Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.027764

ARTICLE

PanopticUAV: Panoptic Segmentation of UAV Images for Marine Environment Monitoring

Yuling Dou[1], Fengqin Yao[1], Xiandong Wang[1], Liang Qu[2], Long Chen[3], Zhiwei Xu[4], Laihui Ding[4], Leon Bevan Bullock[1], Guoqiang Zhong[1] and Shengke Wang[1,*]

[1]School of Computer Science and Technology, Ocean University of China, Qingdao, 266100, China

[2]North China Sea Environmental Monitoring Center, State Oceanic Administration, Qingdao, 266000, China

[3]Department of Informatics, University of Leicester, Leicester, LE1 7RH, UK

[4]Research and Development Department, Shandong Willand Intelligent Technology Co., Ltd., Qingdao, 266102, China

*Corresponding Author: Shengke Wang. Email: neverme@ouc.edu.cn

Received: 14 November 2022 Accepted: 19 April 2023 Published: 22 September 2023

ABSTRACT

UAV marine monitoring plays an essential role in marine environmental protection because of its flexibility and convenience, low cost and convenient maintenance. In marine environmental monitoring, the similarity between objects such as oil spill and sea surface, *Spartina alterniflora* and algae is high, and the effect of the general segmentation algorithm is poor, which brings new challenges to the segmentation of UAV marine images. Panoramic segmentation can do object detection and semantic segmentation at the same time, which can well solve the polymorphism problem of objects in UAV ocean images. Currently, there are few studies on UAV marine image recognition with panoptic segmentation. In addition, there are no publicly available panoptic segmentation datasets for UAV images. In this work, we collect and annotate UAV images to form a panoptic segmentation UAV dataset named UAV-OUC-SEG and propose a panoptic segmentation method named PanopticUAV. First, to deal with the large intraclass variability in scale, deformable convolution and CBAM attention mechanism are employed in the backbone to obtain more accurate features. Second, due to the complexity and diversity of marine images, boundary masks by the Laplacian operator equation from the ground truth are merged into feature maps to improve boundary segmentation precision. Experiments demonstrate the advantages of PanopticUAV beyond the most other advanced approaches on the UAV-OUC-SEG dataset.

KEYWORDS

Panoptic segmentation; UAV marine monitoring; attention mechanism; boundary mask enhancement

1 Introduction

Deep learning has caused a stir in various fields [1–4]. In recent years, deep learning has been gradually applied in Marine environmental monitoring. Panoptic segmentation based on deep learning [5] was first proposed by Facebook Research Institute and Heidelberg University in Germany. Combined with UAV technology, it makes up for the deficiency of remote sensing images, and can accurately identify the distribution range of Marine objects by using different flight altitudes and

flight angles. As a result, the demand for intelligent data analysis collected by drones is increasing. As shown in Fig. 1, UAV images with polymorphic objects captured by UAV pose a major challenge to UAV Marine monitoring. The traditional target detection algorithm can not identify the target accurately.

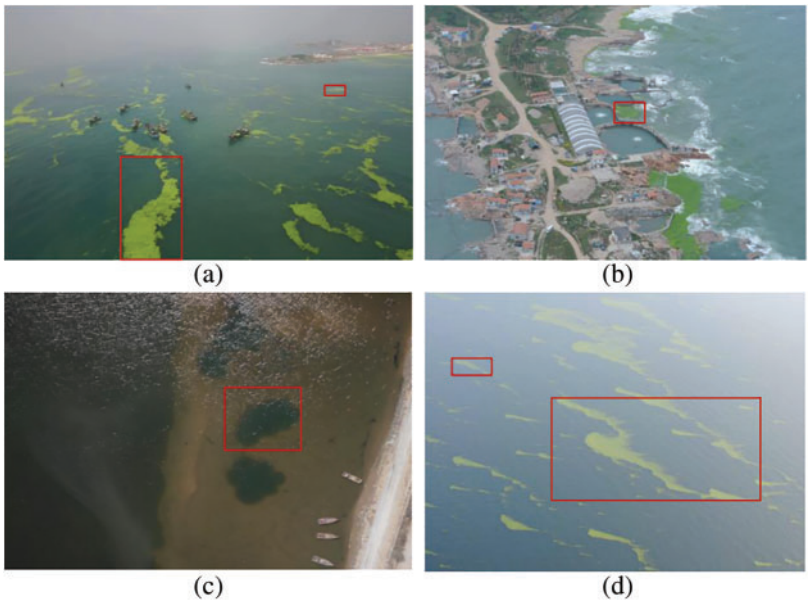

Figure 1: UAV images with polymorphic objects captured by drones pose a significant challenge to UAV object detection, while traditional detection algorithms cannot accurately identify objects. Panoptic segmentation is a vision task of comprehensive image parsing, which can identify different shapes of objects very well. Hence, we solve the problem of UAV marine monitoring with panoptic segmentation

Panoramic segmentation is a full-pixel visual analysis task, which divides the input image into different areas [6,7] according to certain standards, and it unifies instance segmentation [8,9] and semantic segmentation [10,11]. This task focuses on the things and stuff classes that split objects. Instances objects, such as foreground objects, cars, pedestrians, animals, and tools, can all be counted. stuff refers to an uncountable background, such as grass, sky and roads. The existing methods are mainly targeted at realistic scenarios, such as automatic driving datasets [12,13], and have achieved outstanding results. Panoramic segmentation can well identify shapes of different objects [14]. Therefore, the use of panoptic segmentation algorithm for image analysis may have a more comprehensive application and greater significance in the UAV scenes. However, most of the existing panoramic segmentation methods are for real-world scenes, such as unmanned technology. These images are very different from ocean images from drones [15]. UAV visual data [16] has polymorphic objects and large scale, and many small objects may be missing in panoramic segmentation. Therefore, we need to design stronger feature representation networks to improve the segmentation accuracy of UAV Marine images.

