ITC 1/54

Information Technology
and Control
Vol. 54 / No. 1/ 2025
pp. 99-114
DOI 10.5755/j01.itc.54.1.36913

FACENet: A Fusion Atrous and Channel Enhancement Network for Remote Sensing Image Instance Segmentation

Received 2024/04/08 | Accepted after revision 2024/11/11

HOW TO CITE: Zhao, S., Liu, Z., Cheng, S., Zhang, L., Chen W. (2025). FACENet: A Fusion Atrous and Channel Enhancement Network for Remote Sensing Image Instance Segmentation. *Information Technology and Control*, 54(1), 99-114. https://doi.org/10.5755/j01.itc.54.1.36913

FACENet: A Fusion Atrous and Channel Enhancement Network for Remote Sensing Image Instance Segmentation

Shenhua Zhao

College of Big Data and Information Engineering, Guizhou University, Guiyang, 550025, China

Ziyan Liu

College of Big Data and Information Engineering, Guizhou University, Guiyang, 550025, China;
State Key Laboratory of Public Big Data, Guizhou University, Guiyang, 550025, China

Shitong Cheng, Lihui Zhang, Weidong Chen

College of Big Data and Information Engineering, Guizhou University, Guiyang, 550025, China

Corresponding author: Ziyan Liu, e-mail: gzucomm@gmail.com

The instance segmentation task has been widely used in remote sensing. However, existing remote sensing instance segmentation models may lead to incomplete mask segmentation in complex and diverse background environments. In addition, commonly used feature fusion methods struggle to handle instances of different sizes well and predominantly suffer from loss of semantic information, failing to segment the mask accurately. To solve these problems, we propose a fusion atrous and channel enhancement network (FACENet) for the remote sensing image (RSI) instance segmentation. Specifically, we first replace the FPN with the FACE-FPN, which produces a more detailed pyramid by increasing the receptive field at the feature level. Second, we propose a semantic enhancement module for mining the rich semantic information of the underlying features. Then, we enhance the model's adaptability to complex object deformations by introducing deformable convolution. Experiments on the iSAID, NWPU VHR-10, and HRSID datasets demonstrate that our proposed FACENet outperforms SOLOv2 in terms of average accuracy by 5.1%, 12.9%, and 7.6%, respectively, and beats other instance segmentation models.

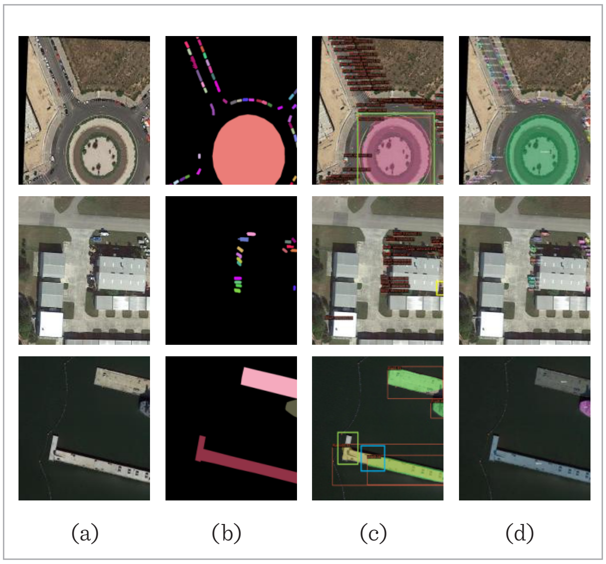
KEYWORDS: Remote sensing image, instance segmentation, SOLOv2, feature fusion, semantic enhancement.

1. Introduction

Deep learning, a fast-growing technology, is considered the most effective remote sensing image processing method, unlike conventional methods such as artificial recognition. Image processing methods based on convolutional neural networks [17] extract and process complex abstract features, resulting in high recognition accuracy and robustness. Therefore, it has been widely used in the field of remote sensing, such as image fusion [19], image classification [26], object detection [36], semantic segmentation [16], and instance segmentation [44].

Figure 1

Visual illustration of iSAID results in RSI instance segmentation. (a) Original images, (b) GT, (c) Mask R-CNN, (d) FACENet



Instance segmentation can distinguish objects of different categories and individuals of the same category in a given image based on individual pixel properties [10], i.e., each foreground object has a different mask. The instance objects in remote sensing images have a range of scales, directions, spectral characteristics, and a lot of interfering noise because most of the images are shot from a top-down perspective and have a variety of complicated backgrounds. In Figure 1, there is a significant scale difference between the first row of densely arranged cars and the roundabout, which causes a lack of clarity in the boundary after segmentation. In addition, the lighting shadows in the second row make the segmentation more difficult. Moreover, insufficient visual features determine the instance's boundary and shape. For example, the partial display of the harbor and ship in the third row can prevent the model from accurately assessing the actual scale of the instances. These issues make the instance segmentation of RSI more challenging. Lin et al. [18] propose Feature Pyramid Networks (FPN) to build semantic information at different scales using a top-down hierarchical structure with side connections. PANet [20] extends FPN with a fusion path, making it easier to transfer information from the bottom layer to the top layer. BiFPN [27] improves detection accuracy by introducing learnable weights. Although the methods described above are effective in feature fusion, they still suffer from inadequate feature fusion [9] and poor detection of irregularly shaped objects.

To address the abovementioned issues, we propose fusion atrous and channel enhancement network for instance segmentation of RSI. Firstly, we apply the FACE-FPN to remote sensing image instance segmentation. This approach significantly mitigates the attenuation of channel information, bolsters the receptive field, and possesses remarkable capabilities for multi-scale feature extraction and fusion. Secondly, to enhance the segmentation performance of the model, we propose a semantic enhancement module designed to capture richer contextual semantic information and nuanced boundary details. Finally, deformable convolution is incorporated into the SOLOv2 header to enhance the network's deformability by adding dynamic offsets. Furthermore, experimental analysis and comparison of the iSAID [41], NWPU VHR-10 [4], and HRSID datasets [34] demonstrate the framework's effectiveness.

Our main contributions are summarized as follows:

1. We introduce FACENet, a multi-scale feature fusion network that achieves noise reduction and feature refinement by effectively bridging contextual semantic and channel information across multiple feature levels.
2. We propose FACE-FPN to improve the network's ability to detect and recognize multi-scale objects by enhancing the feature layer's channel information and receptive field range.

3 We propose a semantic enhancement module that effectively harnesses semantic and texture information to its fullest potential. Furthermore, to enhance the network's flexibility, applicability, and ability to adapt to a diverse array of complex object shapes, we integrate deformable convolution into the SOLOv2 detector head.

4 We experimentally evaluate the proposed method on iSAID, NWPU VHR-10, and HRSID datasets and demonstrate that FACENet performs better in remote sensing image segmentation.

2. Related Works

2.1. Remote Sensing Image Instance Segmentation

In addition to identifying every instance in the image, instance segmentation provides more accurate contour information than object detection, which has a wide range of applications in several fields. Instance segmentation is generally divided into two-stage and single-stage approaches, especially the two-stage approach subdivided into top-down and bottom-up approaches. Mask R-CNN [11] is a classic two-stage algorithm that adds a mask prediction branch to Faster R-CNN [25] to enable instance segmentation. In addition, the two-stage algorithms are PANet, Cascade Mask R-CNN [2], and RefineMask [43]. Single-stage algorithms can perform detection and segmentation tasks in parallel, allowing for end-to-end result outputs such as YOLACT [1], SOLO [31], and QueryInst [7]. Table 1 gives an overview of the different instance segmentation algorithms.

