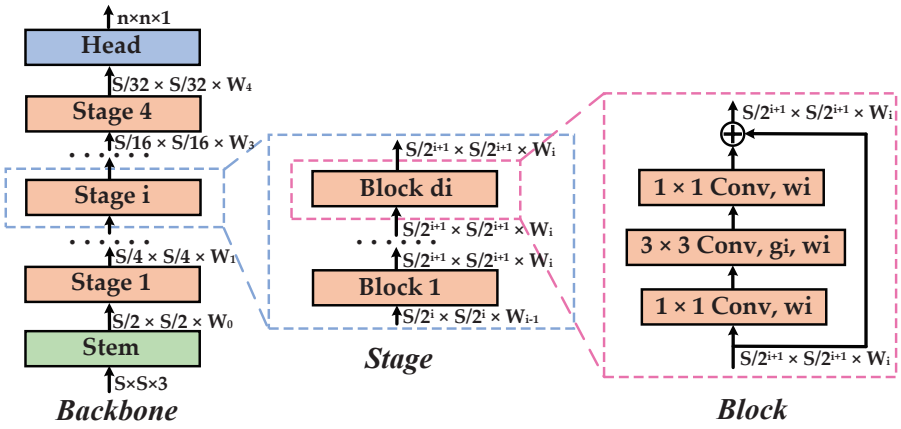

Figure 2. The network architecture and structural details of RegNetx. Following the classic scheme of the stage, block, and residual bottleneck structure, RegNetx optimizes the width and depth of network blocks compared to the classic backbone networks such as ResNet.

Table 1. Comparison of ResNet and RegNetx. Please note that S_1 to S_4 represent stage 1 to stage 4 of the backbone network, respectively.

Backbone Network	Stage Output Width				Num of Blocks				Group Ratio
	S_1	S_2	S_3	S_4	S_1	S_2	S_3	S_4	g_i
ResNet-50	256	512	1024	2048	3	4	6	3	✘
ResNet-101	256	512	1024	2048	3	4	23	3	✘
RegNetx-3.2GF	96	192	432	1008	2	6	15	2	48
RegNetx-4.0GF	80	240	560	1360	2	5	14	2	40

3.1.2. Multi-Level Feature Extraction Network

In top-down instance segmentation networks, FPN shows notable performance of multi-scale instance segmentation. As the edges and instance parts of low-level features can improve the localization capability of FPN, we introduce the bottom-up path augmentation (BPA) for FPN to improve the semantic representation of output feature maps. The lowest level of FPN is regarded the same as BPA. For the upper layer $\boldsymbol{B}_i$ of BPA, it is constructed from $\boldsymbol{B}_{i-1}$ and the FPN layer $\boldsymbol{F}_i$ via:

$$\boldsymbol{B}'_i = \boldsymbol{F}_i + Conv_{3\times3}(\boldsymbol{B}_{i-1}; \theta_1), \tag{5}$$

$$\boldsymbol{B}_i = Conv_{3\times3}(\boldsymbol{B}'_i; \theta_2), \tag{6}$$

where θ_1 and θ_2 represent the weight for each 3×3 convolution layer. As an extension of FPN, the output $\boldsymbol{B}_i$ of BPA is regarded as the output multi-level feature map of AFEN. As illustrated in Figure 3, the backbone network and multi-level feature extraction network constitute the overall network architecture of our AEFN.

3.2. The RoI Extractors

As for top-down instance segmentation methods, RPN is responsible for preliminarily predicting the candidate region proposals, which initially screens out the positive samples among the predictions. To map the coordinate-based region proposals to the multi-level feature from FPN, [31] proposed the RoI extractor which selects matched region proposals for each output level of FPN and pools them with RoI Pooling to generate RoIs for object detection. Based on the previous exploration of researchers, we have designed corresponding RoI extractors for our CPISNet.

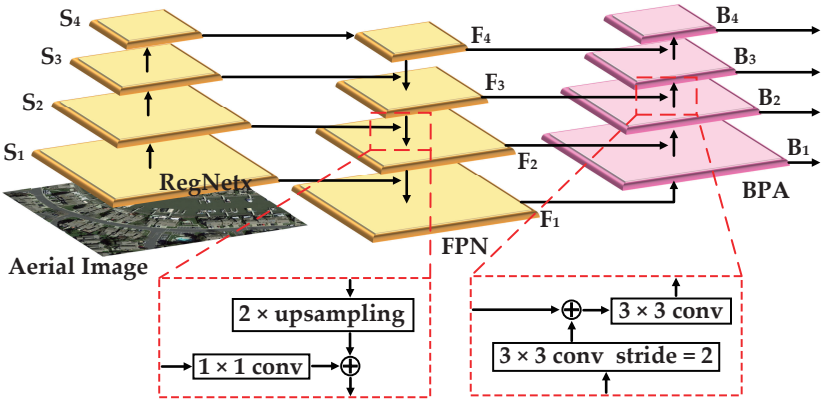

Figure 3. The network architecture of AFEN.

3.2.1. Single RoI Extractor

Generally, in the top-down instance segmentation methods, the speed of mask prediction is limited to object detection as it executes the detect-then-segment formula. The decrease of object detection speed will iteratively slow down the segmentation speed, and ulteriorly influence the network speed. Therefore, we adopt the single RoI extractor for each stage of our subsequent object detection network here.

Assuming the output multi-level feature map from FPN are $\{\boldsymbol{F}_0, \boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$, and the initially screened out $i-th$ bounding box from RPN is denoted as $\{x_b, y_b, x_t, y_t\}$ in Cartesian coordinate, where $\{x_b, y_b\}$ and $\{x_t, y_t\}$ represent the bottom left and top right coordinate of the bounding box, respectively. Therefore, area S_i of the $i-th$ bounding box is calculated as:

$$S_i = (x_t - x_b) * (y_t - y_b). \tag{7}$$

Following the above-mentioned bounding box area S_i, the level of $i-th$ bounding box is calculated as:

$$k = \left\lfloor log_2(\sqrt{S}/56) \right\rfloor, \tag{8}$$

where k is related to the $k-th$ level of FPN level; the denominator 56 denotes the smallest threshold scale of 56^2 for $0-th$ level mapping that is defined by the canonical ImageNet [44] pre-training size. Following the schedule, each bounding box is mapped to a certain level of FPN. Next, the bounding box and correspondig FPN level are pooled by RoIAlign [14] to generate the RoI via:

$$\boldsymbol{RoI}_i = RoIAlign(\boldsymbol{F}_i; (x_b, y_b, x_t, y_t)), \tag{9}$$

where $\boldsymbol{RoI}_i$ represents the $i-th$ RoI pooled by the $i-th$ bounding box and $\boldsymbol{F}_i$. In this paper, we present the single RoI extractor (SRoIE) to extract the RoIs prepared for object detection branch. The architecture of SRoIE is shown in Figure 4.

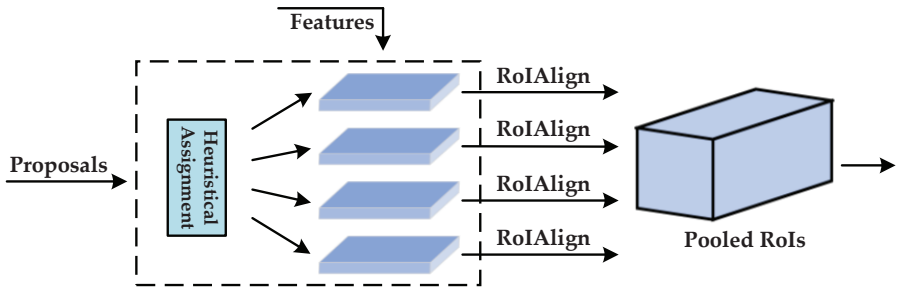

Figure 4. Illustration of SRoIE. It allocates the RPs from FPN with heuristical assigment criterion (the formulation in Equation (8)).

3.2.2. Elaborated RoI Extractor

Distinguished from the heuristically selected schedule in SRoIE, we select the pre elaborate, aggregate, and post elaborate schedule to construct our elaborated RoI extractor (ERoIE). The architecture of ERoIE is illustrated in Figure 5.