Because of the above problems, we propose a new module called PanopticUAV. First, we select ResNet50 [17] to extract features. The polymorphic objects and large intraclass variability require scale and receptive field adaptation. Due to the fixed size of standard CNN extraction features, we employ deformable convolution [18] in ResNet50 to obtain a feature map with a flexible receptive

field. Second, we use FPN [19] to fuse features. The objects are complex and diverse, which leads to inaccurate boundary information. We generate boundary masks from the ground truth and merge them into P5 feature maps to improve the boundary segmentation precision. Then, we combine context information to fuse features in UAV scenarios, where targets are small and features are few. Specifically, we use the CBAM [20] attention mechanism to further feature fusion and reduce false detection cases. Finally, we gain a high-resolution feature from FPN [19] and then generate encoded features by convolution. Additionally, the kernel generator produces each kernel weight of an object instance or stuff category. The kernel weight is multiplied by the encoded feature to obtain segmentation results for each thing and stuff. Our method performs well on the collected UAV datasets, UAV-OUC-SEG. Extensive experiments verify the effectiveness of PanopticUAV for UAV sea images. In short, the main contributions of this article are listed as follows:

1. We collect and annotate UAV sea images, forming a dataset, UAV-OUC-SEG.

2. We propose a panoptic segmentation method for UAV marine image recognition, named PanopticUAV. To some extent, it addresses the problem that traditional object detection algorithms cannot accurately identify objects with polymorphic objects.

3. Boundary mask enhancement and the CBAM attention mechanism are integrated in Panoptic-UAV for better application to UAV marine image recognition.

2 Related Work

This section presents some of the work related to panoptic segmentation and UAV image parsing. In Section 2.1, we introduce the role of panoptic segmentation, existing panoptic segmentation models and their advantages and disadvantages. In Section 2.2, we introduce and analyze the research work related to UAV image parsing.

2.1 Panoptic Segmentation

Panoptic segmentation is a global, uniform, and pixel-level approach that requires assigning semantic labels to each pixel and distinguishing different individual IDs for the same semantics [14], whose evaluation metrics are proposed by [5]. Due to the complexity and uniformity, panoptic segmentation appeals to many researchers for excellent work.

From deep learning [21,22] framework design, existing methods can be split into three forms: top-down, bottom-top, and united. Most advanced approaches address the panoptic segmentation problem from a top-down perspective. Specifically, PanopticFPN [23] uses the mask R-CNNs [8] to extract overlapping instances, and then a branch of semantic segmentation is added. Then, some postprocessing methods are used to resolve the overlap of the mask, and the results of the semantic and instance segmentation are fused. The results overlap the semantic and instance segmentation results and require postprocessing. Hence, the two-stage method is generally slower. UPSNet [7] achieves a panoptic head at the end of the PanopticFPN [23] to better integrate the stuff and things branches. In contrast, the bottom-top methods add an instance segmentation branch in the semantic segmentation method. DeeperLab [6] put forward a single-shot, bottom-up approach for complete image analysis. A fully convolutional approach extracts features for instance segmentation and semantic segmentation. There is a fusing process that fuses semantic and instance information. As the instance information is category-irrelevant, the category corresponding to the mask is selected by voting. PanopticDeepLab [24] uses an encoder backbone network with semantic segmentation and instance segmentation and adds a null convolution to the last module of the backbone network to obtain a denser feature map. A dual ASPP dual decoder structure is used for feature fusion, and finally, the semantic segmentation and

instance segmentation results are fused to produce the final segmentation results. These methods are better in terms of speed while maintaining accuracy. PanopticFCN [25], a method for handling things and stuff simultaneously, enables accurate end-to-end panoptic segmentation. DETR [26] applies the transformer to a computer vision task and achieves good panoptic segmentation results. MaskFormer [27] regards panoptic segmentation as a mask classification task that adopts a transformer decoder module to achieve mask prediction.

These methods aim for image parsing in natural scenes, such as COCO, driverless scenes, and Cityscapes. UAV sea images with morphological diver-sity, large scale, and small targets present new challenges for panoptic segmentation. New networks need to be designed to solve the problems that exist in UAV scenes.

2.2 UAV Image Parsing

Image parsing decomposes images into successive visual patterns, such as textures and detection targets, covering segmentation, detection, and recognition tasks. Image parsing is first performed using a Bayesian-based [28] framework. Introducing instance-based panoptic quality (PQ) evaluation into several benchmarks, panoptic segmentation is a task of comprehensive understanding of the image, which is also of great importance for practical applications.