The above instance segmentation algorithms have achieved success in natural and urban scenes. However, performance degradation issues arise when these methods are directly applied to remote sensing images [14], [15]. Fang et al. [6] propose Spectral-Spatial FPN, a feature pyramid network for hyperspectral images, which integrates spectral and spatial features through the bidirectional fusion structure and the attention module. To address the severe scale variation problem in remote sensing images, Liu et al. [21] proposed a context aggregation network to aggregate global contexts

Table 1
The overview of different instance segmentation algorithms

Type	Algorithm	Advantages	Disadvantages
Two-stage	Mask R-CNN	Semantic Segmentation with Faster R-CNN	Rely on frame accuracy
	PANet	Bottom-up enhancement paths; Adaptive feature pooling	-
	MS R-CNN [12]	Add the mask scoring strategy	-
	Cascade Mask R-CNN	A cascade detector to refine the features	Large Target Edge Prediction Roughness
	RefineMask	Fine-grained features can compensate for the loss of specific information	Higher computational costs
Single-stage	YOLACT	Real-time instance segmentation using the parallel processing architecture	Overlap the target position is difficult
	SOLO	Direct segmentation based on the center position and object size	Poor accuracy of small target detection
	QueryInst	Build the model using one-to-one correlation between query and instance	Long training time
	CondInst [29]	The dynamic network output mask directly	Large object instance lack segmentation details
	Mask2former [3]	Introduction of masking techniques and self-attention mechanisms	Inconsistent mask prediction between decoder layers

in the feature, spatial, and instance domains, respectively. LFG-Net [33] captures more texture information by enhancing the receptive field of the underlying features. Yin et al. [39] exploited atrous residual blocks for multilayer and multiscale feature fusion to achieve rapid detection. NAS-HRIS [45] implements remote sensing image segmentation through neural architecture search, which uses the differentiable searching process to learn end-to-end searching rules. Although remote sensing image instance segmentation algorithms have achieved remarkable results, most only apply to a particular remote sensing scene, such as buildings and SAR ships. Therefore, further research and exploration of new methods are required to improve the accuracy and generalization of remote sensing image instance segmentation.

2.2. Feature Fusion

A wide range of viewpoints in high-resolution RSI can result in dense instances appearing in a small portion of a single image, and smaller objects and the contour boundaries of an object frequently fail to achieve good segmentation when the network extracts features [38]. Therefore, enhancing the interaction between shallow and high-level features of multi-scale features is necessary. In convolutional neural networks, the semantic information from shallow features is frequently extracted through a limited number of convolution kernels. The high-level features gain more channels and semantic information as the network grows in depth. However, the learned features must be channel downscaled during the fusion process, which means the channel and semantic information are lost [22]. In addition, direct summation of features with significant semantic differences can lead to reduced expressiveness and relevance of multiscale features. To solve these problems, we propose FACE-FPN to enhance the channel and semantic information representation and perform better segmentation in complex scenarios.

2.3. Atrous Convolution

In image segmentation, convolution and pooling are commonly used to reduce the size of the feature map and increase the receptive field, and subsequent operations require an up-sample to restore the image size [35]. However, the above approach is bound to cause information loss, especially in remote sensing scenarios where detailed information is critical. Yu et al. [40] propose atrous convolution, which can solve the problem. Atrous convolution expands the receptive field of the convolution kernel by introducing an atrous rate and aligning it with the size of the output feature map. It allows the model to capture more contextual information, thus enhancing the network's understanding of the complex structure of RSIs. Furthermore, adjusting atrous rates enhances the network's adaptability without increasing the number of parameters. However, simple stacking of atrous convolutions leads to information loss, which harms the segmentation results. Fu et al. [8] use the nested cascade model to connect atrous convolutions with different atrous rates to provide more practical information. Ma et al. [24] adopt the adaptive atrous rate strategy so that the network adaptively adjusts receptive field size according to the category area. To improve accuracy, FLPK-BiSeNet [28] performed feature fusion via an atrous pyramid pooling layer.

3. Method

3.1. Overall Architecture

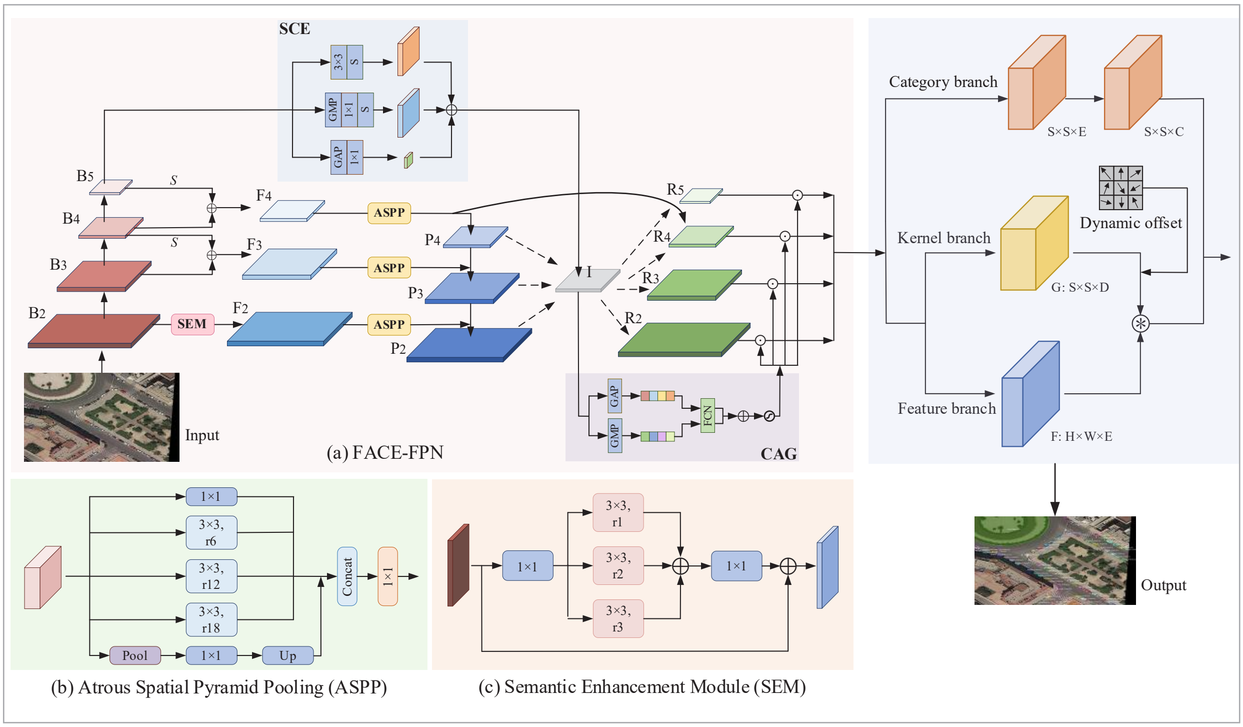
We propose a network that provides as much information as possible to address the challenges of the low completeness of RSI segmentation and inconspicuous contour information in complex scenes. Figure 2 illustrates the FACENet structure, which builds upon the SOLOv2 framework. Firstly, the network splits the input $X \in \mathbb{R}^{H\times W\times 3}$ into the $S \times S$ grid structure and extracts the features through the bottom-up backbone to obtain hierarchical feature tensors $B_i \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$, where $i \in \{1,2,3,4,5\}$, $S_i \in \{2,4,8,16,32\}$ and $C_i \in \{64,246,512,1024,2048\}$. Then, the extracted features are fused by FACE-FPN to obtain information-rich multi-scale features. Among them, the global context information obtained by the high-level feature B_5 through the sub-pixel context enhancement (SCE) module is weighted in the form of weights to the output of FACE-FPN. The head is mainly composed of Category, Kernel, and Feature branches. Among them, the Category branch is primarily responsible for predicting the probability of different instance categories of the grid, for a dataset of C categories, the output dimension of this prediction network is $S \times S \times C$. The Kernel branch is used to learn the convolution kernel $G \in \mathbb{R}^{S\times S\times D}$. Specifically, the feature extraction process involves four convolutional layers. The feature dimension obtained by adding normalized coordinates within the initial convolutional layer is $H \times W \times (D + 2)$, and the final convolutional layer is used for prediction. Furthermore, introducing dynamic offsets gives the kernel features the ability to adapt to spatial deformation. The Feature branch integrates different layers of features into 1/4 to learn the expressive power of mask features. Then, the final instance segmentation result is obtained by realizing non-maximum suppression through efficient Matrix NMS. In our proposed algorithm, the original FPN has difficulty extracting adequate discriminative information from the noise during the fusion process, which can weaken the interaction between the high-level semantic and the underlying edge contour information. Therefore, we integrate FACE-FPN into the network instead of the original FPN to enhance the information flow between different semantic levels, especially to fully integrate high-level global information. In addition, we designed a pluggable module SEM to explore the detailed information of the underlying features and improve the edge profile of the instances.

