Fast-DecoupledNet: An Improved Multi-branch Edge Enhanced Semantic Segmentation Network

Junyu Xue[1], Zhiyuan Zhang[2*]

[1]College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China.
[2]The 91001 Unit of PLA, Beijing, China.
[*]Corresponding author, E-mail: generalzzy@139.com

Abstract. There are existing semantic segmentation methods that incorporate the idea of edge detection by using multi-branch networks to focus on edges and subjects separately, but there are a large number of unfocused aspects and limited improvement. In this paper, we suggest the Fast-DecoupledNet, a semantic segmentation network. We design Edge Feature Extractor to extract the target's edge features more accurately, and the global features obtained by joint downsampling are computed to obtain the subject features and the final features. In addition, we use a shallower ResNet as the backbone network to reduce computational complexity while ensuring computational accuracy. Experiments on the *Deepglobe Land Cover Classification* dataset show that our proposed method achieves 72.59 F-score and 77.64% mIoU.

1. Introduction

1.1. Research Significance of Semantic Segmentation

A key tool in computer vision is semantic segmentation, which tries to assign labels to each pixel in the original image and comprehend the image in terms of its individual pixels. The technique divides multiple pixels in an image that belong to the same set of labels into a region that gives the corresponding semantic category. With the continuous modernization and intelligence of life, semantic segmentation has become the technical basis of several industries and is widely used in remote sensing images, autonomous driving, medical imaging, and other fields. For example, in the field of autonomous driving, semantic segmentation can be applied to distinguish obstacles, driving areas, etc.; in the field of medical images, it can be applied to distinguish the types of diseases in different parts of human beings. Therefore, semantic segmentation is of great research importance and has wide application value.

Deep learning has demonstrated its potent abilities in numerous sectors recently, including remote sensing. It significantly raises the precision and effectiveness of semantic segmentation. At the same time, it also brings a lot of problems and challenges. First, semantic segmentation requires a lot of data for training, but the available datasets are small in quantity and low in quality. At the same time, the types and sizes of datasets vary, and the ultra-high resolution images have high requirements on the computer's arithmetic performance. In real complex situations, semantic segmentation needs to overcome two major challenges: the variability between similar items and the similarity between different classes of items, which is often accompanied by various disturbances such as oversized, undersized, and fragmented items. The effect of semantic segmentation will have a significant impact on the subsequent image processing, and further research on it in the segmentation region is necessary.

1.2. Existing Semantic Segmentation Methods and Drawbacks

Convolutional neural networks are the foundation of today's widely used semantic segmentation approaches. U-net [1] is one of the most effective strategies, which is based on fully convolutional networks for data augmentation of the dataset, incorporating more scales, but is not effective for segmenting large objects. DeeplabV3+ [2] uses null convolution and improves the Decoder structure to improve the segmentation accuracy. Some later works also adopt the self-attention mechanism to enhance the representation learning and feature extraction ability of the model, and use the residual connection to enhance the depth and stability of the model. Later work appeared using two-stream CNN, adding a separate branch to learn edge information to improve the segmentation of fine objects. Subsequently, Li et al [3] proposed a new paradigm for semantic segmentation by decoupling the high-semantic level feature graph into two parts: subject features and boundary features. The subject features are generated by a flow-based approach, learning offsets to deform the internal pixel features of the target and the boundary features can be obtained by subtracting the subject features from the output feature map. However, they work only to extract features under the supervision of loss function, and the extraction effect is limited. Also, the issue of feature extraction order is not further considered. For semantic segmentation and edge detection, Zhen et al. [4] introduced a collaborative multi-task learning architecture that stores shared latent semantics to facilitate interaction between tasks. But it focuses more on the coupling of the two tasks rather than enhancing the semantic segmentation effect with edge detection.

In summary, previous work has had the idea of joint edge detection and semantic segmentation, but there are still a large number of unconsidered problems.

Our main contributions are summarized as follows:

- We propose a new network called Fast-DecoupledNet for semantic segmentation, which achieves the best synthesis results on the corresponding dataset.
- We designed the Edge Feature Extractor module, which can extract edge features more accurately and compute subject features in combination with global features.
- We used a more shallow ResNet as the backbone network to maintain segmentation accuracy while reducing computational complexity.