Drones play an essential role in marine environmental monitoring because of their flexibility and convenience. How to efficiently and accurately understand UAV sea imagers is a pressing issue. Existing image parsing for UAV images is mainly focused on object detection [29], crowd counting [30], object tracking [31], etc. In the popular drone dataset, VisDrone [32], images are acquired in different cities and are mainly gathered for image detection and target tracking tasks in computer vision. Some semantic segmentation datasets, such as FloodNet [33] and UAVid [34] are primarily concerned with specific scenarios. Therefore, there is no UAV sea dataset for panoptic segmentation. We collect and label UAV images, forming a dataset, UAV-OUC-SEG, which will be published shortly.

3 PanopticUAV

We provide a detailed description of the PanopticUAV approach. First, we summarize the structure of the panoramic UAV, then describe the composition of each part of the backbone network, boundary mask enhancement, attention mechanism, and prediction head. Finally, we calculate the loss function.

3.1 Overview

Fig. 2 shows the structure of PanopticUAV. First, we use ResNet50 as the backbone to extract features. As the standard CNN extraction features have a fixed size, we employ deformable convolution in ResNet50 to obtain a flexible receptive field. Second, we use FPN to fuse features P2, P3, P4, and P5. We then generate the P5 boundary mask by using the Laplacian operator from the ground truth, and merge it with the P5 feature maps to improve the boundary segmentation precision. After that, we use the CBAM attention mechanism to combine context information in P2, P3, P4, and P5 features. Finally, we obtain high-resolution feature maps from AP2, AP3, AP4, and AP5, and generate encoded features by convolution. In addition, the kernel generator produces a kernel weight for each object instance or stuff category. We multiply the kernel weight by the encoded feature to obtain segmentation results for each thing and stuff.

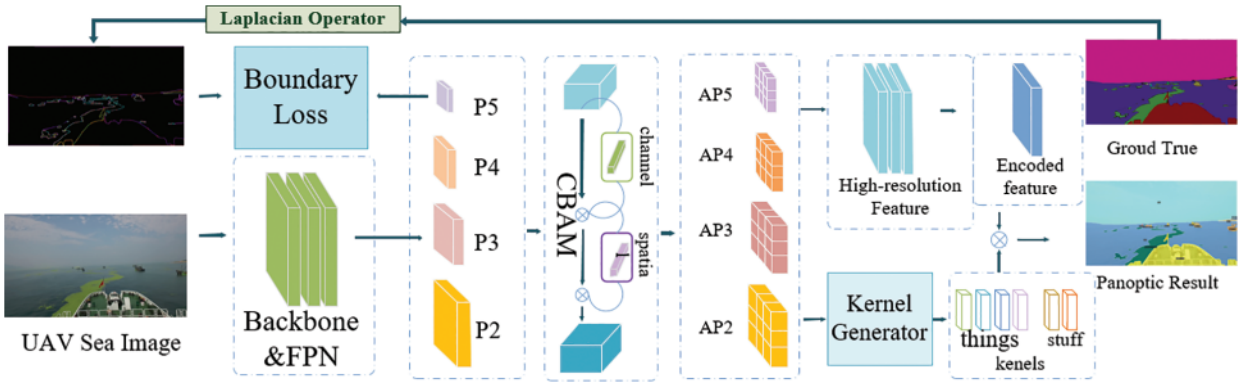

Figure 2: Our PanopticUAV adopts deformable ResNet50 as the backbone to extract features and FPN to fuse features, P2, P3, P4, and P5. Then, the P5 boundary mask is generated by the Laplacian operator from the ground truth and merged into the P5 feature map to improve the boundary segmentation precision. Then, the CBAM attention mechanism combines context information to obtain AP2, AP3, AP4, and AP5. Finally, we gain a high-resolution feature from the fused feature map. Then we generate encoded features by convolution. Additionally, the kernel generator produces each kernel weight of an object instance or stuff category. The kernel weight is multiplied by the encoded feature to obtain segmentation results for each thing and stuff

3.2 Network Pipeline

We introduce the network pipeline for PanopticUAV, including the details of the implementation of each part and the corresponding loss functions.

3.2.1 Backbone

We use ResNet50 as the common backbone, but the UAV sea images present rich morphological diversity and many small objects. Traditional CNNs use a fixed-size kernel to extract features, which limits the shape of the receptive field. Therefore, we use deformable convolution in ResNet50, which adds an extra direction parameter to each element in the convolution kernel. This allows the kernel to be extended to cover a wider range during training.

Deformable convolution involves performing both deformable convolutional and pooling operations in 2-dimensions on the same channel. The regular convolution operation can be divided into two main parts: (1) sampling the input feature map using a regular grid R and (2) performing a weighting operation where R determines the size of the receptive field.

$$R = \{(-1,1),(-1,0),\ldots(0,1),(1,1)\} \tag{1}$$

For position P_0 on the output feature map, the calculation is performed by the following equation:

$$y(P_0) = \sum_{P_n \in R} W(p_n) \cdot X(P_0 + P_n) \tag{2}$$

In this equation, R refers to 3 × 3 regions, X refers to the input, W refers to the corresponding weight, and Pn refers to each offset within the range of the P0 convolution kernel. Deformable convolution is different from conventional convolution in that it introduces an offset ΔPn for each

point based on conventional convolution. The offset is generated by the input feature map and another convolution, and it changes the size of the receptive field. This is shown in the following equation:

$$\mathrm{y}\left(\mathrm{P}_0\right)=\sum_{P n \in R} W\left(\mathrm{p}_{\mathrm{n}}\right) \cdot X\left(P_0+P_n+\Delta P_n\right) \tag{3}$$