3.2. Multi-scale Feature Fusion Network

Figure 2(a) illustrates our proposed FACE-FPN framework. Firstly, the RSI generates rich feature representations after the backbone, i.e., $B_2 - B_5$. A series of multi-scale features can be generated for FPN using 1×1 convolution and upsampling layer by layer downward. However, the operations mentioned above lead to the loss of channel information, especially for high-level features. Sub-pixel convolution enables the generation of high-resolution features from low-resolution features by employing convolution and inter-channel reorganization techniques. Additionally, leveraging sub-pixel connections for high-level features facilitates learning intricate details and texture information within pixels, resulting in noise reduction and an enhanced resolution of the extracted features. Higher resolution shows the contours and features of the instance more clearly and helps to segment the target more accurately.

Figure 2

The structure of FACENet. FACE-FPN maps the semantic information to the integrated map I by SCE and weights the features R using CAG. *S* denotes sub-pixel convolution



(a) FACE-FPN

(b) Atrous Spatial Pyramid Pooling (ASPP)

(c) Semantic Enhancement Module (SEM)

Precisely, feature B_5 after sub-pixel convolution is summed with B_4 after 1×1 convolution to obtain feature F_4 and, similarly, feature F_3. This process can be expressed as:

$$F_i = \begin{cases} \varphi(B_i) + S(B_{i+1}) & i = 3,4 \\ \varphi(B_i) & i = 2 \end{cases} \quad (1)$$

where φ represents 1×1 convolution to adjust channels, S denotes sub-pixel convolution with factor r set to 2, and *i* denotes the pyramid levels index.

Subsequently, the obtained features undergo Atrous Spatial Pyramid Pooling (ASPP) to generate richer multi-scale information $F_2^{'}$ – $F_4^{'}$. B_5 is transmitted by the SCE module to fully utilize high-level features and then integrate with P_2 – P_4 to obtain the integrated map *I*. Features R_2 – R_5 are generated by interpolation and maximum pooling, where features R_2 – R_4 are constructed as feature layers at each scale by a top-down hierarchy with $F_2^{'}$ – $F_4^{'}$ side connections. The process of decoupling *I* into the feature *R* is as follows:

$$R_i = \begin{cases} Adaptive\_MaxPool(I) & i = 5 \\ Adaptive\_MaxPool(I) + F_i^{'} & i = 4 \\ upsample(I) + F_i^{'} & i = 3,2 \end{cases} \quad (2)$$

The decoupled features are fused at multiple scales with contextual information captured at different scales, enabling the consideration of objects of varying scales and shapes. In addition, the channel weights extracted from *I* by the channel attention guidance (CAG) module act on the constructed feature layers to generate the final multi-scale features, respectively.

3.2.1. Atrous Spatial Pyramid Pooling

ASPP comprises multiple parallel dilation convolutions with varying dilation rates to obtain more scale feature information by broadening the network's receptive field. Figure 2(b) depicts the module, which includes a 1×1 convolution and three dilation convolutions with different dilation rates, respectively. Furthermore, global average pooling, convolution, and upsampling operations integrate global contextual information into the feature map. Finally, the features generated from the above parallel operations are concatenated together, and their channel number is recovered by 1×1 convolution.

3.2.2. Sub-pixel Context Enhancement

During the fusion process, since the advanced features are standardized to match the number of channels of the underlying features, a certain degree of channel and local information is inevitably lost. Furthermore, this loss of information intensifies with continuous upsampling, resulting in the loss of even more semantic details. Therefore, the sub-pixel context enhancement module is introduced to better use high-level feature-rich channel information. Figure 2(a) depicts the features undergoing a parallel three-branch convolution and pooling operation. Specifically, the upper branch extracts the local information of the features through convolution and 2x sub-pixel upsampling operations; The middle branch acquires the $W \times H \times 16C$ features through global maximum pooling (GMP) and 1×1 convolution operations, followed by 4x sub-pixel upsampling aiming at obtaining the rich context information; The lower branch obtains features with the global information through global average pooling (GAP) and convolution operations. Finally, the three generated features are summed to get the final result. SCE expands the receptive field range of B_5 and captures effective contextual information, enabling the network to utilize the semantic information of B_5 effectively. This process can be described as follows:

$$\mathbf{SCE}(B_5) = S(C_{3\times3}(B_5)) + S(\varphi(GMP(B_5))) + \varphi(GAP(B_5)) \quad (2)$$

where φ represents 1×1 convolution and S denotes sub-pixel convolution with factor r set to 2. $C_{3\times3}$ represents 3×3 convolution.

3.2.3. Channel Attention Guidance Module

To address the aliasing problem of FACE-FPN during cross-scale fusion, we introduced the channel attention guidance module. As shown in Figure 2(a), the features perform the GAP and GMP operations separately in a parallel manner. Then, the obtained features are subjected to fully connected layers, and the results of the two operations are summed. Finally, the sigmoid function obtains the features by fusing the contextual information from different spaces. The module applies the channel weights extracted from the integrated map I to feature R, coalescing the valuable information and enhancing the network's performance. The process can be formulated as:

$$\mathbf{CA}(I) = \sigma(fc_1(AvgPool(I)) + fc_2(MaxPool(I))) \quad (4)$$

$$H_i = \mathbf{CA}(I) \odot R_i, \tag{5}$$

where $\mathbf{CA}()$ denotes the channel attention guidance function, σ represents the sigmoid function, fc_1 and fc_2 denote a fully connected operator, $\odot$ represents the dot-product operator, H is the input of the segmentation head, and i denotes the pyramid levels index.

