Performance and Analysis of FCN, U-Net, and SegNet in Remote Sensing Image Segmentation Based on the LoveDA Dataset

Shuhao Yang

College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, 300000, China

Abstract. Remote sensing image segmentation is a vital method in image analysis that significantly contributes to the extraction of surface information and aids in land use planning. This study utilizes the LoveDA dataset to investigate the segmentation performance of three classic deep learning models: Fully Convolutional Networks(FCN), U-Net, and SegNet, in both urban and rural scenarios. By partitioning the Urban-Rural dataset of LoveDA for training and testing, it was determined that SegNet excels in detail restoration and boundary handling, while U-Net demonstrates robust adaptability across various scenarios. In contrast, FCN, with its simpler architecture, shows lower segmentation accuracy in certain contexts. This paper offers a comprehensive comparison of the strengths and weaknesses of different models in remote sensing image segmentation and proposes enhancements in model structure and data preprocessing optimization. The findings provide valuable insights for improving the performance of semantic segmentation models and are of significant importance for the precise analysis and practical applications of remote sensing images.

1 Introduction

Remote sensing images are defined as images captured of the Earth's surface by airborne or satellite sensors. These images provide auxiliary ground information related to roads, meteorology, agriculture, and more. Through the analysis of these images, people can enhance land use in both urban and rural areas, as well as detect maritime traffic routes for vessels [1,2].

However, the high resolution and multispectral characteristics of remote sensing images pose significant challenges for processing, particularly in terms of efficiently extracting information and accurately performing semantic segmentation. In light of these challenges, image analysis plays a crucial role in the processing of remote-sensing images. Semantic segmentation, a fundamental image analysis technique, classifies each pixel within an image, thereby facilitating the extraction of more valuable information. This makes it a topic deserving further investigation [3].

Corresponding author: yangshuhao@mail.tust.edu.cn

In recent years, advancements in remote sensing imaging technology have led to the availability of more detailed and abundant remote sensing images. Traditional machine learning models have faced challenges in enhancing performance when applied to semantic segmentation tasks [4]. In contrast, deep learning models, after being trained on extensive datasets, have demonstrated superior performance, establishing themselves as the mainstream approach for semantic segmentation in remote sensing image analysis. Long et al. [5] proposed Fully Convolutional Networks (FCN), which replaced fully connected layers with convolutional layers, thus enabling pixel-level classification of images of any size. This development marked a significant advancement in the deep learning-based extraction of road information from remote sensing images. Subsequently, Ronneberger et al. [6] introduced U-Net, which notably improved segmentation accuracy by incorporating a symmetric encoder-decoder structure with skip connections, particularly excelling in the handling of fine-grained structures and boundaries. However, deep learning models are known to exhibit performance variability during testing due to differences in training datasets [7]. Consequently, whether existing deep learning models can maintain or improve their performance on new large-scale datasets remains an area of interest warranting further investigation.[8]

In response to this issue, this study utilizes the LoveDA remote sensing image semantic segmentation dataset, establishing independent training sets for urban and rural areas. The study trained and validated the performance of the FCN, U-Net, and SegNet models on these sets and examined their inference results on the test set. The models are deployed on a Nvidia RTX 4080s server, where the dataset is divided using k-fold cross-validation. All models are trained with identical parameters, and their Intersection over Union (IoU) ratios on the test set are compared, with the results analyzed in conjunction with the LoveDA dataset.

This study aims to compare the semantic segmentation performance of FCN, U-Net, and SegNet across various scenarios, specifically focusing on urban and rural datasets within the LoveDA dataset. The semantic information in these datasets may vary due to inherent geographical characteristics. By identifying performance discrepancies and correlating them with the unique characteristics of each model, this research seeks to offer insights into potential directions for improving semantic segmentation models when applied to remote sensing image training sets.