With deformable convolution, the sampled position becomes an irregular shape due to the introduction of the offset ΔPn, which is often represented as a fractional number. We achieve this through bilinear interpolation, which is given by the following equation:

$$\mathrm{y}\left(P_0\right)=\sum_q G(q, p) \cdot X(q) \tag{4}$$

Here, p represents an arbitrary decimal position. Q traverses all the integer space positions of the characteristic graph x. G is two-dimensional. It can be separated into two one-dimensional convolution cores.

$$p=p_0+p_n+\Delta p_n \tag{5}$$

$$\mathrm{G}(\mathrm{q}, p)=g\left(q_{\mathrm{x}}, p_x\right) \cdot g\left(q_y, p_y\right) \tag{6}$$

$$g(q, p)=\max (0,1-|q-p|) \tag{7}$$

3.2.2 Boundary Mask Enhancement

We obtain feature maps P2, P3, P4, and P5 from the ResNet50 backbone and FPN. However, due to the complexity and diversity of the images, as well as the inaccuracy of the boundary, we obtain the boundary mask from the ground-truth using the Laplacian operator Eq. (5). We then merge the boundary information with a feature map by calculating boundary losses. Specifically, P2, P3, P4, and P5 all calculate boundary losses with the boundary mask. Our experiments have shown that fusing the boundary information with the P5 feature maps yields better results.

We were inspired by BisenetV2 [10] to use a Laplacian operator kernel L to generate a boundary mask. We selected three different strides (2×, 4×, and 8×) to obtain mask maps of different scales, which were then upsampled to a uniform size and fused using a 1 × 1 convolution. We obtained P5 through a 3 × 3 convolution, batch norm, ReLU, and another 3 × 3 convolution, and adjusted its shape to match that of the boundary mask through bilinear interpolation. Finally, we used binary cross-entropy and dice loss to jointly optimize the boundary mask by learning about P5 and the boundary mask.

$$L=\begin{pmatrix}-1 & -1 & -1\\ -1 & 8 & -1\\ -1 & -1 & -1\end{pmatrix}, \tag{8}$$

$$L_{\text{boundary}}\left(p_d, g_d\right)=L_{dice}\left(p_d, g_d\right)+L_{bce}\left(p_d, g_d\right), \tag{9}$$

where p_{d} denotes the predicted p5 feature, and g_{d} denotes the corresponding boundary mask, L_{dice} denotes the Dice loss, and L_{bce} denotes the binary cross-entropy loss.

During the training phase, we use the boundary mask enhancement module to obtain better weights. However, we remove this module during the inference phase.

3.2.3 Attention Module

UAV sea images contain objects with rich morphological diversity, and some targets are not always accurately segmented. To reduce the error rate, it is important to incorporate context information. The attention module can help to make objects more distinctive, which is why we use the CBAM network to fuse features further. The CBAM attention mechanism consists of a spatial attention module and a channel attention module, as shown in Fig. 2. The channel attention mechanism (green) first obtains weighted results from the output features, which are then fed into the spatial attention mechanism (purple) to give a final weighting to the result. By applying CBAM, the new feature map can be weighted in both channel and space, which greatly improves the interconnectivity of each feature in the channel and space, and is more conducive to extracting effective features of the target.

In the channel attention module, input features F_0 are subjected to global max pooling and global average pooling based on width and height. Then, the output features are fed to the MLP, which is added and activated by sigmoid. The channel attention features are multiplied by the input features F_0 elementwise to generate the required features F_c by the spatial attention module. The equation:

$$\mathrm{M_c}(\mathrm{F}) = \mathrm{F_0} \times (\sigma\,(\mathrm{MLP}\,(\mathrm{AvgPool}\,(\mathrm{F_0})) + \mathrm{MLP}\,(\mathrm{MaxPool}\,(\mathrm{F_0})))) \tag{10}$$

The spatial attention module takes the features F_c as the input features. First, a global max pooling and global average pooling based on the channel are performed, and then the two results are contacted based on the channel. The sigmoid generates the spatial attention features, and then the features are multiplied by the input features F_c to obtain the final features F_8.

$$\mathrm{M_s}(\mathrm{F}) = \mathrm{F_c} \times \left(\sigma\left(\mathrm{f}^{7\times7}\left([\mathrm{AvgPool}\,(\mathrm{F_c})\,;\mathrm{MaxPool}\,(\mathrm{F_c})]\right)\right)\right) \tag{11}$$

Following the CBAM module, we obtain different feature maps, AP2, AP3, AP4, and AP5.

3.2.4 Prediction Head

We integrate the high-resolution features from AP2, AP3, AP4, and AP5 to generate encoded features through a convolution layer. Then, we use a kernel generator to produce kernel weights for each level feature APi for both things and stuff. To identify each object class and position from the features APi, we use the center point to position each individual thing. However, stuff is indistinguishable from individual things. Therefore, we represent the location of each instance separately by using the entire area of stuff and the object center of things. Finally, we use a focal loss to calculate the position loss,

$$L_{\mathrm{pos}}^{\mathrm{th}} = \sum_{\mathrm{i}} FL\left(L_{\mathrm{i}}^{\mathrm{th}}, Y_i^{th}\right)/N_{th} \tag{12}$$

$$L_{pos}^{\mathrm{st}} = \sum_{i} FL\left(L_i^{st}, Y_i^{st}\right)/W_iH_i \tag{13}$$

$$L_{pos} = L_{pos}^{th} + L_{pos}^{st} \tag{14}$$

where FL $(\cdot, \cdot)$ represents the focal Loss. N is the number of things, and W_iH_i is the stuff region.