3.3. Semantic Enhancement Module

We introduce a semantic enhancement module to address the challenges posed by the intricate background of remote sensing images, which often results in blurred edges, missing segmentations, and aliased masks for instances. The SEM in Figure 2(c) optimizes network performance by capturing finer-grained and more detailed semantic information. The proposed module enhances the semantic expression of the underlying features by capturing semantic information at different scales through three parallel atrous convolutions instead of the 1×1 convolution of the underlying features. Moreover, the residual connection can better preserve the local and global information of the original features, improving the network's capability to recognize and localize the object. The overall process can be mathematically expressed as follows:

$$x = \varphi(B_2) \tag{6}$$

$$\mathbf{SEM}(B_2) = \varphi(C_{3\times3\_1}(x) + C_{3\times3\_2}(x) + C_{3\times3\_3}(x)) + B_2, \tag{7}$$

where φ represents 1×1 convolution, $C_{3\times3\_i}$ denotes the 3×3 convolution with dilation factor i, where $i = 1,2,3$.

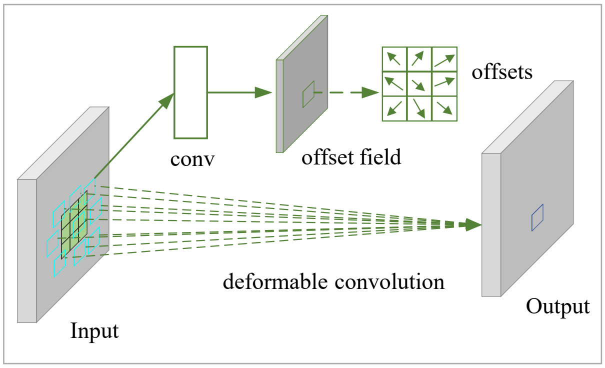
3.4. Deformable Convolution

As we know, the object instances in the remote sensing image often encounter problems such as arbitrary direction, large-scale span, and irregular shape. However, vanilla convolution is less sensitive to object shape changes, and the generalization ability is insufficient to significantly boost the model's capability to segment geometric shape changes of object instances in complex scenarios. Therefore, DCNv2 [46] is merged into the head to learn the geometric deformations of the instances by introducing dynamic offsets in the receptive field, and weights are added to each sampling point to improve the model's ability to model complex deformation changes. Figure 3 illustrates the schematic of deformable convolution.

For a feature at feature map p is $y(p)$, the formula is shown in Equation (8):

Figure 3

Schematic of deformable convolution



$$y(p) = \sum_{k=1}^{K} \omega_k \cdot x(p + p_k + \triangle_{p_k}) \cdot \triangle_{m_k}, \tag{4}$$

where K denotes the number of sampling positions of the convolution kernel; ω_k and p_k denote the weight of the kth position and the pre-set offset, respectively; $\triangle_{p_k}$ and $\triangle_{m_k}$ represent the learnable offset of the kth position and the modulation scalar, respectively; and $x(p + p_k + \triangle_{p_k})$ is the feature value after the offset at p.

3.5. Loss Function

The loss function consists of two parts, the category and the mask prediction losses, and is calculated as shown in Equation (9).

$$L = L_{cate} + \lambda L_{mask}, \tag{9}$$

where the category loss function is Focal Loss. The mask prediction loss function is Dice Loss [32], which aims to solve the problem of positive and negative sample imbalances. λ is set to 3. Equation (10) is the formula for the mask prediction loss function L_{mask} .

$$L_{mask} = \frac{1}{N_{pos}} \sum_k 1_{\{p^*_{i,j}>0\}} d_{mask}(m_k, m^*_k), \tag{10}$$

where $i = [k / S]$, $j = k \bmod S$, N_{pos} is the number of positive samples, p^* and m^* denote the category and mask truth values, respectively, 1 for the indicator function, and d_{mask} represents the Dice Loss, calculated as shown in Equation (11).

$$L_{Dice} = 1 - \frac{2\sum_{x,y}(\mathrm{p}_{x,y}\cdot \mathrm{q}_{x,y})}{\sum_{x,y}\mathrm{p}_{x,y}^{2} + \sum_{x,y}\mathrm{q}_{x,y}^{2}}. \quad (11)$$

4. Experiment

4.1. Experimental Details

Our experimental environment is Ubuntu 20.04, based on MMDetection, an open-source detection toolkit developed by PyTorch, and the experimental hardware platform uses an AMD Ryzen 9 5900HX and an Nvidia GTX3080 GPU. During training, SGD was used as the model optimizer, with the initial learning rate set to 0.0025, momentum to 0.9, and weight decay set to 0.0001.

4.2. Datasets

The iSAID dataset contains 2086 high-resolution images with 15 categories. Since the scale ratio of the original images of this dataset varies a lot, the original images are cropped to 800×800, and the training, validation, and test sets obtained are 18100, 5896, and 19377, respectively.

The NWPU VHR-10 dataset is a 10-category geospatial object detection dataset created by Northwestern Polytechnical University, which consists of 650 positive samples and 150 negative samples containing only background. The dataset was augmented using the data augmentation method RandAugment [5] and randomly divided into training and test sets in a ratio of 7:3.

The HRSID dataset, which contains 5604 SAR images and 16951 instances, is used for ship detection and segmentation. The image size is 800×800 with resolutions of 0.5m, 1m, and 3m, and the training and test sets are 3642 and 1962, respectively.

4.3. Assessment of Indicators

We evaluate the model using COCO metrics, which include average precision (AP), AP_{50}, AP_{75}, AP_S, AP_M, and AP_L. Specifically, AP denotes the AP at different IoU thresholds, AP_{50} and AP_{75} denote the AP values at IoU thresholds of 0.50 and 0.75. The corresponding symbols AP_S, AP_M, and AP_L indicate the average precision for small, medium, and large-size objects.

4.4. Main results

In this section, we analyze the model's performance qualitatively and quantitatively. First, we conduct comparison experiments on three datasets. Then, we conduct enough ablation experiments on the NWPU VHR-10 and HRSID datasets to confirm our model's validity.

4.4.1. Comparative Experiment Results on iSAID

Table 2 compares the results of our method and other popular instance segmentation methods, including top-down methods (Mask R-CNN, Mask Scoring R-CNN, YOLACT), cascade methods (PointRend [13] and Cascade Mask R-CNN), direct methods (RDSNet [30], SOLO, and SOLOv2 [32]), and query-based methods (QueryInst). The comparison results indicate that our method improves SOLOv2 performance by 5.1%, 5.0%, and 6.2% in AP, AP_{50}, and AP_{75}, respectively. Furthermore, our model outperforms other mainstream methods when dealing with complex scenarios, with a clear advantage over different single-stage algorithms. The last two columns assess the model's complexity in terms of parameter count and computation. The comparison results show that our approach achieves higher accuracy despite the absence of a discernible advantage in terms of model complexity. High accuracy means that the model can capture and recognize detailed information in the image more accurately, reducing the possibility of missing segmentation and false alarms, which is crucial for real-world application scenarios that demand high-precision data. Since the high spatial resolution of remote sensing images makes it possible for even subtle mis-segmentation to cause significant real-world ground errors, high accuracy can support more refined applications. In addition, our method does not exhibit model complexity far beyond that of other algorithms but instead achieves high accuracy based on a certain level of complexity that can be applied to real-world remote sensing image applications.