2. Related Work

2.1. Semantic Segmentation

Traditional machine learning methods such as random forests have been used to perform semantic segmentation. With the ongoing advancement of deep learning methodologies, most of the research methods are now based on FCNs. This was the first network based on CNN architecture and achieved a huge accuracy leap, using deconvolution for upsampling. Chen et al. [5] proposed GLNet in 2019, which innovatively adopts a semantic segmentation method that integrates global and local information and effectively aggregates features. It uses global branches for rough segmentation, and then uses local branches for fine segmentation after obtaining the foreground region, but this consideration is relatively rough. Shan et al [6] improved GLNet's branching structure and proposed the first local feature fusion method, which allowed the cropped chunks to learn features from the surrounding. MBNet [7] designed a multi-branching structure to solve the multi-resolution input problem and also introduced an attention mechanism to enhance the fusion effect.

2.2. Muti-branch Network

The downsampling operation in fully convolutional networks leads to the blurring of edge features, so multi-branch networks have been proposed for optimizing details. Yu et al [8] proposed a two-branch network that uses different branches to parse the context and details and finally uses a feature fusion module to fuse them. After this several works have been proposed for such architectures, introducing attention mechanisms or improving their branches.

Our approach also uses a two-branch structure to capture global features through the ASPP branch, with the difference that we use the Edge Feature Extractor branch to detect edge features. And we design a shallower backbone network that can reduce computational complexity while ensuring the extraction of higher quality edge features.

2.3. Edge Optimization

Back in 2016, Chen et al [9] explicitly extracted edge cues as constraints and allowed CNNs to learn edge feature maps. At the same time, there have been many efforts to improve edge extraction through better structural modeling, but they do not directly process the boundary pixels. Li et al [3] extracted the subject and edge features of the image separately inspired by Gaussian filtering, which improved the internal consistency of the object as well as the edge segmentation effect. However, Decoupled SegNet [3] only supervises the training by a special loss function and finally obtains the edge features computationally. On the other hand, our work directly designs an Edge Feature Extractor to extract edge features, computes the subject features, and then fuses the global features to get the final feature map, which achieves better results on the test set.

3. Method

In this section, our entire framework is initially introduced in 3.1, and later we will focus on the multi-branch network structure and its advantages in 3.2.

3.1. Overview

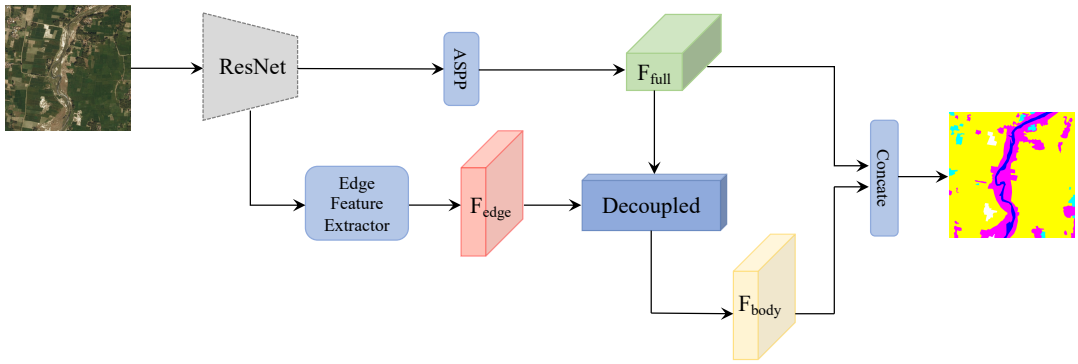

Figure 1. Overview of the whole network architecture.

The overall network structure is shown in Fig. and consists of five parts, which are Edge Feature Extractor, Backbone Network, ASPP Module, Decoupled Module, and Concatenation Module. For the design of the whole network, we borrowed the structure of the state model Deeplabv3+ [2], using ResNet as the backbone network. In particular, we use a more shallow ResNet, which in turn reduces the computational complexity. In addition, we designed Edge Feature Extractor to extract edge information. the Decoupled module processes the resulting features and integrates them to obtain the final feature map.