Objects (e.g., planes, harbors, and helicopters) in aerial images have geometric variations due to overlooking angle, local characteristics, etc., which may impede the network from integrally presenting the shape of an object. Consequently, we choose the dynamic convolutional network (DCN) [45,46] to deal with such variations. Assuming the output multi-level feature maps from FPN are $\{\boldsymbol{F}_0, \boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3\}$, along with the stride of $\{4, 8, 16, 32\}$ (corresponding to the original image) for RPN. All the region proposals (RPs) from RPN are pooled within $\boldsymbol{F}_i$ by RoIAlign via:

$$\boldsymbol{RoI}_{i-th} = RoIAlign(\boldsymbol{F}_i; RPs), \tag{10}$$

where $\boldsymbol{RoI}_{i-th}$ represents the pooled RoIs in $i-th$ level. Here, all the RPs are regarded as the indispensable elements for RoI pooling. Then, each $\boldsymbol{RoI}_{i-th}$ are preliminarily elaborated by the 5×5 dynamic convolution:

$$\boldsymbol{DRoI}_{i-th} = dynamic\_conv_{5\times5}(\boldsymbol{RoI}_{i-th}), \tag{11}$$

where *dynamic_conv* denotes the dynamic convolutional network. More details see [45,46]. Next, the $\boldsymbol{DRoI}_{i-th}$ for each level is aggregated via the element-wise addition:

$$\boldsymbol{RoI} = \sum_{i=0}^{3} \boldsymbol{DRoI}_{i-th}. \tag{12}$$

Finally, we adopt the global context block (GCB) to post elaborate the aggregated RoIs via:

$$\boldsymbol{ERoI} = \boldsymbol{RoI} + Conv_{1\times1}(RL(Conv_{1\times1}(\boldsymbol{\beta}_j; \gamma_1)); \gamma_2), \tag{13}$$

$$\boldsymbol{\beta}_j = \boldsymbol{RoI} * (softmax(Conv_{1\times1}(\boldsymbol{RoI}; \gamma_3))), \tag{14}$$

where γ_1, γ_2, and γ_3 are the weight for each 1×1 convolution; RL represents the consecutive *ReLU* and *Layer Normalization* operation. $\boldsymbol{\beta}_j$ represents the global context feature weighted by softmax function. The output $\boldsymbol{ERoI}$ is regarded as the elaborated RoI feature of our ERoIE.

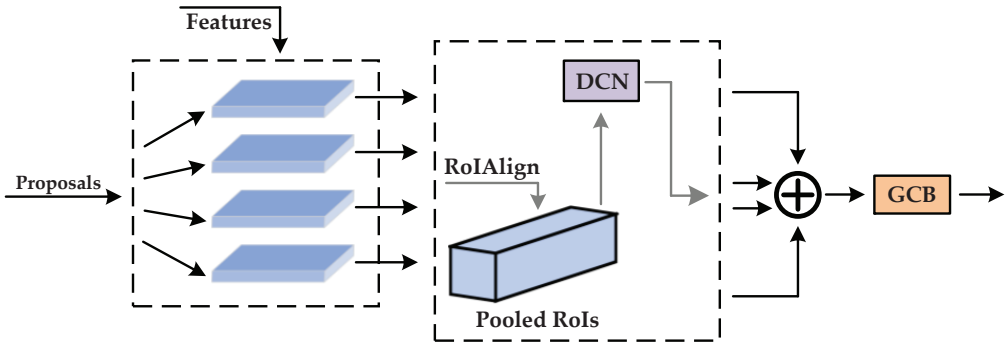

Figure 5. Detailed architecture of ERoIE. Each level of the pooled RoI is preliminarily elaborated by DCN, aggreagted by element-wise addition, and post elaborated by GCB to generate the RoI for mask prediction.

3.3. Proposal Consistent Cascaded Architecture for Instance Segmentation

Cascaded architecture is first introduced in object detection. Cai et al. [32] proposed a stage by stage object detector termed Cascade R-CNN, which leverages the output of previous stage to meet the demand of high-quality sample distribution of next stage.

Similar to the formula of extending Faster R-CNN to Mask R-CNN, Cascade Mask R-CNN attaches a mask branch paralleling to the object detection branch in each stage to exert instance segmentation, which can be formulated via:

$$\boldsymbol{\mathcal{R}}_t^{Box} = \mathcal{P}_b(\boldsymbol{x}; Pb_{t-1}), \tag{15}$$

$$\boldsymbol{\mathcal{R}}_t^{Mask} = \mathcal{P}_m(\boldsymbol{x}; Pb_{t-1}), \tag{16}$$

$$Pb_t = \mathcal{B}_t(\boldsymbol{\mathcal{R}}_t^{Box}), \tag{17}$$

$$Pm_t = \mathcal{M}_t(\boldsymbol{\mathcal{R}}_t^{Mask}), \tag{18}$$

where $\boldsymbol{\mathcal{R}}_t^{Box}$ and $\boldsymbol{\mathcal{R}}_t^{Mask}$ represent the pooled bounding box RoI features by bounding box RoI extractor $\mathcal{P}_b$ and the pooled mask RoI features by mask RoI extractor $\mathcal{P}_m$ in the $t-th$ stage, respectively. $\boldsymbol{x}$ is the multi-scale feature map from FPN. Pb_t and Pm_t denote the predicted bounding box and predicted mask by bounding box branch $\mathcal{B}_t$ and mask branch $\mathcal{M}_t$, respectively. Obviously, $\mathcal{M}_t$ is individually generated in each stage, causing computationally inefficient.

To exploit the reciprocal relationship between detection and segmentation in cascaded architecture, Chen et al. [36] proposed HTC to interweave them for a joint stage by stage processing. Based on the merits of Cascade Mask R-CNN as Equations (15)–(17), HTC connects the mask branch of each stage as Equations (19) and (20):

$$\mathcal{F}_t = Conv_{1\times1}(\mathcal{M}_t(\mathcal{P}_m(\boldsymbol{x}; Pb_{t-1})); \omega_t), \tag{19}$$

$$\boldsymbol{\mathcal{R}}_t^{Mask} = \mathcal{F}_t + \mathcal{P}_m(\boldsymbol{x}; Pb_t), \tag{20}$$

where $Conv_{1\times1}(*; \omega_t)$ represents 1×1 convolution with the weight ω_t. $\mathcal{F}_t$ is the mask information flow from stage $t-1$ to stage t. $\boldsymbol{\mathcal{R}}_t^{Mask}$ denotes the interweaved mask feature for mask prediction. Intuitive comparison of the cascaded architecture in Cascade Mask R-CNN and HTC is illustrated in Figure 6a,b. Unfortunately, these two cascaded architectures ignore the sample IoU distribution consistency of mask prediction, which potentially exacerbate the instance segmentation precision [47]. In this paper, we introduce the proposal consistent cascaded (PCC) architecture to realize high-quality instance segmentation for high-resolution aerial images with a novel cascaded architecture. The network architecture of PCC is shown in Figure 6c.

In PCC architecture, we inherit the architecture of cascaded bounding box stages in Cascade Mask R-CNN but abandon the additional mask branch in each detection stage to eliminate the disparity of the sample's IoU distribution when training and testing. As an alternative, we attach the mask branch to the last stage of detection branch. The pipeline is formulated as follows:

$$\boldsymbol{\mathcal{R}}_t^{Box} = \mathcal{P}_b(\boldsymbol{x}; Pb_{t-1}), \tag{21}$$

$$Pb_t = \mathcal{B}_t(\boldsymbol{\mathcal{R}}_t^{Box}), \tag{22}$$

$$\boldsymbol{\mathcal{R}}^{Mask} = \mathcal{P}_m(\boldsymbol{x}; Pb_2), \tag{23}$$

$$Pm = \mathcal{M}'_n(\boldsymbol{\mathcal{R}}^{Mask}), \tag{24}$$

where $\boldsymbol{\mathcal{R}}^{Mask}$ is pooled with the refined bounding box in the last stage. $\mathcal{M}'_n$ is the mask branch which contains n consecutive blocks with stacked convolutions. Each block $\mathcal{M}'$ contains two 3×3 convolution with shortcut connection via:

$$\mathcal{M}' = \boldsymbol{\mathcal{R}}_n^{Mask} + Conv_{3\times3}(Conv_{3\times3}(\boldsymbol{\mathcal{R}}_n^{Mask}; \theta_1); \theta_2), \tag{25}$$

where $\boldsymbol{\mathcal{R}}_n^{Mask}$ denotes the input of the $n-th$ block; θ_1 and θ_2 are the weight for each 3×3 convolution. At the structural level, PCC does not just ensure instance segmentation to be performed on the basis of precise localization, but also eliminates the intermediate noisy

boxes of mask prediction. Moreover, moderately adjusting the depth of mask branch can tweak the quality of mask predictions.