Then, we fuse different kernel weights for the same objects. Finally, every kernel weight is multiplied by the encoded feature gaining the mask of each instance object. The predicted masks calculate the loss with the ground truth. We adopt dice loss,

$$L_{\mathrm{seg}} = \sum_{\mathrm{j}} WDice\left(P_j, Y_j^{seg}\right)/(M+N) \tag{15}$$

where Y_j^{seg} refers to the ground truth for the j_{th} predicted mask P_j. The N denotes the number of objects.

3.2.5 Loss Function

We need to calculate the position loss, boundary mask loss, and segmentation loss in the training stage. Therefore, the total loss function is as follows:

$$L_{\text{total}} = \lambda_{\text{pos}} L_{\text{pos}} + \lambda_{\text{seg}} L_{\text{seg}} + \lambda_{\text{boundary}} L_{\text{boundary}} \tag{16}$$

where λ_{pos}, λ_{seg} and λ_{seg} are set to constants 1, 2 and 1, respectively, in our work.

4 Experiments

We first present the dataset UAV-OUC-SEG dataset and experimental settings for PnopticUAV. Then, we report the ablation study on UAV-OUC-SEG to reveal the effect of each component. Finally, comparisons with previous methods on UAVSEG are presented.

4.1 Dataset

At present, there are no existing UAV datasets available for panoptic segmentation of ocean scenarios. To address this gap, we have collected and produced a UAV sea image dataset called UAV-OUC-SEG, which contains a total of 813 images. Of these, 609 images are in the training set, 103 are in the validation set, and 101 are in the testing set. The dataset includes a total of 18 categories, consisting of 7 classes of things (person, bicycle, car, hydrant, trashbin, wheelchair, and boat) and 11 classes of stuff (sakura, road, sidewalk, vegetation, building, enteromorpha, sea, sky, land, oil, and seaweed). The dataset covers various scenarios, including campus, enteromorpha, oil spill, and sea grass, which are particularly important for marine environmental monitoring.

4.2 Experimental Setting

Our training and evaluation were implemented in GPU 3090 X 4. The software versions include python 3.7, PyTorch 1.8.0, CUDN 11.1, and numpy 1.21.2. Resnet50 with the deformable convolution strategy was used as our backbone, performing pretraining weight initialization in ImageNet. FPN was used to fuse the features. We used an SGD optimizer for training. The initial learning rate was set to 0.001. It employed the polynomial learning rate policy where the current learning rate equals the initial one multiplied by $\left(1 - \left(\frac{\text{iter}}{max - iter}\right)^{\text{power}}\right)$. The power was 0.9, and the max-iter was 90000. Due to GPU memory limitations, the batch size was 4, momentum was 0.9, and weight decay was 0.0001 to train the model. We used the panoptic segmentation assessment metric PQ to evaluate the model.

4.3 Ablation Study

We conducted several experiments to evaluate the effectiveness of our model. First, we used the baseline method, PanopticFCN, to test it on the UAV-OUC-SEG dataset. We then added the deformable convolution module, the CBAM module, and the boundary mask enhancement to demonstrate the effectiveness of these approaches. The results of our experiments are shown in Table 1. The PanopticFCN+D denotes the addition of deformable convolution, which improved PQ and PQst by 1.10% and 2.11%, respectively. However, PQth decreased by 1.32%, possibly due to the small foreground objects in the drone scene. The PanopticFCN+B denotes the boundary mask enhancement strategy, which resulted in an additional 3.15%, 0.25%, and 2.71% improvement in PQ, PQth and PQst, respectively. These results show that the boundary mask enhancement strategy is effective.

Table 1: Panoptic quality (PQ), segmentation quality (SQ), and recognition quality (RQ). Ablation study: PanopticFCN is the baseline, PanopticFCN+D denotes adding deformable convolution, PanopticFCN+B denotes adding boundary mask enhancement strategy, and PanopticFCN+A denotes adding the CBAM attention mechanism

Method	PQ	SQ	RQ	PQ_{th}	SQ_{th}	RQ_{th}	PQ_{st}	SQ_{st}	RQ_{st}	AP	mIOU	fwIOU	mACC	pACC
PanopticFCN	49.16	77.27	62.66	46.53	71.64	63.43	49.91	78.88	62.44	18.27	59.77	83.69	73.96	90.15
PanopticFCN+D	50.06	77.06	63.83	45.21	70.41	59.74	52.02	78.87	65.00	16.01	**61.54**	84.73	**74.23**	**91.21**
PanopticFCN+B	**52.31**	**77.61**	**65.16**	46.78	**74.88**	61.74	**52.62**	78.40	**66.14**	18.58	59.89	84.33	73.31	90.67
PanopticFCN+A	50.01	77.45	63.48	**48.05**	71.69	**65.76**	50.56	**80.01**	62.79	18.69	61.11	**84.77**	74.02	**91.21**
PanopticUAV	**52.07**	**77.54**	**64.89**	**47.45**	**72.14**	**63.96**	**52.41**	**79.08**	**65.73**	**19.34**	**60.27**	**84.49**	**73.43**	**90.86**

Furthermore, the PanopticFCN+A represents the CBAM attention mechanism. We observed improvements in PQ, PQth and PQst over the baseline of 0.85%, 1.52%, and 0.65%, respectively. In particular, the attention mechanism improved the segmentation accuracy for small targets by incorporating contextual information. Overall, our approach, PanopticUAV, achieved the best results on the UAV-OUC-SEG dataset, with PQ, PQth, and PQst of 52.07%, 47.45%, and 52.41%, respectively.