3.2. Decoupled Module

The features F of the whole image can be decoupled into two parts, F_{body} and F_{edge}. Using the same assumptions as in Decoupled SegNet [3], F_{body} and F_{edge} conform to the addition rule. The multi-branching module is mainly used to process the edge features extracted by Edge Feature Extractor, which are subtracted from the global features extracted by the ASPP module to obtain the body features. Finally, the subject features and the global features obtained by the previous downsampling are integrated and output.

In the traditional Decoupled SegNet [3], its final features are generated under the supervision of two loss functions of subject and edge. Among them, the edge features are obtained by subtracting the global features from the subject features. We employ an Edge Feature Detection module to directly extract edge features, improving efficiency. At the same time, we separately optimized for the edge branch with an acceleration process and used a more shallow network.

3.3. Lose Function

$$L = \lambda_1 L_{body}(S_{body}, \hat{S}) + \lambda_2 L_{edge}(S_{edge}, \hat{S}) + \lambda_3 L_{final}(S_{final}, \hat{S}) \quad (1)$$

where $S_{body}$and S_{edge} are the semantic segmentation graphs predicted according to subject features and edge features, and $\hat{S}$ is the ground truth with label values.

4. Experiment

4.1. Dataset

The *DeepGlobe Land Cover Classification* dataset is a publicly available, high-resolution, sub-meter satellite image of predominantly rural areas. The dataset contains 1146 satellite images of 2448×2448 pixel size. RGB data with a pixel resolution of 0.5 m is present in each image. It is used for multi-class segmentation tasks where the model needs to detect urban, agricultural, pasture, forest, water, barren, and unknown regions in the image. The dataset is extremely challenging due to the diversity of land cover types and the number of annotations.

4.2. Experimental Results and Analysis

Based on the above four evaluation indicators, we have carried out a wealth of experiments. We used Precision, Recall, F-score, and mIoU as evaluation metrics.

Table 1. Deepglobe test set boundary region segmentation results.

Method	Precision	Recall	F-score	mIoU(%)
Ours	75.10	70.24	72.59	77.64
Decoupled SegNet [3]	68.59	60.51	61.74	73.75
MBNet [7]	70.99	54.75	58.83	72.12
UhrsNet [6]	74.93	56.68	61.35	73.62
Deeplabv3+ [2]	73.91	58.88	63.66	74.88
DlinkNet34 [10]	74.59	60.76	64.93	75.30
ConDinet++ [11]	68.69	70.44	67.96	76.58

Compared with ConDinet++ and other models, the earlier proposed models such as DlinkNet34 and Deeplabv3 achieve higher Precision scores, but the Precision scores of all models are lower than ours. MBNet and UhrsNet have low Recall scores. The Recall scores of ConDinet++ and our model both exceed 70. F-score is a thorough evaluation of Precision and Recall as it is the harmonic mean of both. At this point, only our model exceeds 70 points. ConDinet++, which has the highest score among other models, has only 67.96. Our baseline Decoupled SegNet has a slightly worse Precision score, but a higher Recall score. On the same dataset, our model has a small gap with SOTA in the Recall metric, but has the highest Precision, Recall, and mIoU. Overall, our model performs the best.

4.3. Ablation Experiment

In this section, we will experiment with the effect of a more shallow backbone network and Edge Feature Extractor. Our backbone network is ResNet, and we will also test the effect of the ratio of the loss functions of global features, edge features, and subject features on the results.

The impact of Shallow Backbone. Since changing the number of convolutional layers has a small effect on the segmentation accuracy, in this section we mainly compare the computational complexity of ResNet with different numbers of layers. We used ResNet with 6, 12, 24, and 50 layers of convolutional neural network, respectively, and the results are shown in Table 2. The shallower layers of the backbone network greatly reduce the computational complexity while maintaining accuracy. In Fast-DecoupledNet, we used ResNet with 6 convolutional layers.

Table 2. Comparison of computational complexity of ResNet with different layers.