2 Data and methods

2.1 Dataset

The LoveDA dataset was introduced by Wang et al. in 2021 and encompasses a geographical area of approximately 536 square kilometres. It comprises aerial images from three different Chinese cities, categorized into seven land cover types. Each aerial image has a resolution of 1024×1024 pixels, with each pixel representing a real-world area of 0.3m×0.3m. Notably, the LoveDA dataset distinguishes itself from existing datasets by including two types of regions: Urban and Rural. In this study, the LoveDA training set will be utilized to train and evaluate three models: FCN, U-Net, and SegNet. Following established practices, approximately 70% of the dataset samples will be allocated for training, 10% for validation, and the remaining 20% for testing. The dataset comprises a total of 2,900 RGB images, with the specific data distribution ratios detailed in Table 1 [9].

Table 1. Overview of Training, Validation, and Test Set Splits in the LoveDA Dataset

Data Split	Domains		Proportion	Total
	Urban	Rural		
Train	1000	1000	68.97%	2000
Val	150	150	10.34%	300
Test	300	300	20.68%	600

Building upon previous research, this study adopts Intersection over Union (IoU) as the evaluation metric. IoU quantifies the overlap between predicted results and ground truth labels, rendering it a crucial metric for evaluating image segmentation performance. The calculation formula (1) is as follows: [10].

$$\mathrm{IoU} = \frac{|M_e \cap M_g|}{|M_e \cup M_g|} \tag{1}$$

In the formula, M_e represents the predicted result, while M_g denotes the ground truth label. The term $|M_e \cap M_g|$ refers to the intersection of pixels, which represents the correctly predicted pixels, whereas $|M_e \cup M_g|$ indicates the union of pixels, signifying the total area covered by both the predicted and actual pixels.

2.2 Methods

The experiment is conducted using the TensorFlow 2.8 deep learning framework for the construction, training, and testing of network models. The training process employs NVIDIA GeForce RTX 4080 GPUs with CUDA version 12.1. Input image and mask sizes are configured to 512x512, and the batch size is set to 8. The initial learning rate is established at 0.01, with cosine annealing implemented for learning rate decay. The Adam optimizer is utilized for updating the model weights. To mitigate overfitting, data augmentation techniques are employed to enhance the dataset before inputting it into the models, including random flipping, scaling, and colour space transformations. This approach increases data diversity and improves the generalization capability of the target models.

The experiment evaluates three classic semantic segmentation models: FCN, U-Net, and SegNet, which exhibit significant differences in their network structures and characteristics. FCN, introduced by Long et al. in 2015, was the first fully convolutional network. It replaces the fully connected layers found in traditional convolutional neural networks (CNNs) with convolutional layers, enabling the processing of input images into feature maps by the network. This model employs a multi-layer feature fusion approach, integrating shallow features with deep features, which results in a more cohesive combination of high-level semantic information and low-level spatial information, thereby enhancing segmentation accuracy and facilitating pixel-level predictions. Additionally, the incorporation of skip connections in FCN improves its capacity to capture edge details, leading to superior performance in segmenting fine structures [11].

U-Net, proposed by Ronneberger et al. in 2015, is extensively utilized in the domain of medical image segmentation. Its architecture features a symmetric encoder (downsampling path) and decoder (upsampling path). The encoder is responsible for extracting feature information from the image, while the decoder progressively restores the spatial resolution. Similar to Fully Convolutional Networks (FCN), U-Net incorporates skip connections between corresponding layers of the encoder and decoder. This design enables the direct transfer of the encoder's features to the decoder, thereby ensuring that critical spatial information is preserved during the upsampling process. The architecture of U-Net is particularly advantageous for accurate boundary segmentation and complex morphological

target segmentation, which accounts for its prevalent application in medical image segmentation tasks [12][14].

SegNet, introduced by Badrinarayanan et al. in 2015, incorporates an encoder architecture analogous to the VGG16 network, employing a sequence of convolutional and pooling operations to extract features. The decoder subsequently restores spatial resolution through progressive upsampling. A notable innovation of SegNet is its utilization of max-pooling indices to inform the decoder's upsampling process. This approach effectively preserves essential edge information while minimizing computational complexity, thereby enabling the model to sustain high segmentation accuracy. SegNet demonstrates strong performance in scene parsing and real-time applications [13][11].