4.4.2. Comparative Experiment Results on NWPU VHR-10 and HRSID

Tables 3 and 4 present the model's quantitative analysis of the NWPU VHR-10 and HRSID datasets. Compared to the baseline SOLOv2 model, our optimized

Table 2
Segmentation performance on iSAID dataset

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params(M)	FLOPs
Mask R-CNN	35.4	59.2	37.7	17.2	41.3	47.4	63.17	351.64
Mask Scoring R-CNN	35.5	58.4	37.6	17.3	41.3	47.2	79.729	337.92
PointRend	33.4	56.3	35.1	16.0	39.4	45.3	60.215	181.248
Cascade Mask R-CNN	34.4	57.5	36.5	16.5	40.3	46.2	96.09	470.0
YOLACT [37]	22.3	43.3	19.9	8.4	31.8	40.4	54.286	91.716
RDSNet [42]	29.3	51.3	29.5	12.6	40.2	47.9	62.04	228.42
QueryInst	28.6	46.8	30.4	12.5	35.5	43.7	196.608	250.88
Luo et al. [23]	29.4	54.5	27.8	15.5	37.8	42.0	43.83	-
SOLO	24.7	44.9	24.2	7.9	32.7	47.8	54.91	442.13
SOLOv2	31.1	54.2	31.4	13.7	38.3	43.1	65.84	284.44
Ours	36.2	59.2	37.6	15.7	43.6	52.4	93.28	379.85

Table 3
Segmentation performance on NWPU VHR-10 dataset

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLACT	41.6	75.2	39.6	23.4	39.7	54.3
QueryInst	38.3	61.2	40.0	19.2	37.0	50.2
SOLO	14.2	20.9	15.0	14.2	17.1	28.6
SOLOv2	32.9	56.4	31.8	13.1	27.3	47.7
Ours	45.8	74.8	44.4	23.3	42.6	55.7

Table 4
Segmentation performance on HRSID dataset

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLACT	33.3	65.4	31.5	33.7	37.3	14.2
SOLO	34.8	61.0	39.1	33.2	50.0	19.6
SOLOv2	33.3	59.3	37.6	32.5	42.0	17.4
Ours	40.9	68.9	48.2	38.6	61.2	29.8

approach achieves a remarkable enhancement in performance, boasting a 12.9% increase in accuracy on the NWPU VHR-10 dataset and a 7.6% improvement on the HRSD dataset. It achieves 4.2% and 6.1% performance gains compared to the sub-optimal algorithm. Moreover, FACENet consistently surpasses SOLOv2 across multiple evaluation metrics, including AP_{50}, AP_{75}, and across varying scales (AP_S, AP_M, and AP_L), unequivocally demonstrating the superiority and efficacy of the algorithm.

4.4.3. Ablation Study

To validate our model's efficacy and robustness, we meticulously evaluated each constituent module on different datasets. As demonstrated in Tables 5 and 6, our network achieves an AP of 45.8% and 40.9% on the two respective datasets, showcasing our approach's remarkable performance and adaptability. More specifically, our FACE has improved AP performance by 4.8% and 5.4%. In addition, with the combined assistance of SEM and FACE-FPN, our network enhances the AP performance by 7.8% and 6.9%, respectively, as they both improve and interact with the information flow of high-level and bottom-level features. Furthermore, dynamic offsets allow for a more extensive coverage area, encompassing the entire of the instance's features and pertinent background information. This approach is pivotal for the network to adeptly handle objects exhibiting intricate geometrical deformations, enhancing its capability to accurately segment and comprehend diverse shapes and configurations within complex scenes. In conclusion, our proposed algorithm effectively boosts the RSI instance segmentation performance, demonstrating its superiority and potential for real-world applications.

The semantic enhancement module fully exploits the semantic information, allowing the network to achieve more delicate segmentation effects. We validate our proposed method by performing an ablation experiment on the SEM. Table 7 clearly illustrates that the single branch and parallel structures fail to capture intricate feature information with a high degree of granularity. Conversely, a reasonable residual structure is crucial in optimizing the network's performance to extract and utilize features more efficiently.

Table 5

Effects on NWPU VHR-10 dataset

Baseline	SEM	FACE-FPN	DCNv2	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√				32.9	56.4	31.8	13.1	27.3	47.7
√	√			34.0	57.7	33.0	14.3	29.7	41.0
√		√		37.7	63.9	37.0	14.8	34.8	44.6
√			√	34.8	60.9	32.6	10.9	31.1	43.4
√	√	√		40.7	67.9	40.3	23.0	36.5	50.6
√	√	√	√	45.8	74.8	44.4	23.3	42.6	55.7

Table 6

Effects on HRSID dataset

Baseline	SEM	FACE-FPN	DCNv2	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√				33.3	59.3	37.6	32.5	42.0	17.4
√	√			38.9	67.0	45.0	37.3	53.6	17.2
√		√		38.7	66.6	44.9	36.9	56.8	16.8
√			√	35.3	61.5	40.8	34.0	49.0	12.7
√	√	√		40.2	68.0	47.8	38.2	59.3	25.0
√	√	√	√	40.9	68.9	48.2	38.6	61.2	29.8

Table 7
Comparison of different semantic components in SEM

Baseline	SEM	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
√	Case 1	32.2	55.1	30.4	9.9	28.1	41.6
√	Case 2	31.3	53.7	28.4	13.3	26.4	41.8
√	Case 3	33.9	56.2	34.9	12.2	29.3	41.8
√	SEM*	32.8	56.4	30.4	14.2	27.5	42.0
√	SEM	34.0	57.7	33.0	12.3	29.7	41.0

Table 8
Comparison of different FPNs

Methods	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
FPN	32.9	56.4	31.8	13.1	27.3	47.7
HRFPN	29.2	51.2	29.6	8.7	24.0	45.2
PAFPN	31.3	53.2	31.0	10.2	26.1	46.2
BiFPN	17.0	23.7	19.7	1.2	9.7	39.5
FACE-FPN	37.7	63.9	37.0	14.8	34.8	44.6

To verify the detection and segmentation performance of FACE-FPN, we choose several advanced FPN structures for comparative analysis. Table 8 demonstrates the poor segmentation of these FPNs with an AP value that is even lower than that of vanilla FPN. It also shows that although these FPNs have exhibited commendable performance on natural images, they have poor transmission capability on remote sensing images. FACE-FPN achieves a promising 4.8% AP increment compared to vanilla FPN, demonstrating the rationality and effectiveness of the structure.