Model	Computational complexity (FPS)
ResNet6	5.6
ResNet12	8.7
ResNet24	15.0
ResNet50	23.0

The effect of Edge Feature Detection module. We remove the Edge Feature Detection module and choose to use the traditional up and down sampling. We use the previous four indicators to measure, and the values of the four indicators Precision, Recall, F-score and mIoU are reduced to 0.4, 0.3, 0.34 and 30% respectively. The effect is significantly reduced.

The effect of loss function ratio. In line with other works, we change the ratio of the L_{full}, L_{edge}, and L_{body}, and measure the effect by mIoU. As the results in Table 3 show, when the loss function ratio is changed to 1:2:2 or 1:2:4, the values of mIoU all decrease and the effect becomes worse.

Table 3. Effects of different loss function ratios on mIoU.

Loss Ratio	mIoU(%)
1:1:1	77.64
1:2:2	77.54
1:2:4	77.54

4.4. Visual Analysis

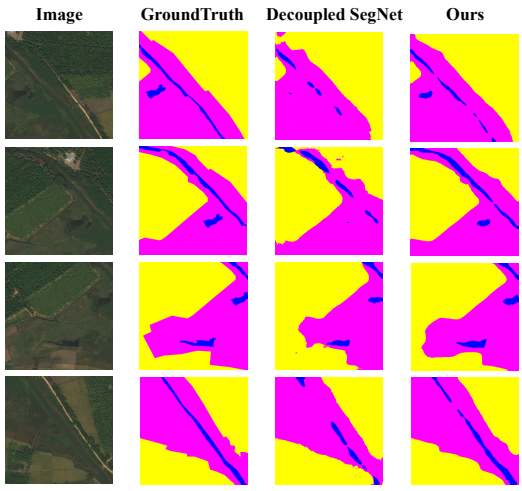

Figure 2. Example of segmentation results on *Deepglobe Land Cover Classification* dataset. Compared with Decoupled SegNet, our method improves the results for these three cases.

By comparison, the results of our model are significantly better. In terms of edge processing, our results are more precise and fluid. In terms of the whole image, our results are richer in detail and have higher accuracy.

5. Conclusion

In this paper, we propose a novel multi-branch network for semantic segmentation with fused edge enhancement. We extract the edge features in the image by Edge Feature Extractor and combine the global features acquired by the ASPP module to subtract the subject information. We use a modified ResNet as the backbone network, and shallow processing reduces the computational complexity of the entire network, making it lighter. We achieved state-of-the-art results on the *DeepGlobe Land Cover Classification* dataset and proved the effectiveness of the whole network.

References

[1] Ronneberger, O., Fischer, P., & Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference. Munich, Germany. pp. 234-241.

[2] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV). pp. 801-818.

[3] Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., Tan, S., & Tong, Y. (2020) Improving Semantic Segmentation via Decoupled Body and Edge Supervision (arXiv:2007.10035).

[4] Zhen, M., Wang, J., Zhou, L., Li, S., Shen, T., Shang, J., ... & Quan, L. (2020) Joint semantic segmentation and boundary detection using iterative pyramid contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13666-13675.

[5] Chen, W., Jiang, Z., Wang, Z., Cui, K., & Qian, X. (2019) Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8924-8933.

[6] Shan, L., Li, M., Li, X., Bai, Y., Lv, K., Luo, B., ... & Wang, W. (2021) Uhrsnet: A semantic segmentation network specifically for ultra-high-resolution images. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 1460-1466.

[7] Shan, L., & Wang, W. (2022) MBNet: A Multi-Resolution Branch Network for Semantic Segmentation Of Ultra-High Resolution Images. In: 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2589-2593.

[8] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018) Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325-341.

[9] Chen, L. C., Barron, J. T., Papandreou, G., Murphy, K., & Yuille, A. L. (2016) Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4545-4554.

[10] Zhou, L., Zhang, C., & Wu, M. (2018) D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 182-186.

[11] Yang, K., Yi, J., Chen, A., Liu, J., & Chen, W. (2021) ConDinet++: Full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images. IEEE Geoscience and Remote Sensing Letters, 19, 1-5.