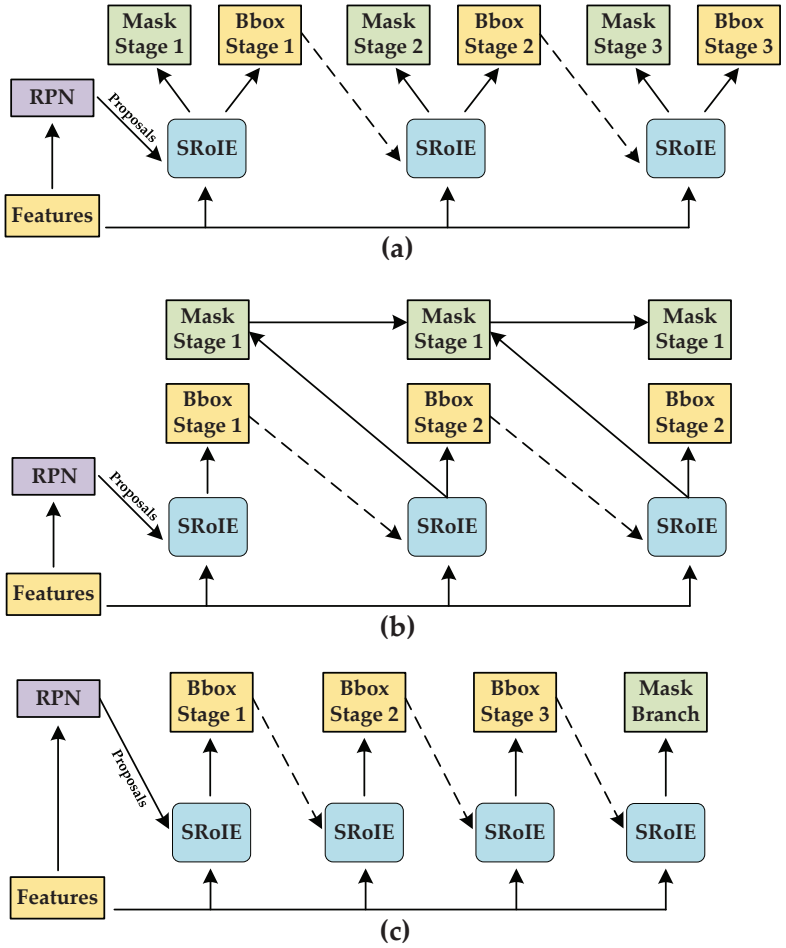

Figure 6. Comparison of the cascaded architectures in Cascade Mask R-CNN, HTC, and PCC. The panels are listed as: (**a**) The cascaded architecture in Cascade Mask R-CNN. (**b**) The cascaded architecture in HTC. (**c**) The PCC in our CPISNet.

4. Experiments

In this section, we will separately introduce the datasets, loss functions, evaluation metrics, and implementation details. Next, experiments on these prerequisites are implemented to verify the effectiveness of our proposed CPISNet.

4.1. The Datasets

We select two mainstream instance segmentation dataset of high-resolution aerial images for experiments, including the Instance Segmentation in Aerial Images Dataset (iSAID [17]) and NWPU VHR-10 instance segmentation dataset [11].

4.1.1. The iSAID

iSAID is the first benchmark instance segmentation dataset for aerial images, akin to Microsoft Common Objects in Context (MS COCO) dataset in the natural scene, with dense but detailed annotation. It contains 2806 large-scale images interspersed with 15 categories, and there are 655,451 object instances in total. To satisfy the demand of real-world application, the object categories including small vehicle (SV), large vehicle (LV), plane (PL), storage tank (ST), ship (SH), swimming pool (SP), harbor (HB), tennis court (TC), ground track field (GTF), soccer ball field (SBF), baseball diamond (BD), bridge (BR), basketball court (BC), roundabout (RA), and helicopter (HC) are chosen according

to the frequency of occurrence and earth observation importance. iSAID is divided into three parts: 1/2 original images for constructing the training set, 1/6 original images for the validation set, and 1/3 original images for the test set. As existing methods cannot handle the large spatial dimension of high-resolution images in iSAID, we split each image into 800×800 pixels through a sliding window with the stride of 200 pixels in length and width. Consequently, there are 18,732 images for training, 9512 images for validation, and 19,377 images for testing.

4.1.2. The NWPU VHR-10 Instance Segmentation Dataset

The NWPU VHR-10 instance segmentation dataset is the extended version of NWPU VHR-10 dataset [48,49] by [11], which provides the pixel-wise annotation for each instance. There are 10 object categories including airplane (AI), baseball diamond, ground track field, vehicle (VC), ship, tennis court, harbor, storage tank, basketball court, and bridge in total. The dataset consists of 650 very high-resolution (VHR) aerial images with targets and 150 VHR images with pure background. In our experiments, it is divided into the training set (70% images) and the test set (30% images) for training and testing, respectively.

4.2. Evaluation Metrics

Following the instance segmentation in natural scenes, we adopt the MS COCO evaluation metrics to evaluate the effectiveness of the methods. Similar to object detection, the AP of instance segmentation result is defined over the IoU, which is calculated through the overlap ratio of predicted mask and ground truth mask:

$$IoU_{mask} = \frac{M_p \cap M_g}{M_p \cup M_g}, \tag{26}$$

where M_p and M_g denote the predicted mask and the ground truth mask, respectively. Based on a certain IoU threshold, the precision and recall value is defined by the instance-wise classification results via:

$$Precision = \frac{TP}{TP + FP}, \tag{27}$$

$$Recall = \frac{TP}{TP + FN}, \tag{28}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively. Meanwhile, the AP of the predicted results is calculated through:

$$AP = \int_0^1 P(r)\, dr, \tag{29}$$

where P is the precision value, r is the recall value. Generally, the AP value is calculated by averaging 10 IoU threshold, where the IoU threshold value ranges from 0.5 to 0.95 with the stride of 0.05. In addition to the AP, MS COCO evaluation metrics also include the single threshold AP for instance AP_{50} (IoU = 0.5) and AP_{75} (IoU = 0.75). Moreover, AP_S, AP_M, and AP_L are responsible for measuring the AP of small (area$< 32^2$ pixels), medium ($32^2 <$area$< 96^2$ pixels), and large (area$> 96^2$ pixels) instance, respectively.

4.3. The Loss Functions

For simplicity, we choose cross entropy loss function for object classification, which is defined as:

$$L_{cls}(x, class) = -x_{class} + log(\sum_{i=0}^{c-1} e^{x[i]}), \tag{30}$$

where *class* is the ground truth category label; x and c denote the predicted probability of a certain category and the number of categories, respectively. The $smooth_{l1}$ loss is responsible for regressing the bounding boxes via:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N} smooth_{l1}(p_i - g_i), \tag{31}$$

$$smooth_{l1}(x) = \begin{cases} 0.5x^2, & if|x| < 1 \\ |x| - 0.5, & otherwise \end{cases}, \tag{32}$$

where p_i is the predicted bounding box, and g_i is the ground truth bounding box. N denotes the number of the predicted bounding boxes. Following [14], we select binary cross entropy (BCE) loss for mask prediction, which can be represented via:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(T_{i(x,y)}log(\hat{P}_{i(x,y)}) + (1 - T_{i(x,y)})log(1 - \hat{P}_{i(x,y)})), \tag{33}$$

$$\hat{P}_{i(x,y)} = \frac{1}{1 + e^{-P_{i(x,y)}}}, \tag{34}$$

where P_i denotes the predicted pixel with coordinate (x, y) in the predicted mask; T_i is ground truth with coordinate (x, y) in the ground truth mask.