In the LoveDA dataset, the Urban dataset is characterized by a multitude of labels and intricate road networks, whereas the Rural dataset typically presents open landscapes, single-label images, and roads that are often obscured by shadows. The three models—FCN, U-Net, and SegNet—each exhibit distinct network structures and characteristics. These structural differences result in varying performance levels on the Urban and Rural datasets. A comprehensive analysis of these differences is crucial for understanding the strengths and weaknesses of each model, as well as their suitability for different scenarios, which is one of the primary objectives of this experiment [9].

3 Results Analysis and Recommendations

3.1 Results Analysis

The experiments were conducted using an NVIDIA RTX 4080 computing server. To ensure a fair comparison, identical training parameters were established for all models. The experiments utilized the LoveDA dataset and involved the U-Net, SegNet, and FCN models. After 5000 epochs of training, the models were able to make predictions on the LoveDA test set. The Intersection over Union (IoU) results for each model on the split test sets are presented in Table 2.

Table 2. Performance Comparison of Semantic Segmentation Models on LoveDA Dataset

Models	Backbone	Rural mIoU	Urban mIoU	mIoU
FCN8s	VGG16	53.78%	55.34%	54.28%
SegNet	VGG16	**59.27%**	**58.03%**	**58.96%**
U-Net	ResNet50	58.17%	57.21%	57.80%

This section will further discuss the performance of each model, analyzing their strengths and weaknesses in terms of semantic segmentation based on the LoveDA dataset, with a particular emphasis on the differences in segmentation outcomes between the Urban and Rural test sets.

The data presented in Table 2 indicates that all three models—FCN, SegNet, and U-Net—exhibit fundamental capabilities in semantic segmentation. Among these, SegNet outperformed the others on the LoveDA dataset, achieving a mean Intersection over Union (mIoU) of 58.96%. Its remarkable spatial restoration ability is pivotal for detailed segmentation tasks involving remote sensing images. U-Net, utilizing the more robust ResNet50 as its backbone, attained a mIoU of 57.80%, which is slightly lower than that of

SegNet. Furthermore, the distinct characteristics of various scenes within the dataset significantly impacted model performance. In particular, the mixed training involving the Urban and Rural datasets highlighted the substantial differences between scenes and the multi-scale nature of the targets, which favoured SegNet's indexed upsampling technique, enhancing its adaptability to these datasets. Conversely, FCN8s, constrained by its earlier architecture and the limitations of VGG16, did not perform as well as the other two models.

When comparing the performance of the three models on the Rural and Urban test sets, FCN exhibited superior performance on the Urban test set. This is likely due to the Urban dataset containing more pronounced segmentation boundaries, such as those found in houses and buildings. In contrast, the Rural dataset presents simpler remote sensing imagery, where features like trees and vegetation complicate semantic interpretation. FCN, with its relatively simpler network architecture and reliance on multi-layer feature map fusion, faced challenges in accurately processing these scenes. In contrast, SegNet and U-Net, with their more intricate architectures, demonstrated more consistent performance across both the Rural and Urban datasets. Nevertheless, both models yielded better results on the Rural test set compared to the Urban test set, likely because the Urban dataset's detailed features, including roads and irregularly shaped buildings, posed additional challenges for the models in capturing finer details.

The performance differences among these models underscore the varying requirements of distinct remote sensing scenarios, with SegNet demonstrating the greatest versatility for mixed-scene segmentation.

To provide a more intuitive and clearer comparison of the segmentation performance across different experimental networks, the segmentation visualization results for each network are illustrated in Fig. 1.