4.4.4. Qualitative Results

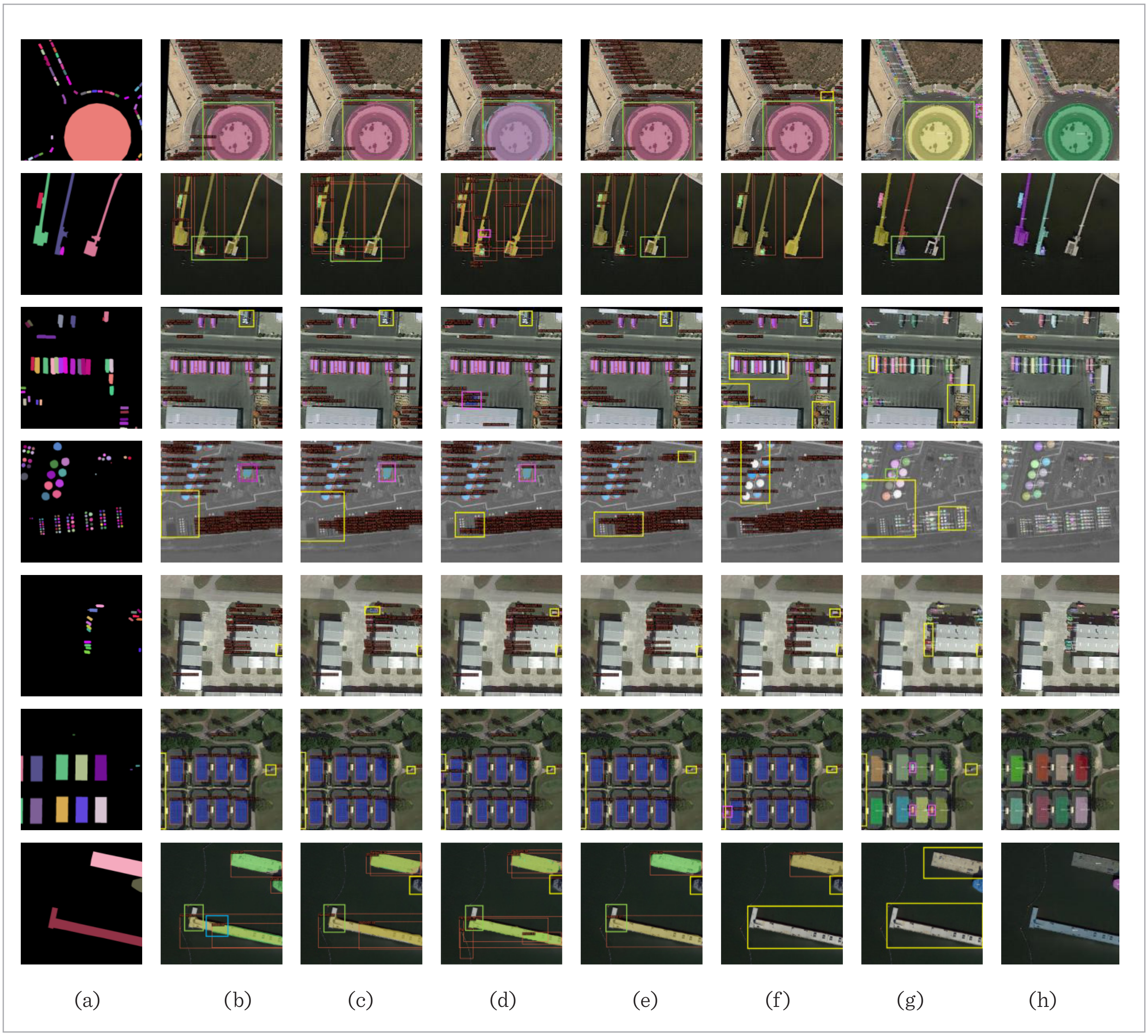
In this section, we perform visual analysis, and the visualization results demonstrate the validity and rationality of our proposed model for RSI instance segmentation. The rectangles marked blue, green, yellow, and purple represent aliasing masks, poor masks, missing segmentation, and false alarms.

iSAID: Figure 4 displays the results of our visualization on the iSAID dataset. Notably, other example segmentation models struggle to deliver satisfactory results when confronted with challenging scenarios featuring small or densely packed segmentation objects, as exemplified in rows 3, 4, 5, and 6. In stark contrast, our model excels in these situations, demonstrating its robustness and capability to segment even the most intricate details accurately. Moreover, conventional models often suffer from inadequate edge segmentation and the generation of aliased masks in scenarios where a single object dominates the scene or a significant scale disparity exists between different classes, as seen in the first and last rows. Conversely, our model excels in these complex conditions, delivering superior segmentation outcomes. In conclusion, the visualization results demonstrate the superiority of the FACENet algorithm in complex remote sensing scenarios, achieving finer-grained segmentation.

NWPU VHR-10: The visual segmentation effects of our proposed FACENet and SOLOv2 are compared in Figure 5. Our observations show that the original

Figure 4

Visualization results on iSAID. (a) GT, (b) Mask R-CNN, (c) Mask Scoring R-CNN, (d) PointRend, (e) Cascade Mask R-CNN, (f) QueryInst, (g) SOLOv2, (h) Ours

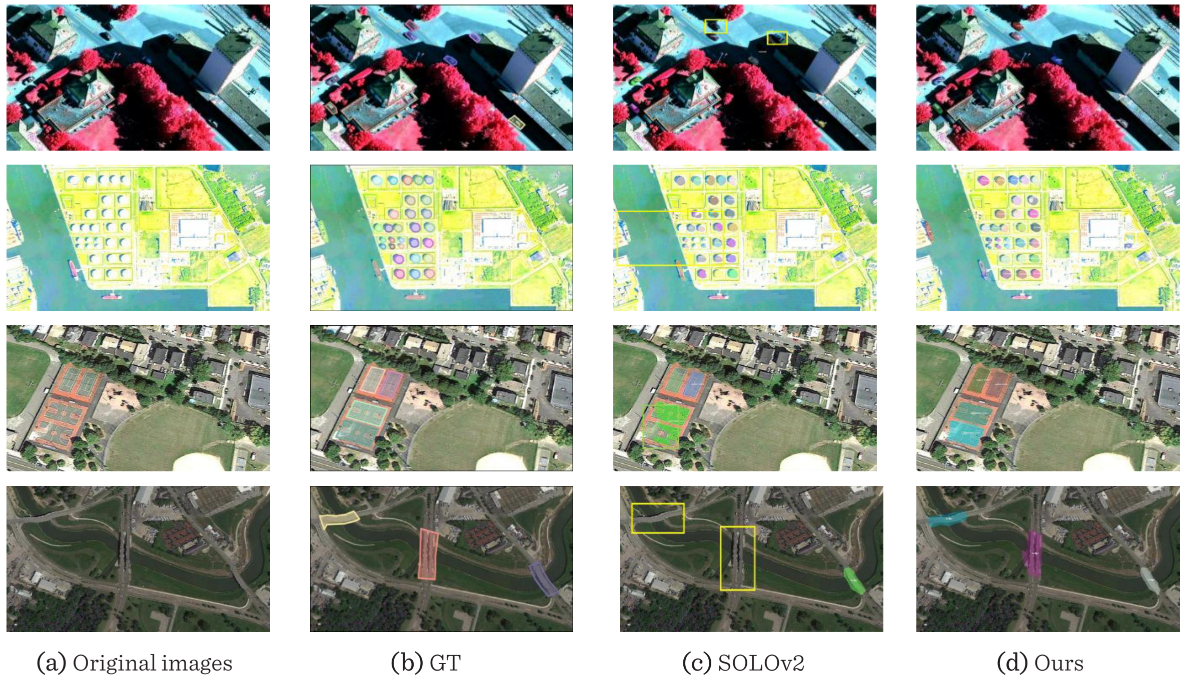


SOLOv2 exhibited challenges in accurately detecting objects within scenes characterized by small and densely packed objects, leading to missed detections. Fortunately, our innovative FACENet framework effectively addresses these limitations in remote sensing instance segmentation, ensuring more comprehensive and accurate segmentation.

