Academic Editor: Richard Dansereau

Copyright 2024 Pengfei Shi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Instance segmentation is a task that involves pixel-level classification and segmentation of each object instance in images.
CNN-based methods have achieved promising results in natural image instance segmentation. However, the noise interference, low resolution, and blurred edges bring more significant challenges for sonar image instance segmentation. To solve these problems, we propose the Effective Strategy for Sonar Images Instance Segmentation . We introduce ASception, a new network combining Atrous Spatial Pyramid Pooling and Extreme Inception .

Introduction

Instance segmentation constitutes a task focused on categorizing object types and predicting pixel-level results for individual instance. In constrast to object detection, instance segmentation clearly depicts the boundary of each object instance and generates detection and segmentation results. Convolutional neural networks have propelled the applications of deep learning methods in image instance segmentation. At present, there are two main kinds of instance segmentation methods based on deep learning methods: top-down and bottom-up detection methods. Topdown detection methods mainly include FCIS , Mask RCNN , Mask Scoring R-CNN , HTC , etc. The main bottom-up detection methods are YOLACT , SOLO, and so on. The designs of the aforementioned models are all based on optical images. However, due to the complexity of underwater environments and the attenuation of light waves, the quality of optical images obtained by optical equipment in underwater environments is relatively low .

FIGURE 1: Overall architecture of proposed ESSIIS framework.

fishery to other fisheries.
The aforementioned methods either focus on reducing the number of parameters in the network at the expense of accuracy or achieve higher detection accuracy at the cost of significantly increased computational complexity. Our goal is to maximize segmentation accuracy while maintaining a low parameter count. To achieve this, we adopt a strategy that splits instance segmentation into two parallel tasks and improves the modules in the method specifically addressing the characteristics of sonar images. It may strike a relative balance between speed and accuracy in the network.

ResNet to construct a new feature extraction network, ASception-ResNet, which improves the feature learning ability for multi-scale sonar objects.
We utilize the deformable convolution instead of the traditional convolution in the feature extraction network, which makes the receptive field more flexible to facilitate feature extraction.

Proposed Method

Network Architecture of ESSIIS. The overall framework of the method in this paper is shown in Figure 1. First, the sonar image is input, and then the sonar feature maps containing different scale information are extracted by the feature extraction module ASception-ResNet. Following this, the extracted features of different dimensions are fused using the bidirectional fusion module, so that the feature maps contain richer sonar object location and semantic information. This lays the foundation for subsequent sonar image object location and segmentation tasks. The Protonet branch is composed of several convolution layers and is used to generate mask

vectors.
and the output mask dimension is 138 138 32.

Classification loss and regression loss are the same as Mask

ASception-ResNet. ASPP is a technique applied to image semantic segmentation tasks. It can effectively capture contextual information at different scales.

FIGURE 2: The structure of ASception.
objects in sonar images varies due to the interaction of sound waves, distance changes, and physical characteristics of the objects. ASPP employs convolution layers with larger receptive fields to capture multiscale contextual information and adapt to the changes in object scales. Additionally, ASPP includes the global average pooling layer. It can obtain contextual information for the entire image and extract overall features from the feature maps. This acquisition of overall features significantly enhances the ability of the network to accurately classify and recognize objects.

Xception with ASPP to design a new network architecture called

ASception. ASception can effectively extract and represent target features at different scales and model the spatial variations and shape features of the targets. Moreover, we incorporated ResNet into ASception, enabling ASception-ResNet to leverage the advantages of both ASPP and Xception, while also possessing a powerful feature-learning capability. These characteristics make ASception-ResNet more suitable for sonar image processing tasks compared to other feature extraction networks.
The novel feature extraction network ASception-ResNet is divided into five layers C1, C2, C3, C4, and C5. The first four layers C1-C4 use ResNet to extract the features of sonar objects step by step. The fifth layer C5 uses the ASception proposed in this paper, and its structure is shown in Figure 2. First of all, the sonar feature maps output by C4 is respectively passed through three convolutional layers with expansion rates of 6, 12, and 18 and a 1 1 pooling layer, to obtain four groups of features with different receptive fields.
Then, the four groups of features are connected with the input to obtain the output feature O1. The number of channels of O1 is five times that of the input feature map. After the dimension of O1 is reduced by 1 1 convolution, we obtain the features with the same number of channels as the input. Finally, 3 3 convolution processing is performed on the feature channels after dimension reduction to obtain features with richer semantics. The optimized feature extraction network has better feature extraction ability in the instance segmentation task.
Deformable Convolution. Most CNN-based detection methods take ResNet50 as the feature extraction network.

FIGURE 3: Deformable convolution: traditional convolution, deformable convolution, and deformable convolution.

Deformable convolution adds an offset to the traditional convolution operation, which transforms the convolution into an irregular convolution. At the same time, a weight coefficient is added to distinguish whether the region we introduce is the region of interest to us. Therefore, the output of a deformable convolution is as follows: yð pÞ ¼ ∑ wk · xð p þ pk þ pk Þ · mk