IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10642684

A SEMANTIC SEGMENTATION METHOD FOR SAR IMAGE WITH ASSISTANCE OF SELF-SUPERVISED SCENE CLASSIFICATION

Yang Cheng[1], Chenxuan Li[1], Zenghui Zhang[1], Wenxian Yu[1]

[1]Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, Shanghai, 200240, China

ABSTRACT

Unlike natural images, synthetic aperture radar (SAR) images exhibit a more scattered and uneven spatial distribution of objects, making semantic segmentation of SAR images a valuable topic of research. This paper presents a SAR image semantic segmentation method that incorporates the attention mechanism assisted by self-supervised scene classification. The self-supervised scene classification provides coarse scene classification at a higher semantic level, while the attention mechanism utilizes high-level semantic features to guide fine-grained classification of lower-level spatial structures. Overall, this approach improves the pixel-level classification performance of SAR images. We validate this method on the WHU-OPT-SAR dataset and compare its performance with previous works, providing a detailed analysis of its effectiveness.

***Index Terms*—** SAR, Semantic Segmentation, Attention Mechanism, Contrastive Learning

1. INTRODUCTION

Synthetic aperture radar (SAR) is a significant microwave imaging technology which is widely employed on various flight platforms such as aircraft and satellites. Its ability of surface penetration, coupled with its ability to provide continuous ground observation under all weather conditions, renders it a crucial tool in remote sensing applications. Due to the low resolution ratio of satellite imagery and the complex distribution of ground objects, a single remote sensing image often covers a large land area. Often, there is no fixed geometric or spatial structure in the distribution of objects, presenting a scattered pattern. These factors pose challenges to the semantic segmentation of remote sensing images.

One of the traditional methods for semantic segmentation is the Fully Convolutional Network (FCN). It utilizes a symmetrical encoder-decoder network architecture to extract high-level features, which are subsequently decoded to match the original image pixel resolution. To mitigate the loss of spatial information in high-level features, the Deeplabv3[1] employs dilated convolution kernels of varying scales to ensure a broad receptive field while minimizing downsampling rates. Another approach, PSPNet[2], adopts global average pooling to fuse multi-scale information, thereby enhancing contextual understanding during high-level feature extraction. The Squeeze-and-Excitation (SE) structural block[3] employs global pooling to generate channel attention mechanisms, with the aim of strengthening distinct feature channels to make targeted attention and context comprehension for different target categories. However, due to the scattered and uneven distribution of land cover in SAR images, unlike natural images, the application of image-level pooling to remote sensing images may not be able to ideally capture the characteristics of the target area.

In this paper, in order to extract high-level semantic features from images while implicitly constraining the distribution of categories in SAR images, we adopt the constractive clustering method[4], using a self-supervised scene clustering auxiliary task to further enhance the semantic representation capability of high-level features. Additionally, we refer to the LANet structure[5], to refine the pooling granularity by introducing local attention and fusion attention mechanisms. The local attention mechanism enhances contextual understanding of the scene, while the fusion attention mechanism establishes the association between high-level semantic features and low-level spatial features to enhance the semantic representation capability of the latter.

2. METHODOLOGY

2.1. Unsupervised Scene Clustering Auxiliary Task

SAR images exhibit scattered terrain distribution, and there are significant differences in pixel distribution between urban and rural terrains. Therefore, it is necessary to perform coarse scene classification in advance, as it can enhance the high-level features and therefore enhance the performance of SAR image semantic segmentation. As a result, we refer to the method proposed by[4], using the method of contrastive clustering. As shown in Fig. 1, we assume a batch of N images, and construct their positive samples through image augmentation and K nearest neighbors (KNN) clustering. As a result, we have N positive pairs for instance-level learning. Besides, we plan to divide M clusters for the cluster-level learning.