4.4. Implementation Details

All the models in our experiments are coded with Pytorch framework. A single RTX 3090 with 24 GB memory is adopted for training and testing the models. We select the stochastic gradient descent (SGD) as the optimizer for each model. In the training phase, with the initial learning rate of 0.0025, each model is trained for 12 epochs with mini-batch size of 2, and the learning rate is decreased by 0.1 at $8 - th$ and $11 - th$ epochs. As for image size, each image in the NWPU VHR-10 instance segmentation dataset is resized to the size of 1000×600 pixels for training and testing. Moreover, soft non-maximal suppression (Soft-NMS) [50] with the threshold of 0.5 is selected as the bounding box filter. The increasing IoU thresholds for each stage of the cascaded architectures are set at 0.5, 0.6, and 0.7, respectively.

4.5. Ablation Experiments

In this section, we conduct comprehensive experiments on AFEN, ERoIE, and PCC to verify the effects of our proposed CPISNet. All the experiments are based on the Mask R-CNN (meta top-down instance segmentation formula) with ResNet-101 backbone network. Moreover, we select the iSAID validation set to test our instance segmentation results.

4.5.1. Effects of CPISNet

The instance segmentation results of AFEN, ERoIE, and PCC are individually reported in Table 2. Quantitatively, AEFN, ERoIE, and PCC perform well in segmenting aerial objects (gain 0.6%, 0.9%, and 1.9% AP increments, respectively). With regard to CPISNet, it yields 2.6% AP increments than vanilla Mask R-CNN under the same training and testing conditions. With various AP indicators, CPISNet even gains 3.1% AP_{50} increments and 5.3% AP_L increments, respectively.

Table 2. Effects of CPISNet. Please note that all results are evaluated on iSAID validation set.

Model	AFEN	ERoIE	PCC	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN				36.0	58.4	38.8	22.7	**43.3**	49.7
Modules	✓			36.6	59.3	39.6	23.8	43.1	51.7
		✓		36.9	59.2	39.9	23.1	44.0	52.1
			✓	37.9	60.2	41.0	24.0	45.3	53.8
CPISNet	✓	✓	✓	**38.6**	**61.5**	**41.4**	**25.7**	**45.6**	**55.0**

4.5.2. Experiments on AFEN

We have selected several classic feature extraction structures (ResNet, FPN, HRNet [51,52], HRFPN [52]) to verify the effectiveness of AFEN. RegNetx with 3.2-gigabyte flops (3.2-GF) retains the architecture setting of $w_0 = 88$, $w_a = 26.31$, $w_m = 2.25$, $g_i = 48$, $d = 25$, and RegNetx with 4.0 GF means $w_0 = 96$, $w_a = 38.65$, $w_m = 2.43$, $g_i = 40$, $d = 23$. As shown in Table 3, HRNet and RegNetx both serve as the efficient backbone network for instance segmentation in high-resolution aerial images. The structure of HRNetw32-HRFPN and RegNetx3.2GF-FPN achieve 0.3% AP, 0.1% AP better than ResNet101-FPN, respectively. While our proposed AFEN can achieve higher mask prediction precision: 0.3% AP gain from BPA under RegNetx-3.2GF backbone, and 0.6% AP gain with AFEN-4.0GF compared to ResNet101-FPN. Results of ablation experiments indicate that AFEN is efficient in high-quality feature extraction for high-resolution aerial images.

Table 3. Ablation experiment on AFEN. Please note that all results are evaluated on iSAID validation set.

Feature Extraction Structures	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-101 + FPN	36.0	58.4	38.8	22.7	**43.3**	49.7
HRNetv2-w32 + HRFPN	36.3	58.7	39.0	**24.4**	42.5	51.1
RegNetx-3.2GF + FPN	36.1	59.0	38.3	23.9	43.1	49.4
AFEN-3.2GF	36.4	59.1	38.9	24.1	42.9	51.2
AFEN-4.0GF	**36.6**	**59.3**	**39.6**	23.8	43.1	**51.7**

4.5.3. Experiments on ERoIE

In this subsection, we implemented three stages of experiments, including effects of the preliminarily elaborated module, effects of the post elaborated module, and effects of the integral ERoIE, to verify the rationality of ERoIE.

1. *Stage 1: Effects of the Preliminarily Elaborated Module*

On the basis of experiments in [53], we follow the criterion of selecting the most effective convolution layer for the preliminarily elaborated module, and set DCN with the kernel size of 5 here to be consistent with the previous statement in Section 3.2.2. Ulteriorly, we compare the effects of the single-level and fused-level elaborated strategy when implementing DCN as the preliminarily elaborated module (element-wise experiments). Please note that DCN is additionally selected as the default post elaborated module here. Assuming the pooled RoI features from RoIAlign are $\boldsymbol{B}_2$, $\boldsymbol{B}_3$, $\boldsymbol{B}_4$, $\boldsymbol{B}_5$, the corresponding feature maps are recorded as $\boldsymbol{B}_2 - level$, $\boldsymbol{B}_3 - level$, $\boldsymbol{B}_4 - level$, $\boldsymbol{B}_5 - level$, respectively. As shown in Table 4, DCN for $\boldsymbol{B}_3 - level$ elaboration and post-processing outperforms remaining forms up to 0.3% AP in the single-level elaborated strategy, which is the same as $\boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$ in the fused-level elaborated strategy.

Table 4. Effects of the preliminarily elaborated module. Please note that all results are evaluated on iSAID validation set.

Elaborated Layer	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$\boldsymbol{B}_0 - level$	36.4	58.6	39.4	**23.1**	43.4	51.3
$\boldsymbol{B}_1 - level$	36.6	58.7	39.7	23.0	43.8	51.2
$\boldsymbol{B}_2 - level$	**36.7**	**58.9**	**39.9**	22.9	44.0	51.5
$\boldsymbol{B}_3 - level$	36.5	58.7	39.5	22.7	**44.0**	**52.0**
$\boldsymbol{B}_1 + \boldsymbol{B}_2$	36.5	58.6	39.3	23.0	43.8	50.9
$\boldsymbol{B}_2 + \boldsymbol{B}_3$	36.4	58.5	39.2	22.7	43.5	51.5
$\boldsymbol{B}_1 + \boldsymbol{B}_3$	36.4	58.5	39.3	22.9	43.8	51.6
$\boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$	36.7	58.6	39.7	23.0	43.6	51.6
$\boldsymbol{B}_0 + \boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$	36.4	58.5	39.4	22.3	43.8	50.7

2. *Stage 2: Effects of the Post Elaborated Module*

Stage 2 focuses on evaluating the effects of the post elaborated module to ERoIE. In this context, we individually measure the global enhancement capability of GCB and DCN for post-processing. Without loss of generality, we replace the DCN to GCB for post elaboration in stage 1. As shown in Table 5, effects of GCB are similar to DCN in the single elaborated strategy. In particular, $\boldsymbol{B}_0 + \boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$ in the fused-level elaborated strategy yields 0.6% AP than $\boldsymbol{B}_0 - level$ in the single-level elaborated strategy. Therefore, we select the DCN with fused-level elaborated strategy of $\boldsymbol{B}_0 + \boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$ to preliminarily elaborate the RoIs, and the GCB to post elaborate the aggregated RoIs to formulate ERoIE.

Table 5. Effects of the post elaborated module. Please note that all results are evaluated on iSAID validation set.

Elaborated Layer	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$\boldsymbol{B}_0 - level$	36.3	58.6	39.3	22.9	43.4	51.1
$\boldsymbol{B}_1 - level$	36.4	58.9	39.1	**23.2**	43.5	51.2
$\boldsymbol{B}_2 - level$	36.7	58.8	39.8	22.9	43.7	51.8
$\boldsymbol{B}_3 - level$	36.6	58.7	39.6	22.8	43.7	51.5
$\boldsymbol{B}_1 + \boldsymbol{B}_2$	36.7	59.1	39.7	23.1	43.8	51.7
$\boldsymbol{B}_2 + \boldsymbol{B}_3$	36.5	59.0	39.2	22.8	43.7	51.1
$\boldsymbol{B}_0 + \boldsymbol{B}_1$	36.6	58.9	39.4	23.1	43.9	50.7
$\boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$	36.6	58.7	39.4	22.6	43.9	51.6
$\boldsymbol{B}_0 + \boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3$	**36.9**	**59.2**	**39.9**	23.1	**44.0**	**52.1**

3. *Stage 3: Effects of the Integral ERoIE*

Stage 3 tends to research the effects of the integral ERoIE formula. Results are shown in Table 6. Without appendages, ERoIE has a similar performance to SRoIE. When omitting preliminary elaboration, adding post-processed GCB/DCN can improve 0.3% and 0.6% AP, respectively. As for our integral ERoIE (best result in Table 6), it yields SRoIE 0.9% AP compared to SRoIE, which verifies the effectiveness of DRoIE in instance segmentation for high-resolution aerial images.