HRSID: The outcomes of our visualization on the HRSID dataset are displayed in Figure 6. The second line of results underscores a common challenge encountered when processing coasts with intricate interference: the ship's proximity to the shoreline often blurs the boundary contour, rendering it indistinguishable from the background and resulting in suboptimal segmentation outcomes. Conversely, FACENet stands out by its ability to precisely identify and characterize the ship within the segmentation boundary, even in these complex scenarios. Moreover, our method performs better when the ship boundary is irregular, dense, or the object is tiny.

Figure 5

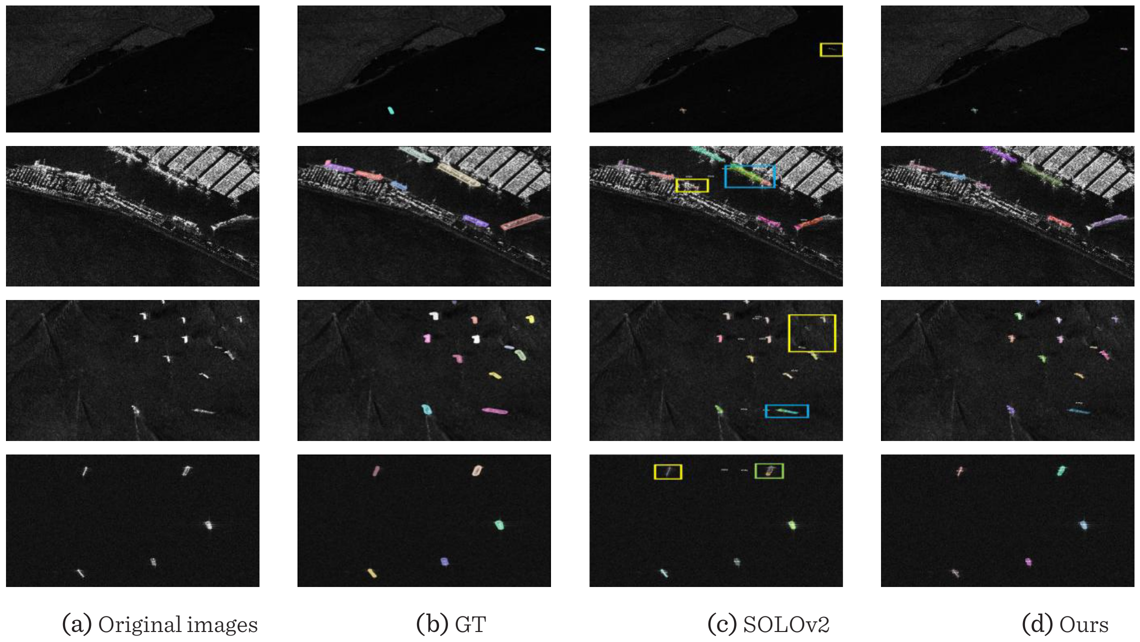
Visualization results on NWPU VHR-10



(a) Original images (b) GT (c) SOLOv2 (d) Ours

Figure 6

Visualization results on HRSID



(a) Original images (b) GT (c) SOLOv2 (d) Ours

5. Conclusion

We propose the FACENet for remote sensing image instance segmentation. Firstly, we replace FPN with FACE-FPN, intending to utilize the channel and detailed information of multilevel features fully. Secondly, SEM enhances the semantic representation of features by capturing finer-grained semantic information. Then, to make the network more sensitive to the geometric deformation of instances in complex scenes, deformable convolution is introduced to improve the network performance. Finally, the experiments demonstrate that our proposed FACENet produces reliable segmentation results in complex scenarios, particularly avoiding the problems of aliasing masks, poorly segmented masks, missing segmentation, and false alarms to some extent, and achieving more delicate segmentation. In the future, we will further optimize the network to enhance small objects' detection and segmentation effects.

Acknowledgement

The authors would like to thank the computing support of the State Key Laboratory of Public Big Data, Guizhou University.

Data Sharing Agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

This work was supported by the Natural Science Foundation of Guizhou Province under Grant No. [2016]1054, the Join Natural Science Foundation of Guizhou Province under Grant No. LH[2017]7226, the Special Project of New Academic Talent Training and Innovation Exploration of Guizhou University under Grant No. [2017]5788, and National Natural Science Foundation of China under Grant No. 12205062.

Conflicts of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Bolya, D., Zhou, C., Xiao, F., Lee, Y. J. YOLACT: Real-Time Instance Segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 9156-9165. https://doi.org/10.1109/ICCV.2019.00925
2. Cai, Z., Vasconcelos, N. Cascade R-CNN: High Quality O-bject Detection and Instance Segmentation. IEEE Transact-ions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483-1498. https://doi.org/10.1109/TPAMI.2019.2956516
3. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, Rohit. Masked-attention Mask Transformer for Universal Image Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 1290-1299. https://doi.org/10.1109/CVPR52688.2022.00135
4. Cheng, G., Han, J., Zhou, P., Guo, L. Multi-class Geospatial Object Detection and Geographic Image Classification Based on Collection of Part Detectors. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 98, 119-132. https://doi.org/10.1016/j.isprsjprs.2014.10.002
5. Cubuk, E. D., Zoph, B., Shlens, J., Le, Q. V. Randaugment: Practical Automated Data Augmentation with A Reduced Search Space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. 3008-3017. https://doi.org/10.1109/CVPRW50498.2020.00359
6. Fang, L., Jiang, Y., Yan, Y., Yue, J., Deng, Y. Hyperspectral Image Instance Segmentation Using Spectral-Spatial Feature Pyramid Network. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-13. https://doi.org/10.1109/TGRS.2023.3240481
7. Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W. Instances as Queries. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 6890-6899. https://doi.org/10.1109/ICCV48922.2021.00683
8. Fu, D., Zeng, X., Han, S., Lin, H., Li, W. Nested Densely Atrous Spatial Pyramid Pooling and Deep Dense Short Connection for Skeleton Detection. IEEE Transactions on Human-Machine Systems, 2023, 53(1), 75-84. https://doi.org/10.1109/THMS.2022.3224552

9. Geng, W., Cao, Z., Guan, P., Ren, G., Yu, J., Jing, F. Adaptive Long-Neck Network with Atrous-Residual Structure for Instance Segmentation. IEEE Sensors Journal, 2023, 23(7), 7786-7797. https://doi.org/10.1109/JSEN.2023.3244818

10. Hafiz, A. M., Bhat, G. M. A Survey on Instance Segmentation: State of the art. International Journal of Multimedia Information Retrieval, 2020, 9, 171-189. https://doi.org/10.1007/s13735-020-00195-x

11. He, K. M., Gkioxari, G., Dollár, P., Girshick, R. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.322

12. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X. Mask Scoring R-CNN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 6402-6411. https://doi.org/10.1109/CVPR.2019.00657

13. Kirillov, A., Wu, Y., He, K., Girshick, R. PointRend: Ima-ge Segmentation as Rendering. Proceedings of the IEEE/C-VF Conference on Computer Vision and Pattern Recogniti-on, 2020, 9796-9805. https://doi.org/10.1109/CVPR42600.2020.00982

14. Li, J. Y., Cai, Y. X., Li, Q., Kou, M. Y., Zhang, T. X. A R-eview of Remote Sensing Image Segmentation by Deep L-earning Methods. International Journal of Digital Earth, 2024, 17(1), 2328827. https://doi.org/10.1080/17538947.2024.2328827