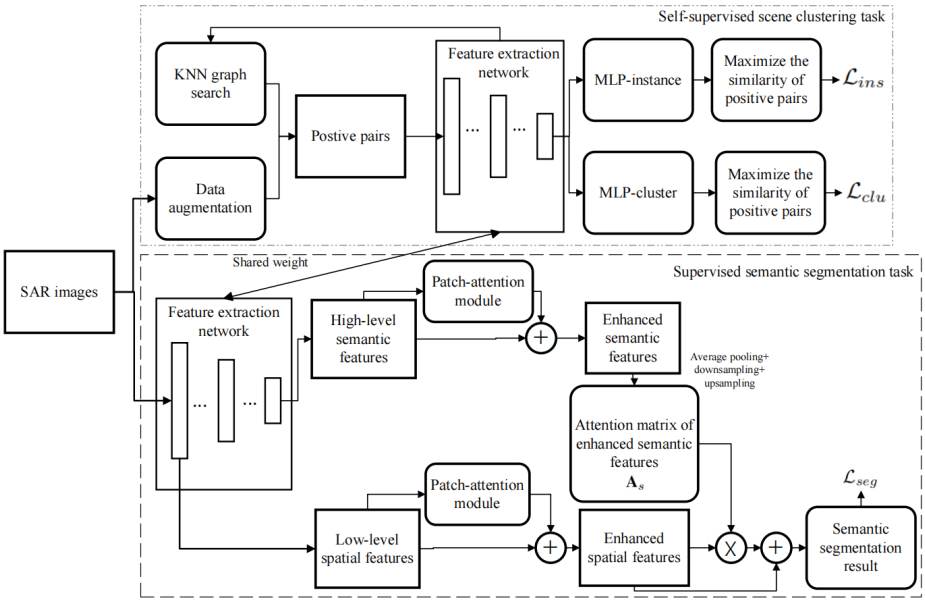

Fig. 1. Overall framework of the proposed method, including the self-supervised scene clustering task and the supervised semantic segmentation task. The self-supervised scene classification auxiliary task further enhances the semantic representation capability of high-level features and implicitly constrains the distribution of object categories in images. The supervised pixel-level semantic segmentation task introduces local attention mechanism and fusion attention mechanism to enhance contextual understanding through high-level semantic features and low-level spatial features.

Subsequently, these images are passed through a feature extraction network for instance-level contrastive learning and cluster-level contrastive learning. Finally, we obtain two loss functions $\mathcal{L}_{ins}$ and $\mathcal{L}_{clu}$, each derived from instance-level contrastive learning and cluster-level contrastive learning. As is mentioned in the work[4], the detailed definition of $\mathcal{L}_{ins}$ and $\mathcal{L}_{clu}$ are as follows:

$$\mathcal{L}_{ins} = \frac{1}{2N}\sum_{i=1}^{N}(\ell_i^a + \ell_i^b) \tag{1}$$

Where $\ell_i^a = \frac{exp(s(x_i^a, x_i^b)/\tau_I)}{\sum_{j=1}^{N}[exp(s(x_i^a, x_j^a)/\tau_I) + (s(x_i^a, x_j^b)/\tau_I)]}$, $s(\cdot)$ represents the cosine distance, τ_I is the temperature coefficient used in instance-level contrastive learning to control the curve of the loss function, and x_i^a, x_j^a are the feature of positive pairs extracted from the feature extraction network. Others on the denominator are negative pairs. ℓ_i^b is similarly defined as ℓ_i^a.

$$\mathcal{L}_{clu} = \frac{1}{2M}\sum_{i=1}^{M}(\hat{\ell}_i^a + \hat{\ell}_i^b) - H(Y) \tag{2}$$

Where $\hat{\ell}_i^a = \frac{exp(s(y_k^a, y_k^b)/\tau_C)}{\sum_{j=1}^{M}[exp(s(Y_i^a, Y_j^a)/\tau_C) + (s(y_i^a, y_j^b)/\tau_c)]}$, τ_C is the temperature coefficient used in cluster-level contrastive learning, and y_i^a, y_j^a are the positive cluster pairs extracted from the feature extraction network. Others on the denominator are negative pairs. $\hat{\ell}_i^b$ is similarly defined as $\hat{\ell}_i^a$. In order to maintain a relatively balanced distribution of cluster categories, we introduce the entropy of cluster assignment probabilities $H(Y)$, as is mentioned in the work[4].

2.2. Supervised Pixel-level Semantic Segmentation Task

2.2.1. Local Attention Mechanism

The attention mechanism enables a comprehensive understanding of both semantic and spatial information from features, which is particularly suitable for SAR images characterized by scattered distributions and diverse semantic information. Therefore, this paper adopts the method proposed in[5] to address the challenge.

We assume that the feature extraction network can extract feature maps with C channels. After passing through the feature network, a single image generates a series of feature maps $\{X_1, \cdots, X_C\}$ of size $(H \times W)$. Subsequently, the feature maps are fed into the patch-attention module. For channel $c, c \in [1, C]$, we firstly compute its blockwise global descriptor z_c using average pooling: $z_c = \frac{1}{h_p w_p}\sum_{i=1}^{h_p}\sum_{j=1}^{w_p} x_c(i,j)$, where h_p, w_p represents the window size for average pooling.