Table 6. Ablation experiment on ERoIE. Please note that all results are evaluated on iSAID validation set.

Effects of Integral ERoIE	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SRoIE	36.0	58.4	38.8	22.7	43.3	49.7
ERoIE without appendages	36.0	58.4	38.7	22.1	43.2	50.3
+post GCB	36.3	58.8	39.2	22.4	43.7	51.4
+post DCN	36.6	59.0	39.6	**23.3**	43.9	51.3
ERoIE	**36.9**	**59.2**	**39.9**	23.1	**44.0**	**52.1**

4.5.4. Experiments on PCC

In this subsection, we have implemented two groups of experiments, including selecting the depth of mask branch and the effects of PCC, to verify the rationality of PCC.

1. *Group 1: Selecting the Depth of Mask Branch*

Distinguished from the scattered mask branch (contains four consecutive convolution layers) in each stage of the cascaded architecture in Cascade Mask R-CNN and HTC, the mask branch in PCC stacks the consecutive convolution layers with shortcut connection within 2 convolution layers (denoted as a block). Here, the depth of mask branch is equal to the number of blocks. As shown in Table 7, with a gradually increasing number of blocks, PCC yields 0.6% AP increments until 8 blocks. Meanwhile, with over 8 blocks, the AP of PCC begins to drop. It is worth mentioning that even with 2 blocks (equal to the number of convolution layers in the scattered mask branch), PCC improves 1.3% AP based on vanilla Mask R-CNN, which additionally verifies the effectiveness of PCC.

Table 7. Ablation experiment on the depth of mask branch. Please note that all results are evaluated on iSAID validation set.

Number of Blocks	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
2	37.3	59.8	40.1	23.8	44.4	53.4
4	37.6	60.2	40.8	23.6	44.9	53.2
6	37.6	60.2	40.7	23.3	45.2	53.3
8	**37.9**	60.2	**41.0**	**24.0**	**45.3**	**53.8**
10	37.7	**60.4**	40.7	23.9	45.1	53.4

2. *Group 2: Effects of PCC*

Group 2 tends to evaluate the superiority of PCC in the structural level. Therefore, we compare the performance of PCC with Cascade Mask Branch (Cascaded architecture in Cascade Mask R-CNN) and Mask Information Flow (Cascaded architecture in HTC). Table 8 lists the instance segmentation results of the cascaded architectures with ResNet-50 and ResNet-101 backbone network. Compared to Cascaded Mask Branch and Mask Information Flow, PCC with ResNet-50 respectively outperforms 1.0% AP and 0.4% AP, which is the same as PCC with ResNet-101. Moreover, PCC maintains significant increments in threshold AP (AP_{50} and AP_{75}) and area AP (AP_S, AP_M, and AP_L).

Table 8. Ablation experiment on the effects of multiple cascaded architectures for instance segmentation in aerial images. Please note that all results are evaluated on iSAID validation set.

Cascaded Architectures	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Cascaded Mask Branch	R-50	36.0	58.0	38.7	23.7	42.9	48.9
	R-101	36.9	59.1	40.3	23.1	44.1	51.6
Mask Information Flow	R-50	36.6	59.1	39.3	23.7	43.7	51.3
	R-101	37.5	60.1	40.5	23.2	44.7	53.6
PCC	R-50	37.0	58.8	40.1	24.1	44.1	52.4
	R-101	**37.9**	**60.2**	**41.0**	**24.0**	**45.3**	**53.8**

4.6. Instance Segmentation Results on iSAID

To measure the instance segmentation capability of CPISNet in the integral model level, we select five state-of-the-art top-down instance segmentation methods, containing Mask R-CNN, Mask Scoring R-CNN (MS R-CNN), Cascade Mask R-CNN (CM R-CNN), HTC, and SCNet, with the default training and testing hyperparameters as in [54], except for the dedicated hyperparameters that introduced in Section 4.4, for a fair comparison with CPISNet. All the state-of-the-art methods adopt ResNet-101 and FPN for multi-scale feature extraction; the momentum and weight decay for SGD are set at 0.9 and 0.0001, respectively. Correspondingly, CPISNet adopts AFEN-4.0GF for multi-scale feature extraction here; the momentum and weight decay for SGD are set at 0.9 and 0.00005, respectively. Meanwhile, we add the frames per second (FPS) and model size to evaluate the practical engineering application ability of each method. As presented in Table 9, our CPISNet achieves the highest 38.6% AP compared to other methods. As for the non-cascaded methods, CPISNet yields 2.6% and 1.7% AP increments than Mask R-CNN and MS R-CNN with similar model size, respectively. Relatively, compared to the cascaded methods, CPISNet still maintains over 1% AP increments (1.7% AP, 1.2% AP, and 1.3% AP increments than CM R-CNN, HTC, and SCNet, respectively) with reduced model size.

Considering the scale variance of objects in high-resolution aerial images, we further introduce the multi-scale training strategy to improve the scale sensitivity of our CPISNet, termed CPISNet*. While training, the aerial images are rescaled to the size of 1200 × 800, 1000 × 800, 800 × 800, 600 × 800, and 400 × 800 pixels. As for testing, the size of aerial images retains 800 × 800 pixels. Without bells and whistles, CPISNet* further improves 0.8% AP with the same model size as CPISNet and slightly inferior FPS. In general, CPISNet* yields Mask R-CNN 3.4% in AP. With various threshold AP indicators, AP_{50} and AP_{75} improve 4% and 3.6%, respectively. Moreover, CPISNet* outperforms vanilla Mask R-CNN 3.9%, 3.3%, and 4.5% in segmenting small, medium, and large objects in high-resolution aerial images, respectively. Qualitatively, we provide the comparison of visualized instance segmentation results of vanilla Mask R-CNN and our proposed CPISNet* in Figure 7. As illustrated in row 2, Figure 7, the instance segmentation of Mask R-CNN retains aliasing masks, missing segmentations, and poorly segmented mask. Fortunately, our proposed CPISNet* can effectively suppress such defects in instance segmentation for high-resolution aerial images.