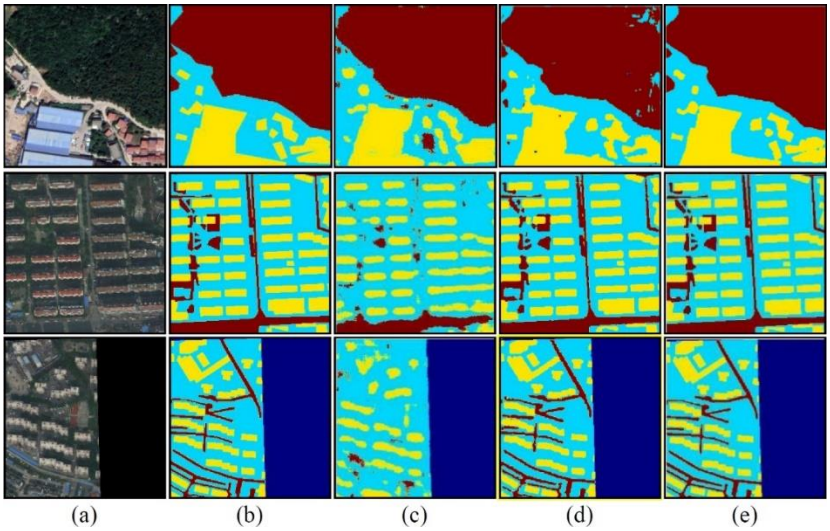

Fig. 1 Visual segmentation results in Urban datasets of LoveDA with different models. (a)RGB images. (b)Ground truth. (c)FCN. (d)U-Net. (e)SegNet (Photo/Picture credit : Original).

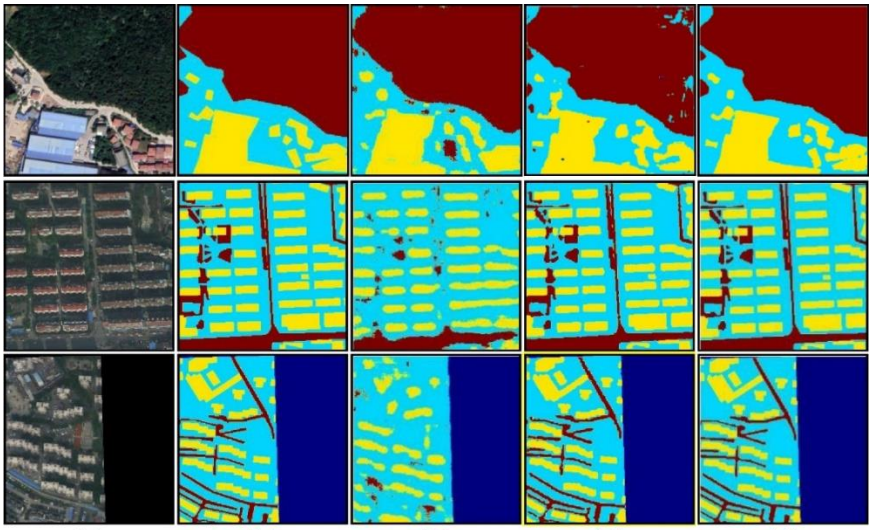

Fig. 2 Visual segmentation results in Rural datasets of LoveDA with different models. (a)RGB images. (b)Ground truth. (c)FCN. (d)U-Net. (e)SegNet(Photo/Picture credit : Original).

Fig.1 demonstrates the model's performance on the Urban test set, with Fig. 1(a) presenting the original true colour image, Fig. 1(b) displaying the segmentation labels, and Figs. 1(c), (d), and (e) depict the semantic segmentation results for FCN, U-Net, and SegNet, respectively. Similarly, Figure 2 showcases the performance of each model on the Rural test set. From Fig. 1, it can be observed that the Fully Convolutional Network (FCN) nearly accurately segments the houses in the Urban dataset; however, its understanding of road details is somewhat limited, as it primarily segments broader road areas. In contrast, U-Net and SegNet exhibit a more detailed segmentation of categories such as houses and roads within the Urban dataset. U-Net effectively preserves spatial detail information due to its 'encoder-decoder' architecture and skip connections, while SegNet enhances feature restoration accuracy by incorporating pooling indices into the decoding process. This improvement in feature restoration contributes to better segmentation performance of small targets, resulting in relatively superior outcomes in the Urban dataset. Nevertheless, despite the overall enhanced performance of U-Net and SegNet, both still face challenges with misclassification in complex urban environments, particularly in regions where building edges intersect with vegetation or roads.