We have experimented on different feature layers separately for boundary mask enhancement, and the results are shown in Table 2. The best experimental results are obtained when a boundary mask is added to the P5 feature layer. We observe +3.15% PQ, +0.25% PQ_{th} and +2.71% PQ_{st} improvements over the PanopticFCN, but the AP only improves by 0.31%. The detailed results are shown in Table 2.

Table 2: Boundary mask experiments, P2, P3, P4, and P5 represent feature maps from FPN

Method	PQ	SQ	RQ	PQ_{th}	SQ_{th}	RQ_{th}	PQ_{st}	SQ_{st}	RQ_{st}	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
PanopticFCN	49.16	77.27	62.66	46.53	71.64	**63.43**	49.91	**78.88**	62.44	18.27	**48.39**	13.11	11.37	40.46	**63.63**
PanopticFCN+P2	49.52	76.79	63.11	42.81	70.27	58.94	51.44	78.66	64.29	16.07	44.59	10.24	9.17	32.84	63.07
PanopticFCN+P3	50.45	76.80	64.37	44.23	72.11	59.56	52.28	78.15	65.74	**19.61**	46.88	**15.60**	11.04	**42.87**	63.32
PanopticFCN+P4	47.39	76.77	60.32	43.70	72.49	58.80	48.45	78.00	60.75	18.52	46.29	12.17	**11.62**	37.01	63.62
PanopticFCN+P5	**52.31**	**77.61**	**65.16**	**46.78**	**74.88**	**61.74**	**52.62**	**78.40**	**66.14**	**18.58**	**47.84**	**11.89**	**10.48**	**38.44**	**61.39**

Fig. 3 shows the qualitative results of our approach. The first column displays UAV sea images, the second column displays visualization images of PanopticFCN, the third column displays our detection images, and the last column displays the ground truth. In the first row, we observe that the baseline method fails to detect land and buildings, while our method successfully detects them. Moreover, our method detects scattered Ulva prolifera, which the baseline fails to detect. In the second row, the baseline method erroneously detects Ulva prolifera as a vehicle, leading to false detections. Our method addresses this issue. In the third row, we notice that our approach produces more precise boundaries compared to the baseline.

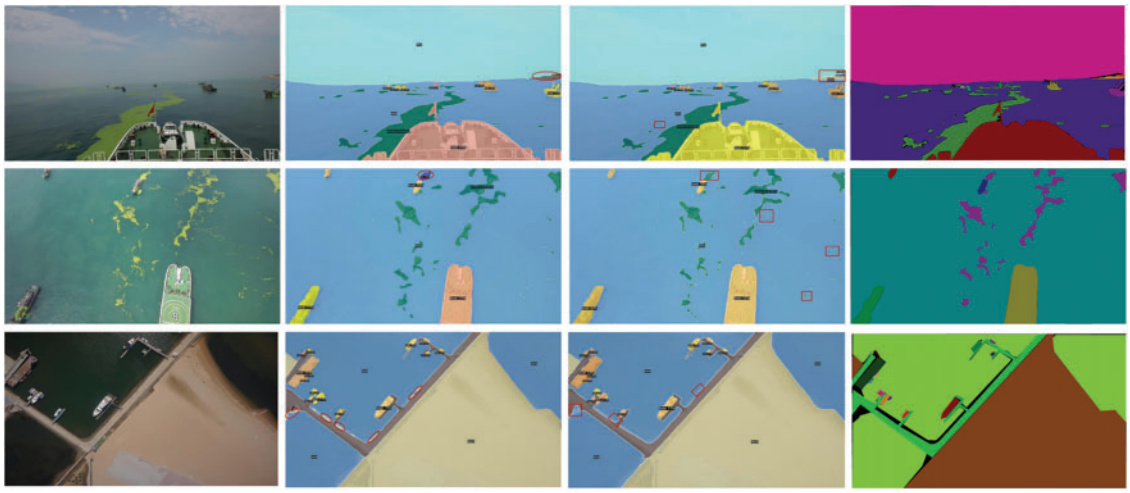

Figure 3: Visualization detection images on UAV-OUC-SEG. The first column shows the UAV sea images, the second column shows the visualization results of PanopticFCN, the third column shows our results, and the last column shows the ground truth

4.4 Comparison Experiments

Table 3 shows the main comparative experiments of our proposed method, PanopticUAV, with other panoptic segmentation methods, including UPSNet [7], PanopticFPN [24], and PanopticFCN [26]. In particular, our approach outperforms the UPSNet approach in the top-down stream with 3.13% PQ, but the PQth of 56.86% is higher than that of our approach, demonstrating that the bottom-up approach performs well for things. Compared to the top-down method, PanopticFPN, our method achieves a 4.79% improvement in PQ. Although PanopticFCN is a unified panoptic segmentation method that performs well in driverless scenes, we have added some strategies based on it. Ultimately, our approach, PanopticUAV, achieves a 2.91% improvement in PQ over the previous baseline.