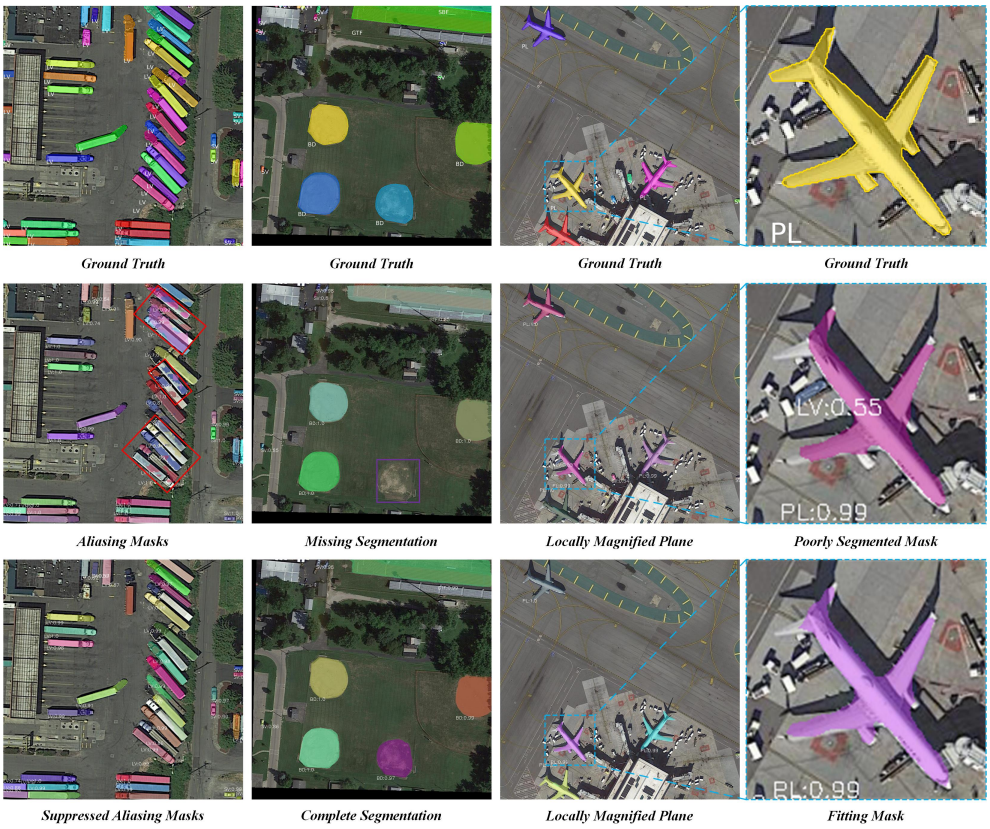

Figure 7. Comparison of visualized instance segmentation results of vanilla Mask R-CNN and our proposed CPISNet* in iSAID. Row 1 to row 3 denote ground truth, Mask R-CNN results, and CPISNet* results, respectively. Please note that red rectangle, purple rectangle, and blue dotted rectangle in row 2 represent aliasing masks of dense objects, missing segmentation, and poorly segmented mask, respectively.

Table 9. Instance segmentation results of the state-of-the-art methods on iSAID validation set.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS	Model Size
Mask R-CNN	36.0	58.4	38.8	22.7	43.3	49.7	13.6	504.2 Mb
MS R-CNN	36.9	58.3	40.3	22.7	44.0	51.9	12.9	634.4 Mb
CM R-CNN	36.9	59.1	40.3	23.1	44.1	51.6	11.5	768.4 Mb
HTC	37.4	60.2	40.1	23.5	44.6	53.5	7.4	791.9 Mb
SCNet	37.3	59.5	40.3	23.3	44.8	52.3	6.7	908.4 Mb
CPISNet	38.6	61.5	41.4	25.7	45.6	**55.0**	6.1	663.3 Mb
CPISNet*	**39.4**	**62.4**	**42.4**	**26.6**	**46.6**	54.2	5.3	663.3 Mb

To measure the meticulous results of CPISNet*, we report the class-wise *AP* of each method for each aerial category in Table 10. Notably, storage tank achieves 80.5% *AP* (highest *AP* among 15 aerial categories) and ship obtains 7.2% *AP* improvement (highest *AP* improvement among 15 aerial categories) in iSAID validation set. Meanwhile, we observe that for some categories, e.g., tennis court and roundabout, CPISNet* yields ∼5% *AP* improvement than Mask R-CNN. Qualitatively, we have visualized the class-wise instance segmentation results in Figure 8. Please note that each subfigure represents the foremost aerial categories. Identical to the quantitative results, CPISNet* is capable of segmenting the hard samples, e.g., densely distributed objects (row 1, column 3–4), small objects (row 2, column 3–4), and objects with nonrigid boundaries (row 1, column 1 and row 3, column 3), in high-resolution aerial images. Quantitative and qualitative results on iSAID validation set indicate our proposed CPISNet is more effective in segmenting aerial objects than state-of-the-art methods.

Table 10. Class-wise instance segmentation results of the state-of-the-art methods on iSAID validation set.

Method	SV	LV	PL	ST	SH	SP	HB	TC	GTF	SBF	BD	BR	BC	RA	HC
Mask R-CNN	40.2	35.0	54.4	77.6	40.1	29.5	21.9	36.9	11.7	4.0	30.9	34.5	46.6	49.2	27.3
MS R-CNN	40.5	35.0	55.9	77.4	41.7	30.7	23.4	37.7	11.8	5.1	31.5	37.6	47.7	49.9	28.1
CM R-CNN	41.1	35.9	54.4	77.7	**43.5**	30.6	22.9	38.6	12.0	4.6	31.6	35.1	48.0	50.2	27.8
HTC	41.4	35.5	54.6	78.6	42.9	32.4	23.3	39.8	12.3	4.5	32.1	36.2	47.9	50.8	28.4
SCNet	41.8	35.5	**56.6**	78.5	41.2	32.6	21.9	39.8	12.1	3.9	31.6	36.4	47.5	51.6	28.9
CPISNet	42.9	**37.8**	54.6	78.8	41.1	36.6	23.9	41.2	**13.0**	**7.6**	33.7	35.9	48.5	53.4	**30.1**
CPISNet*	**43.6**	37.2	55.6	**80.5**	42.8	**36.7**	**25.0**	**41.8**	12.8	5.8	**35.4**	**39.3**	**49.8**	**54.3**	30.0

Following [17], we further measure the generalization ability of the state-of-the-art methods on iSAID test set. Please note that the quantitative results in Tables 11 and 12 are tested on the official evaluation server. As shown in Table 11, compared to vanilla Mask R-CNN, our CPISNet* achieves even better AP improvement in iSAID test set (3.8%) than that in iSAID validation set (3.4%), which reflects the strong generalization ability of CPISNet*. With various AP indicators, CPISNet* still exceeds vanilla Mask R-CNN over 4% increments. Table 12 reports the class-wise AP of the methods. Intuitively, the small vehicle and helicopter challenge the generalization ability of instance segmentation methods due to the small size and unique geometric variations. However, our proposed CPISNet* not only improves the AP of small vehicle and helicopter, but also remains in the ascendancy for other categories, e.g., 6.7% AP increments for the basketball court.

Table 11. Instance segmentation results of the state-of-the-art methods on iSAID test set.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN	36.2	58.6	38.8	38.9	44.2	12.0
MS R-CNN	37.0	57.8	40.5	39.7	46.0	14.3
CM R-CNN	37.1	59.0	40.1	39.8	46.4	12.9
HTC	37.5	59.6	40.8	40.2	47.4	14.2
SCNet	38.1	60.4	41.2	40.9	46.9	12.6
CPISNet	39.1	62.2	42.5	41.8	49.6	**17.6**
CPISNet*	**40.0**	**62.7**	**43.9**	**42.9**	**50.4**	16.5

Table 12. Class-wise instance segmentation results of the state-of-the-art methods on iSAID test set.

Method	SV	LV	PL	ST	SH	SP	HB	TC	GTF	SBF	BD	BR	BC	RA	HC
Mask R-CNN	13.2	29.6	42.9	34.1	46.1	37.4	29.2	75.4	27.1	36.3	51.3	17.6	49.0	43.3	9.6
MS R-CNN	14.0	30.0	43.8	33.9	46.6	37.9	30.1	76.1	29.7	35.7	54.2	17.7	49.9	44.4	11.8
CM R-CNN	14.2	30.4	43.5	34.5	47.1	38.3	30.5	76.6	28.1	37.4	53.0	18.2	50.5	44.2	10.2
HTC	14.5	31.7	43.9	34.8	47.7	38.7	31.0	77.3	29.7	37.9	53.3	18.9	50.2	43.9	9.3
SCNet	14.2	31.7	45.1	35.9	48.0	39.2	31.1	77.0	30.2	36.3	**56.6**	18.7	51.5	46.7	9.7
CPISNet	14.9	32.9	46.2	35.8	49.5	40.6	32.7	77.6	**31.9**	39.3	54.3	19.9	52.9	45.2	**13.2**
CPISNet*	**14.9**	**34.0**	**47.3**	**36.0**	**50.2**	**41.6**	**33.9**	**78.8**	31.6	**40.2**	56.2	**20.2**	**55.7**	**47.6**	12.0