15. Li, K., Wan, G., Cheng, G., Meng, J. W., Han, J. W. Obje-ct Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159, 296-307. https://doi.org/10.1016/j.isprsjprs.2019.11.023

16. Li, Y., Shi, T., Zhang, Y., Ma, J. SPGAN-DA: Semantic-Preserved Generative Adversarial Network for Domain Ad-aptive Remote Sensing Image Semantic Segmentation. IEE-E Transactions on Geoscience and Remote Sensing, 2023, 61, 1-17. https://doi.org/10.1109/TGRS.2023.3313883

17. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12), 6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827

18. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106

19. Liu, N., Li, W., Sun, X., Tao, R. Chanussot, J. Remote Sensing Image Fusion with Task-Inspired Multi-scale Nonlocal-Attention Network. IEEE Geoscience and Remote Sensing Letters, 2023, 20, 1-5. https://doi.org/10.1109/LGRS.2023.3254049

20. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. Path Aggregation Ne-twork for Instance Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn-ition, 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

21. Liu, Y., Li, H. F., Hu, C., Luo, S., Luo, Y., Chen, C. W. Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images. IEEE Transactions on Neural Networks and Learning Systems, 2024, 1-15. https://doi.org/10.1109/TNNLS.2023.3336563

22. Luo, Y., Cao, X., Zhang, J., Cao, X., Guo, J., Shen, H., Wang, T., Feng, Q. CE-FPN: Enhancing Channel Information for Object Detection. Multimedia Tools and Applications, 2022, 81(21), 30685-30704. https://doi.org/10.1007/s11042-022-11940-1

23. Luo, Y., Han, J., Liu, Z., Wang, M., X, G. S. An Elliptic Centerness for Object Instance Segmentation in Aerial Images. Journal of Remote Sensing, 2022, 2022, 1-14. https://doi.org/10.34133/2022/9809505

24. Ma, W., Li, Y., Zhu, H., Ma, H., Jiao, L., Shen, J., Hou, B. A Multi-Scale Progressive Collaborative Attention Network for Remote Sensing Fusion Classification. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8), 3897-3911. https://doi.org/10.1109/TNNLS.2021.3121490

25. Ren, S. Q., He, K. M., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

26. Su, Y., Gao, L., Jiang, M., Plaza, A., Sun, X., Zhang. B. NSCKL: Normalized Spectral Clustering with Kernel-Based Learning for Semisupervised Hyperspectral Image Classification. IEEE Transactions on Cybernetics, 2022, 53(10), 6649-6662. https://doi.org/10.1109/TCYB.2022.3219855

27. Tan, M., Pang, R., Le, Q. V. EfficientDet: Scalable and Efficient Object Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 10778-10787. https://doi.org/10.1109/CVPR42600.2020.01079

28. Teng, Lin., Qiao, Y. L., Shafiq, M., Srivastava, G., Javed, A. R., Gadekallu, T. R., Yin, S. L. FLPK-BiSeNet: Federated Learning Based on Priori Knowledge and Bilater-

al Segmentation Network for Image Edge Extraction. IEEE Transactions on Network and Service Management, 2023, 20(20), 1529-1542. https://doi.org/10.1109/TNSM.2023.3273991

29. Tian, Z., Shen, C. H., Chen, H. Conditional Convolutions for Instance Segmentation. Proceedings of the European Conference on Computer Vision, 2020, 282-298. https://doi.org/10.1007/978-3-030-58452-8_17

30. Wang, S., Gong, Y., Xing, J., Huang, L., Huang, C., Hu, W. RDSNet: A New Deep Architecture for Reciprocal Object Detection and Instance Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07), 12208-12215. https://doi.org/10.1609/aaai.v34i07.6902

31. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L. SOLO: Segmenting Objects by Locations. Proceedings of the European Conference on Computer Vision, 2020, 649-665. https://doi.org/10.1007/978-3-030-58523-5_38

32. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. Advances in Neural Information Processing Systems, 2020, 33, 17721-17732.

33. Wei, S., Zeng, X. F., Zhang, H., Zhou, Z. C., Shi, J., Zhang, X. L. LFG-Net: Low-Level Feature Guided Network for Precise Ship Instance Segmentation in SAR Images. IEEE Transactions on Geoscience and Rem-ote Sensing, 2022, 60, 1-17. https://doi.org/10.1109/TGRS.2022.3188677

34. 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. https://doi.org/10.1109/ACCESS.2020.3005861

35. Xu, G., Liao, W., Zhang, X., Li, C., He, X., Wu, X. Haar Wavelet. Downsampling: A simple But Effective Downsampling Module for Semantic Segmentation. Pattern Recognition, 2023, 143, 109819. https://doi.org/10.1016/j.patcog.2023.109819

36. Yin, S. L. Object Detection Based on Deep Learning: A Brief Review. IJLAI Transactionss on Science and Engineering, 2023, 1(02), 1-6.

37. Ye, W., Zhang, W., Lei, W., Zhang, W., Chen, X., Wang, Y. Remote Sensing Image Instance Segmentation Network with Transformer and Multi-scale Feature Representation. Expert Systems with Applications, 2023, 234, 121007. https://doi.org/10.1016/j.eswa.2023.121007

38. Yin, S. L., Wang, L. G., Shafiq, M., Teng, Lin., Laghari, A. A., Khan, M. F. G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-Weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16, 3583-3598. https://doi.org/10.1109/JSTARS.2023.3241405

39. Yin, S. L., Wang, L. G., Wang, Q. M., Ivanović, M., Yang, J. H. M2F2-RCNN: Multi-functional Faster RCNN Based on Multi-scale Feature Fusion for Region Search in Remote Sensing Images. Computer Science and Information Systems, 2023, 20, 1289-1310. https://doi.org/10.2298/CSIS230315054Y

40. Yu, F., Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. arXiv preprint, 2015, arXi-v:1511.07122. https://doi.org/10.48550/arXiv.1511.07122

41. Zamir, S. W., Arora, A., Gupta, A., Khan, S., Sun, G., Khan, F. S., Zhu, F., Shao, L., Xia, G., Bai, X. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, 28-37.

42. Zeng, X., Wei, S., Shi, J., Zhang, X. A Lightweight Adaptive RoI Extraction Network for Precise Aerial Image Instance Segmentation. IEEE Transactions on Instrumentation and Measurement, 2021, 70, 1-17. https://doi.org/10.1109/TIM.2021.3121485

43. Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X. RefineMask: Towards High-Quality Instance Segmentation with Fine-Grained Features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 6857-6865. https://doi.org/10.1109/CVPR46437.2021.00679

44. Zhang, S., Cao, Y., Sui, B. DF-Mask R-CNN: Direction Field-Based Optimized Instance Segmentation Network for Building Instance Extraction. IEEE Geoscience and Remote Sensing Letters, 2023, 20, 1-5. https://doi.org/10.1109/LGRS.2023.3297839

45. Zhang, M. W., Jing, W. P., Lin, J. B., F, N. Z., Wei, W., Woźniak, M., Damaševičius, R. NAS-HRIS: Automatic Design and Architecture Search of Neural Network for Semantic Segmentation in Remote Sensing Images. Sensors, 2020, 20(18), 5292. https://doi.org/10.3390/s20185292

46. Zhu, X., Hu, H., Lin, S., Dai, J. Deformable ConvNets v2: More Deformable, Better Results. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 9300-9308. https://doi.org/10.1109/CVPR.2019.00953