Table 3: Comparison of the PQ between PanopticUAV and UPSNet, Panoptic-DeepLab and Panoptic FCN on UAV-OUC-SEG

Method	PQ	SQ	RQ	PQ_{th}	SQ_{th}	RQ_{th}	PQ_{st}	SQ_{st}	RQ_{st}
PanopticFPN	47.31	75.61	62.40	43.60	69.70	61.36	50.41	78.08	63.10
UPSNet	48.94	76.00	63.80	**56.86**	**78.63**	**70.90**	50.90	78.50	63.40
PanopticFCN	49.16	77.27	62.66	46.53	71.64	63.43	49.91	78.88	62.44
PanopticUAV	**52.07**	**77.54**	**64.89**	**47.45**	**72.14**	**63.96**	**52.41**	**79.08**	**65.73**

5 Discussion

UAV-based marine image detection is an excellent complement to traditional marine environmental monitoring, and achieving precise segmentation of sea images acquired by drones is an urgent issue. To improve the segmentation accuracy for UAV sea images, we designed a robust feature representation network that utilizes deformable convolutions to obtain a more flexible receptive field suitable for handling the intraclass variability and small objects commonly found in UAV sea images. Additionally, we incorporated the CBAM attention module to parse images more accurately. However,

the segmentation results often suffer from imprecise boundary information due to the complex and diverse nature of the marine scenes. To address this issue, we utilized Laplace operator boundary enhancement to obtain more accurate boundary information that can be fused into the feature maps.

One concern regarding our study is that our method fails to identify some scattered and polymorphic objects in the image. Furthermore, misdetection still occurs in drone images with high-resolution and small objects. Fig. 4 illustrates some examples of segmentation failures of our method, where the first image inaccurately divides the objects, and the second image regards the boat as a road. One possible reason for these limitations is the unevenness and complexity of the dataset. Another reason is the need to further improve the algorithm. Despite these limitations, our proposed approach, PanopticUAV, outperforms most other advanced approaches on the UAV-OUC-SEG dataset.

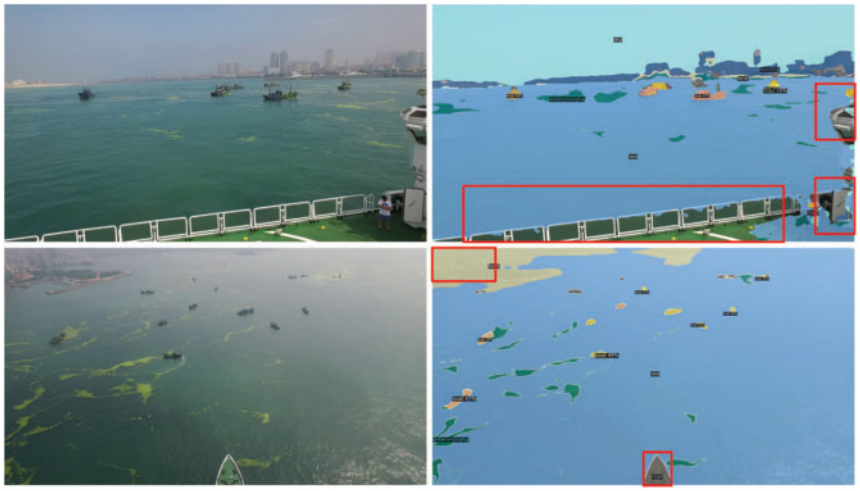

Figure 4: For some complex scenes and sporadic objects, the segmentation results by PanopticUAV are poor, with problems of missed and false detection

6 Conclusion

UAV-based marine monitoring has significant advantages and plays a vital role in marine environmental protection. However, due to the high similarity between some objects in the sea, traditional segmentation algorithms perform poorly, and panoptic segmentation provides a promising solution to this problem. Despite its potential, there is limited research on UAV-based marine image recognition using panoptic segmentation, and no publicly available panoptic segmentation dataset for UAV images exists. To fill this gap, we collected and annotated drone ocean images to create a new dataset called UAV-OUC-SEG.

UAV images captured by drones that contain polymorphic objects pose significant challenges to traditional object detection algorithms, which often fail to accurately identify such objects. To address this issue, we propose a panoptic segmentation method called PanopticUAV, which can accurately identify polymorphic objects to some extent. Our approach incorporates the CBAM attention module to improve the accuracy of image parsing. However, due to the complex and diverse nature of marine scenes, the segmentation results often suffer from imprecise boundary information. To overcome this issue, we utilize Laplacian boundary enhancement to obtain more accurate boundary information that can be fused into the feature maps.

After observation, PanopticUAV method performs well on UAV-OUC-SEG dataset, but there are still some problems for sporadic targets and complex scenes, which need further study.

Acknowledgement: We want to thank "Qingdao AI Computing Center" and "Eco-Innovation Center" for providing inclusive computing power and technical support of MindSpore during the completion of this paper.