Let $H' = h/h_p$ and $W' = W/w_p$, after aggregating the descriptors for each channel, we obtain a C-dimensional feature vector $\mathbf{z}_p \in \mathbb{R}^{C\times H'\times W'}$. Subsequently, the features are dimensionally reduced and restored to C dimensions by using the convolution kernels $Conv_d$ and $Conv_u$, achiev-

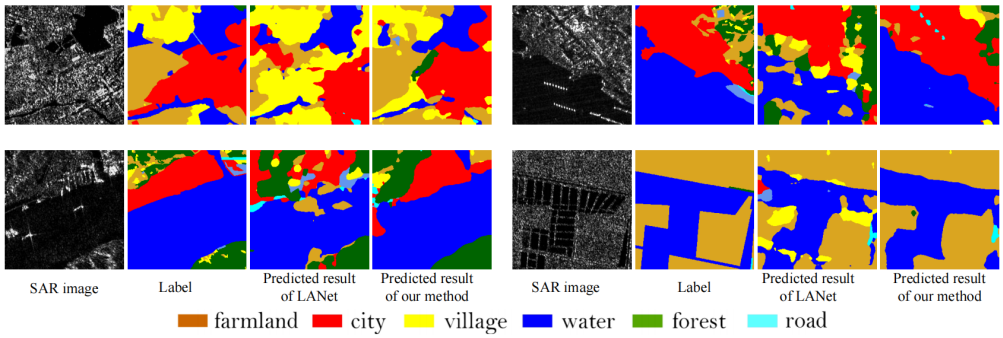

Fig. 2. Comparison of the predicted results of our method with the LANet.

Table 1. Comparison of performance of different semantic segmentation methods.

Accuracy / Method	Average pixel accuracy	Average class precision	Average class recall	Average class F1-score
FCN+SE[3]	0.7223	0.5579	0.5955	0.5631
Deeplabv3[1]	0.7473	0.5823	0.6153	0.5880
PSPNet[2]	0.7518	0.5810	0.6439	0.5955
LANet[5]	0.7347	0.5861	0.6373	0.6034
our method	0.7714	0.6162	0.6714	0.6527

ing context information exchange and fusion. Then we get the image's attention matrix $\mathbf{A}_p$ via: $\mathbf{A}_p = \mathit{Sigmoid}[\mathit{Conv}_u(\mathit{ReLU}[\mathit{Conv}_d(\mathbf{z}_p)])]$.

The image's attention matrix $\mathbf{A}_p$ is subsequently upsampled to the original size $(C \times H \times W)$. It is then multiplied with the input feature map to obtain enhanced features. The process above constitutes the patch-attention module.

We adopt a residual structure to stabilize the training process, where the enhanced features are added to the original feature maps to obtain the final enhanced features.

2.2.2. *Fusion Attention Mechanism*

The motivation of the method is to embed semantic attention from higher-level features into lower-level features, enabling the low-level features of SAR images to acquire contextual and semantic information beyond their limited receptive fields, while preserving spatial details without resolution reduction.

Similar to the patch-attention model, through global pooling and context information exchange,we obtain the attention matrix $\mathbf{A}_f$, which is used to guide the enhancement of the enhanced spatial features $\mathbf{X}_L$ by $\mathbf{X}_L = \mathbf{X}_L + \mathbf{X}_L \times \mathbf{A}_s$. $\mathbf{A}_s$ represents the upsampling operation performed to adapt $\mathbf{A}_f$ to the dimensions of $\mathbf{X}_L$.

2.3. The Loss Function

By employing the conventional pixel-wise cross-entropy to represent the loss function of supervised pixel-level semantic segmentation $\mathcal{L}_{seg}$, the loss function $\mathcal{L}$ of the whole process can be described as: $\mathcal{L} = \mathcal{L}_{ins} + \mathcal{L}_{clu} + \mathcal{L}_{seg}$.

3. EXPERIMENTS

3.1. Dataset and training details

The semantic segmentation experiments were conducted on the WHU-OPT-SAR dataset[6], which has relatively high resolution and segmentation accuracy. The dataset consists of 7 semantic classes, including farmland, city, village, water, forest, road and others. The original WHU-OPT-SAR dataset has images of size 5556×3704, which is not suitable as direct input for deep neural networks. Therefore, we used non-overlapping slicing to obtain image slices in the size of 256×256. This process resulted in a final dataset of $29.4k$ image slices. In this paper, we utilized a ResNet18 architecture as the feature extraction network. The batch size was set to 32. For the KNN clustering task, the hyperparameter K was set to 3, and the number of clusters in constractive learning was set to 10.