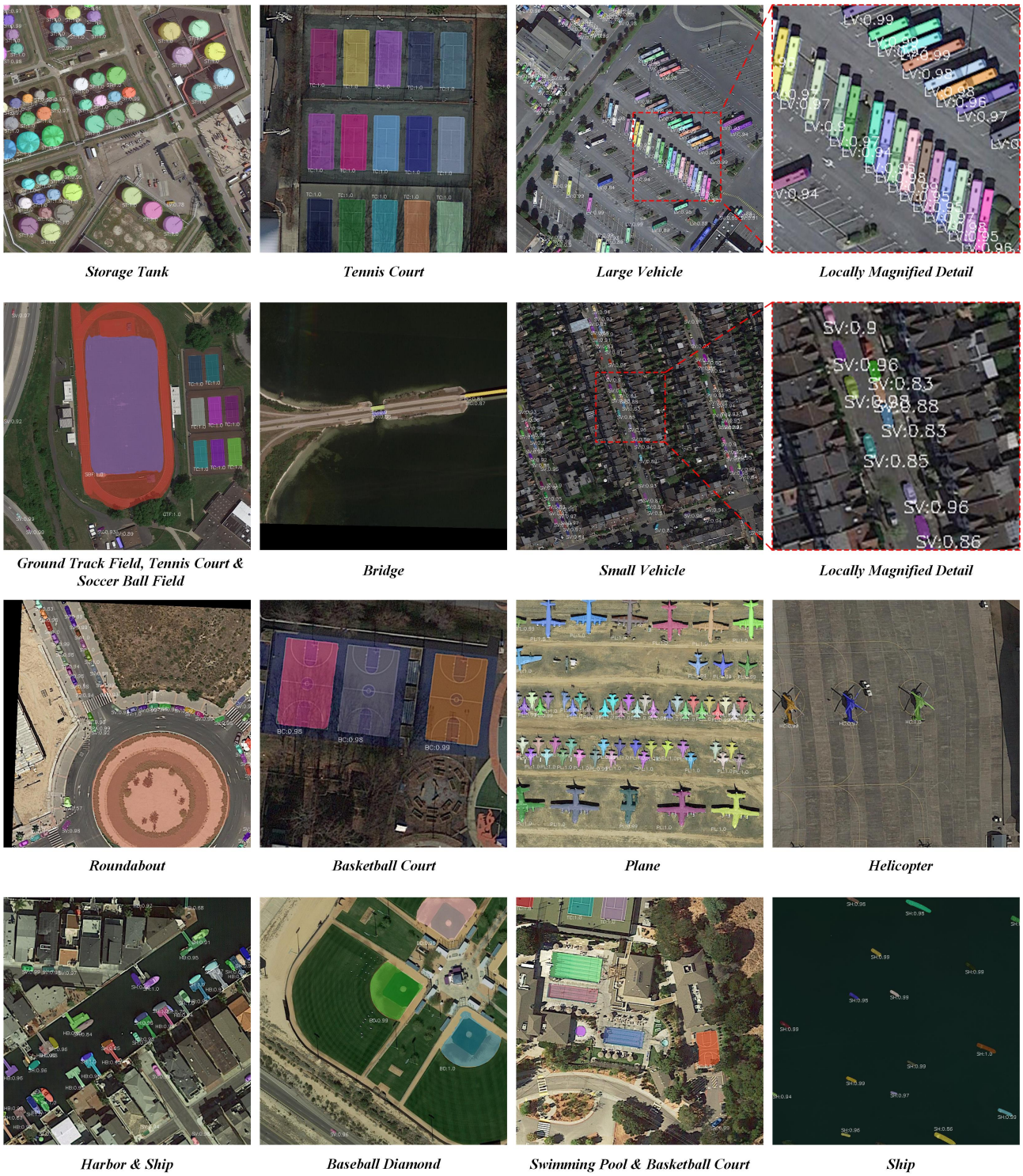

Figure 8. Class-wise instance segmentation results generated by CPISNet*. Red dotted rectangle represents the enlarged detail of small objects.

4.7. Instance Segmentation Results on NWPU-VHR-10 Dataset

Similar to the experiments on iSAID, we supplement the instance segmentation experiments on NWPU VHR-10 dataset to additionally verify the rationality of CPISNet. Still, we select the same control methods as the experiments in iSAID. Considering the image size in NWPU VHR-10 dataset, we define CPISNet* as CPISNet with multi-scale training strategy by rescaling the image size to 1000 × 1200, 1000 × 1000, 1000 × 800, 1000 × 600, and 1000 × 400 pixels. Distinguished from the results on iSAID, CPISNet and CPISNet*

have widened the gap in instance segmentation performance of high-resolution aerial images compared to the state-of-the-art methods. As shown in Table 13, CPISNet* achieves the highest 67.5% AP among the state-of-the-art methods and yields 9.2% AP increments than Mask R-CNN. Compared to SCNet, CPISNet* yields 5.2% AP increments and 27.0% reduced model size but merely 2.1 slowed FPS. With various AP indicators, CPISNet* exceeds over 10% improvements (11.4% in AP_{50} and 16.5% in AP_L) than Mask R-CNN. Moreover, as illustrated in Figure 9, CPISNet* can suppress false alarms, deal with non-rigid boundaries and accurately distinguish the densely distributed objects (unavailable in semantic segmentation).

Table 13. Instance segmentation results of the state-of-the-art methods on NWPU VHR-10 test set.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS	Model Size
Mask R-CNN	58.3	90.9	63.5	46.5	59.6	57.5	12.2	503.9 Mb
MS R-CNN	59.5	90.8	65.2	43.9	61.1	56.8	11.1	634.1 Mb
CM R-CNN	60.4	92.6	67.5	48.1	61.0	63.0	10.6	768.3 Mb
HTC	61.4	92.2	67.0	49.3	62.1	60.8	7.5	791.8 Mb
SCNet	62.3	91.3	69.4	49.8	62.8	68.2	7.1	908.2 Mb
CPISNet	66.1	93.7	73.1	53.3	66.2	**75.5**	5.2	663.1 Mb
CPISNet*	**67.5**	**94.3**	**74.9**	**55.4**	**67.7**	74.0	5	663.1 Mb

Figure 9. Comparison of the visualized instance segmentation results of vanilla Mask R-CNN and our proposed CPISNet* in NWPU VHR-10 instance segmentation dataset. Rows 1 to 3 denote ground truth, Mask R-CNN results and CPISNet* results, respectively. Please note that the red rectangle, orange rectangle, and blue rectangle in row 2 represent aliasing masks of dense objects, false alarm, and poorly segmented mask, respectively.

Next, we report the class-wise instance segmentation results in Table 14. As shown in Table 14, ground track field receives the highest AP value of 92.5% among 10 categories. Moreover, the airplane, tennis court, and basketball court achieve the remarkable AP increments of 14.7%, 14.9%, and 14.0% than Mask R-CNN, respectively. For some particular categories, e.g., bridge with an unbalanced aspect ratio and airplane with the irregular

boundary, the *AP* value has dramatically improved but still very low. Qualitatively, we have visualized the class-wise instance segmentation results in Figure 10. Identical to the quantitative results, the predicted masks of CPISNet* fit the object boundary well. Each category interpreted by CPISNet* completely presents its characteristic.

Table 14. Class-wise instance segmentation results of the state-of-the-art methods on NWPU VHR-10 test set.

Method	AI	BD	GTF	VC	SH	TC	HB	ST	BC	BR
Mask R-CNN	28.4	81.4	84.3	50.6	52.8	59.6	60.7	69.6	69.6	25.8
MS R-CNN	29.6	81.8	85.4	52.5	52.5	61.7	59.6	69.1	72.4	30.3
CM R-CNN	26.3	82.9	86.2	52.5	56.2	64.6	62.9	70.5	72.7	29.4
HTC	28.7	83.3	87.6	54.4	57.9	64.8	63.0	72.3	73.4	28.0
SCNet	32.9	85.8	89.1	55.1	**58.6**	69.5	64.4	70.0	72.9	24.7
CPISNet	41.5	86.2	91.6	57.4	57.6	73.3	**67.6**	74.2	75.7	**35.9**
CPISNet*	**43.1**	**86.2**	**92.5**	**59.7**	58.2	**74.5**	66.6	**74.6**	**83.6**	35.7