Funding Statement: This work was partially supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, the Natural Science Foundation of Shandong Province under Grants Nos. ZR2020MF131 and ZR2021ZD19, and the Science and Technology Program of Qingdao under Grant No. 21-1-4-ny-19-nsh.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Yuling Dou; Data collection: Fengqin Yao, Liang Qu; Analysis and interpretation of results: Xiandong Wang, Yuling Dou; Draft manuscript preparation: Yuling Dou; Investigation: Long Chen, Zhiwei Xu; Visualization: Laihui Ding; Writing-review & editing: Leon Bevan Bullock; Supervision: Guoqiang Zhong, Shengke Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Yan, C., Pang, G., Bai, X., Liu, C., Ning, X. et al. (2021). Beyond triplet loss: Person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia, 24,* 1665–1677.
2. Wang, C., Bai, X., Wang, X., Liu, X., Zhou, J. et al. (2020). Self-supervised multiscale adversarial regression network for stereo disparity estimation. *IEEE Transactions on Cybernetics, 51(10),* 4770–4783.
3. Zhou, L., Bai, X., Liu, X., Zhou, J., Hancock, E. R. (2020). Learning binary code for fast nearest subspace search. *Pattern Recognition, 98(1),* 107040.
4. Zhang, J., Wang, X., Bai, X., Wang, C., Huang, L. et al. (2022). Revisiting domain generalized stereo matching networks from a feature consistency perspective. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12991–13001. New Orleans, LA, USA.
5. Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P. (2019). Panoptic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9396–9405. Long Beach, CA, USA.
6. Yang, T. J., Collins, M. D., Zhu, Y., Hwang, J. J., Liu, T. et al. (2019). DeeperLab: Single-shot image parser. arXiv preprint arXiv:1902.05093.
7. Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M. et al. (2019). UPSNet: A unified panoptic segmentation network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8810–8818. Long Beach, CA, USA.
8. He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. Venice, Italy.
9. Ying, H., Huang, Z., Liu, S., Shao, T., Zhou, K. (2021). EmbedMask: Embedding coupling for instance segmentation. *International Joint Conference on Artificial Intelligence International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, pp. 1266–1273.
10. Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z. et al. (2021). Rethinking BiSeNet for real-time semantic segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9711–9720. Nashville, TN, USA.

11. Tong, X., Ying, X., Shi, Y., Zhao, H., Wang, R. (2021). Towards cross-view consistency in semantic segmentation while varying view direction. *International Joint Conference on Artificial Intelligence International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, pp. 1054–1060.
12. Cordts, M., Omran, M., Ramos, S., Scharwachter, T., Enzweiler, M. et al. (2016). The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223. Las Vegas, NV, USA.
13. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5000–5009. Venice, Italy.
14. Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N. et al. (2021). Panoptic segmentation: A review. arXiv preprint arXiv:2111.10250.
15. Wang, S., Liu, L., Qu, L., Yu, C., Sun, Y. et al. (2019). Accurate Ulva prolifera regions extraction of UAV images with superpixel and CNNs for ocean environment monitoring. *Neurocomputing, 348,* 158–168.
16. Zhang, X., Izquierdo, E., Chandramouli, K. (2019). Dense and small object detection in UAV vision based on cascade network. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 118–126. Seoul, Korea (South).
17. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. Las Vegas, NV, USA.
18. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G. et al. (2017). Deformable convolutional networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773. Venice, Italy.
19. Lin, T. Y., Dollàr, P., Girshick, R., He, K., Hariharan, B. et al. (2017). Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944. Honolulu, HI, USA.
20. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). CBAM: Convolutional block attention module. *European Conference on Computer Vision (ECCV)*, pp. 3–19. Munich, Germany.
21. Wang, C., Wang, X., Zhang, J., Zhang, L., Bai, X. et al. (2022). Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition, 124,* 108498.
22. Bai, X., Wang, X., Liu, X., Liu, Q., Song, J. et al. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition, 120,* 108102.
23. Kirillov, A., Girshick, R., He, K., Dollàr, P. (2019). Panoptic feature pyramid networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6392–6401. Long Beach, CA, USA.
24. Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S. et al. (2020). Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12472–12482. Seattle, WA, USA.
25. Li, Y., Zhao, H., Qi, X., Wang, L., Li, Z. et al. (2021). Fully convolutional networks for panoptic segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 214–223. Nashville, TN, USA.
26. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. et al. (2020). End-to-end object detection with transformers. *European Conference on Computer Vision*, pp. 213–229. Glasgow, UK.
27. Cheng, B., Schwing, A., Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems, 34,* 17864–17875.
28. Tu, Z., Chen, X., Yuille, A. L., Zhu, S. C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision, 63(2),* 113–140.
29. Zhang, R., Xie, C., Deng, L. (2023). A fine-grained object detection model for aerial images based on YOLOv5 deep neural network. *Chinese Journal of Electronics, 32(1),* 51–63.
30. Liang, D., Chen, X., Xu, W., Zhou, Y., Bai, X. (2022). TransCrowd: Weakly-supervised crowd counting with transformers. *Science China (Information Sciences), 65(6),* 48–61.

31. Chen, Z., Zhong, B., Li, G., Zhang, S., Ji, R. et al. (2023). SiamBAN: Target-aware tracking with Siamese box adaptive network. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4),* 5158–5173.
32. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H. et al. (2022). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 44(11),* 7380–7399.
33. Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M. et al. (2021). FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access, 9,* 89644–89654.
34. Lyu, Y., Vosselman, G., Xia, G. S., Yilmaz, A., Yang, M. Y. (2020). UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing, 165(5),* 108–119.