3.2. Results and Discussion

The model performance evaluation considers both pixel-level evaluation and class-level evaluation. Pixel-level evaluation utilizes the average pixel accuracy metric, which measures the accuracy of individual pixels across the entire image. Class-level evaluation employs commonly used evaluation metrics

Table 2. Comparsion of model size and calculations.

Cost \ Method	FCN+SE	Deeplabv3	PSPNet	LANet	Our method
Number of parameters (millions)	2.83	39.05	53.38	2.83	2.91
Memory occupancy (MB)	35.87	164.27	428.78	35.87	35.87
Computation (GFlops)	1.87	11.99	50.51	1.87	1.87

Table 3. Ablation study for assistance of self-supervised auxiliary task.

Auxiliary tasks \ Accuray	Average pixel accuracy	Average class precision	Average class recall	Average class F1-score
No auxiliary tasks	0.7347	0.5861	0.6373	0.6034
Instance contrastive-level tasks	0.7702	0.6137	0.6523	0.6212
cluster-level contrastive tasks	0.7621	0.5922	0.6431	0.6085
Both contrastive tasks	0.7714	0.6162	0.6714	0.6257

such as class average precision, recall, and F1 score. Considering the relatively low proportion nature of the "others" category in the WHU-OPT-SAR dataset, we exclude the "others" classes when calculating class average metrics. Instead, we focus on calculating the average values for the six other main classes representing clear semantic categories.

We compare the performance of our proposed method with other SAR semantic segmentation approaches, and the results are presented in the Table 1. The comparison between our method and the LANet approach can be observed in Fig. 2. Our method achieves more accurate semantic segmentation due to the adoption of a self-supervised coarse scene classification approach, which implicitly constrains the pixel distribution within each scene. For instance, the variations in brightness within the city (the red area) of SAR images often lead to misclassifications as village (the yellow area), forest (the green area), or farmland (the brown area) by traditional semantic segmentation methods. Our method demonstrates promising performance in addressing this issue. Furthermore, SAR images tend to misclassify farmland area as villages. Overall, our approach effectively incorporates semantic contextual information, resulting in better intra-class consistency at the pixel level and higher overall accuracy in semantic segmentation.

Specially, our proposed method utilizes a ResNet18 backbone network, making it a relatively lightweight network. The attention-guided and contrastive learning branches are modular and can be easily incorporated using 1x1 convolutional layers or simple MLP layers, providing flexibility. Other networks compared in this paper such as Deeplabv3 and PSPNet, either employ ResNet50 as their backbone network or incorporate fusion operations like feature pyramid to enhance multi-scale feature concatenation, resulting in larger parameter sizes and more complex network design. The model size and computational complexity comparisons are shown in Table 2, using statistics obtained from the torchstat tool library. Table 2 shows that our method achieves superior performance while maintaining smaller parameter sizes and computational complexity, making it easier to train and apply.

Additionally, we investigate the impact and benefits of instance-level self-supervised tasks and clustering-level self-supervised tasks as auxiliary tasks for semantic segmentation, as shown in Table 3. The results demonstrate that both instance-level contrastive tasks and cluster-level contrastive tasks can provide auxiliary benefits to the main task, but instance-level tasks achieve better performance improvements. Instance-level contrastive tasks are more effective in optimizing feature extraction. In the context of the semantic segmentation main task, where there is a mixture of objects and unclear separability of scene categories, the role of clustering tasks may be somewhat limited. In contrast, instance-level tasks provide more explicit self-supervision signals.

4. CONCLUSIONS

In this paper, we propose a semantic segmentation method for SAR image by applying self-supervised scene classification assistance method and incorporating an attention mechanism to facilitate the fusion of spatial and semantic features. Compared to previous works, our approach achieves performance improvements across multiple evaluation means. These findings may inspire future research work in this field.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 62271311 and 62071333, and in part by the Fundamental Research Funds for the Central Universities under Grant USCAST2022-33.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," in *arXiv preprint arXiv:1706.05587*, 2017.

[2] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Chenxuan Li, Weiwei Guo, Zenghui Zhang, and Tao Zhang, "Self-supervised classification of sar images with optical image assistance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[5] Lei Ding, Hao Tang, and Lorenzo Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426–435, 2021.

[6] Guo Zhang Xue Li, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang, "Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 106, pp. 102638, 2022.