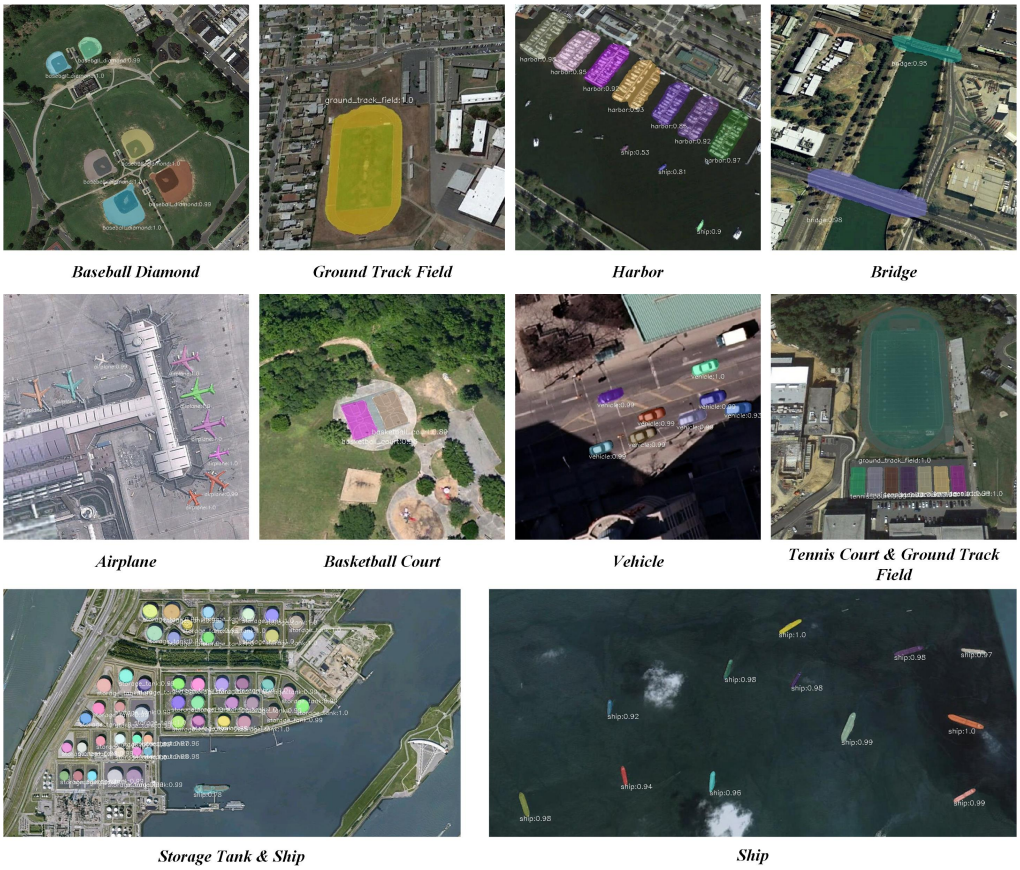

Figure 10. Class-wise instance segmentation results of NWPU VHR-10 instance segmentation dataset generated by CPISNet*.

5. Discussion

Considering the defects of object detection and semantic segmentation in high-resolution aerial images, we employ the instance segmentation to interpret the objects in high-resolution aerial images, which can locate the objects with object boundary, classify the objects in pixel-level, and distinguish the objects within a certain category by a different color. The superiority of instance segmentation in high-resolution aerial images can be observed from Figures 7–10. Despite the effectiveness of CPISNet in segmenting the aerial objects for both iSAID and NWPU VHR-10 instance segmentation dataset, it still encounters difficulties in segmenting the nested objects, e.g., the ground track field and soccer ball

field. Moreover, the objects with small size, large aspect ratio, and irregular boundary challenge the precision of instance segmentation. Future research will focus on tackling the above-mentioned problems and, in addition, improving the FPS of the proposed model under the premise of maintaining high instance segmentation precision in two aspects. First, as the pairwise convolution layers in the mask branch are mapped with shortcut connection, implementing the channel pruning operation may accelerate the procedure of mask prediction (similar to optimizing the backbone network) and further improve the FPS of the model. Second, the reusable bounding box stages in the cascaded architecture of CPISNet may reduce its inference speed. Therefore, to increase the FPS of the model, it is useful to replace the shared fully connected layers to the layers such as global average pooling layer in each bounding box stage.

6. Conclusions

In this paper, we propose a novel instance segmentation network for interpreting multi-category aerial objects, termed CPISNet. The CPISNet follows the top-down instance segmentation formula. First, it adopts the AFEN to extract the multi-level bottom-up augmented feature maps in design space level. Second, ERoIE is designed to extract the mask RoIs via the refined bounding boxes output from PCC and multi-level features output from AFEN. Finally, the convolution block with shortcut connection is responsible for generating the binary mask for instance segmentation. Experimental conclusions can be drawn on the iSAID and NWPU VHR-10 instance segmentation dataset: (1) Each individual module in CPISNet acts on the whole instance segmentation utility; (2)CPISNet* exceeds vanilla Mask R-CNN 3.4%/3.8% *AP* on iSAID validation/test set and 9.2% *AP* on NWPU VHR-10 instance segmentation dataset; (3) The aliasing masks, missing segmentations, false alarms, and poorly segmented masks can be avoided to some extent for CPISNet; (4) CPISNet receives high precision of instance segmentation for aerial images and interprets the objects with fitting boundary.

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

Funding: This research was funded by the National Key Research and Development Program of China (2017YFB0502700), the National Natural Science Foundation of China (61501098), and the High-Resolution Earth Observation Youth Foundation (GFZX04061502).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the anonymous reviewers and editors for their selfless help to improve our manuscript.

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

References

1. 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]
2. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Trans. Image Process.* **2018**, *28*, 265–278. [CrossRef] [PubMed]
3. Zhang, T.; Zhang, X.; Shi, J.; Wei, S. Hyperli-net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 123–153. [CrossRef]

4. Su, H.; Wei, S.; Ming, J.; Wang, C.; Yan, M.; Kumar, D.; Shi, J.; Zhang, X. Precise and robust ship detection for high-resolution sar imagery based on hr-sdnet. *Remote Sens.* **2020**, *12*, 167.
5. Liu, L.; Pan, Z.; Lei, B. Learning a rotation invariant detector with rotatable bounding box. *arXiv* **2017**, arXiv:1711.09405.
6. An, Q.; Pan, Z.; Liu, L.; You, H. Drbox-v2: An improved detector with rotatable boxes for target detection in sar images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 8333–8349. [CrossRef]
7. Mou, L.; Zhu, X.X. Rifcn: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. *arXiv* **2018**, arXiv:1805.02091.
8. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
9. Ding, L.; Tang, H.; Bruzzone, L. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 426–435. [CrossRef]
10. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
11. 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]
12. 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]
13. 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.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
15. 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.
16. 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]
17. Zamir, S.W.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F.S.; 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/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 28–37.
18. Yekeen, S.T.; Balogun, A.-L.; Yusof, K.B.W. A novel deep learning instance segmentation model for automated marine oil spill detection. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 190–200. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
20. Redmon, J.; Farhadi, A. Yolo9000: Better, faster, stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hononlulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
21. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
22. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
23. 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, 8–16 October 2016; pp. 21–37.
24. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
25. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
26. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
27. Yang, X.; Liu, Q.; Yan, J.; Li, A.; Zhang, Z.; Yu, G. R3det: Refined single-stage detector with feature refinement for rotating object. *arXiv* **2019**, arXiv:1908.05612.
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hononlulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
30. Girshick, R. Fast r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Santiago, Chile, 13–16 November 2015; pp. 1440–1448.
31. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
32. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.

33. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hononlulu, HI, USA, 21–26 July 2017; pp. 2359–2367.
34. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
35. 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, 16–20 June 2019; pp. 6409–6418.
36. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4974–4983.
37. 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.
38. 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.
39. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. Solov2: Dynamic and fast instance segmentation. *arXiv* **2020**, arXiv:2003.10152.
40. 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.
41. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing network design spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10428–10436.
42. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hononlulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
43. Radosavovic, I.; Johnson, J.; Xie, S.; Lo, W.-Y.; Dollár, P. On network design spaces for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1882–1890.
44. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
45. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
46. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9308–9316.
47. Vu, T.; Kang, H.; Yoo, C.D. Scnet: Training inference sample consistency for instance segmentation. *arXiv* **2020**, arXiv:2012.10150.
48. 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]
49. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
50. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-nms–improving object detection with one line of code. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
51. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
52. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
53. Rossi, L.; Karimi, A.; Prati, A. A novel region of interest extraction layer for instance segmentation. In Proceedings of the 2020 25th IEEE International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2203–2209.
54. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv* 2019, arXiv:1906.07155.