Fig. 2 illustrates the performance of each model on the Rural test set, which primarily comprises expansive natural environments, including farmland and forests. From Fig. 2, it is clear that the segmentation performance of the Fully Convolutional Network (FCN) in the Rural dataset is inadequate, as it is significantly affected by the high-contrast vegetation and farmland in the background. The segmentation errors predominantly occur at the boundaries of complex vegetation, resulting in inaccurate segmentation outcomes for farmland, roads, and buildings. The FCN's global feature extraction capability is insufficient to capture the subtle category variations present in these intricate natural scenes. In contrast, the network architectures of U-Net and SegNet facilitate the effective retention of spatial semantic information, leading to more consistent segmentation performance in the Rural environment and demonstrating their high adaptability within this relatively straightforward labelled semantic context.

3.2 Suggestions for Model and Dataset Improvement

Based on the experimental results, this paper primarily presents suggestions for potential improvements to the U-Net and SegNet architectures, as well as considerations for enhancing the dataset to facilitate training. Although U-Net and SegNet outperform FCN in the scenarios examined in this study, there remains room for improvement within their existing network architectures.

For U-Net, it would be beneficial to enhance the "encoder-decoder" structure and skip connections to increase the model's depth, thereby improving its adaptability in scenes with complex labels. Regarding SegNet, while retaining the pooling index method during model improvements, future directions could involve refining the Decoder's SAME convolution component to enhance its learning capability during upsampling following pooling. Additionally, incorporating an adaptive feature extraction mechanism could enable the model to focus more precisely on the key areas of the segmentation targets.

Furthermore, improving the quality of training data and enhancing preprocessing methods are effective strategies for boosting the performance of semantic segmentation models. In addition to the data augmentation techniques employed in this study—such as random flipping, scaling, and colour domain transformations—elastic transformations and other data augmentation algorithms could be utilized to increase sample diversity. Methods such as histogram equalization, denoising, and contrast enhancement can improve issues related to image lighting and contrast, thereby reducing the impact on the model in complex environments. Additionally, exploring enhancements in the algorithms used for image augmentation may further mitigate the effects of complex environments on the original RGB images.

4 Conclusion

This study conducts a comparative analysis of the performance of three classic semantic segmentation models—FCN, UNet, and SegNet—on remote sensing image segmentation based on the LoveDA dataset. A comprehensive analysis of their performance in both urban and rural scenarios was conducted. The models were trained and tested on an NVIDIA RTX 4080s server using identical training parameters and optimization strategies, allowing for a comparative assessment of their segmentation performance across different contexts. The experimental results indicate that SegNet slightly outperforms the other models overall, particularly excelling in detail restoration and boundary handling. U-Net follows closely, benefiting from its skip connections and encoder-decoder architecture, which provide stability in complex scenarios. In contrast, FCN struggles with fine segmentation tasks due to its earlier architectural design, rendering it less effective in intricate environments.

This research highlights the performance differences among the models in urban and rural settings, with SegNet demonstrating robust performance in both contexts. Conversely, FCN's limited global feature extraction capabilities hinder its effectiveness in rural scenarios. Furthermore, this paper analyzes the strengths and weaknesses of the three models in various settings, particularly emphasizing the impact of network architecture on segmentation accuracy.

Based on the experimental results, this paper proposes potential avenues for future enhancements in both model architecture and dataset refinement. Further research could concentrate on optimizing the architectures of U-Net and SegNet to improve segmentation accuracy in complex environments, particularly in regions with intricate boundaries. Additionally, refining data augmentation techniques and enhancing image preprocessing methods will contribute to improving the models' generalization capabilities across diverse datasets. This study serves as a reference for future advancements in remote sensing image

segmentation models and holds practical significance for real-world applications.

References

1. A. Temenos, N. Temenos, M. Kaselimi, A. Doulamis, N. Doulamis, 2023. Interpretable Deep Learning Framework for Land Use and Land Cover Classification in Remote Sensing Using SHAP. IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, Art no. 8500105.
2. T. Loran, A. Barros Cardoso da Silva, S. K. Joshi, S. V. Baumgartner, G. Krieger, 2023. Ship Detection Based on Faster R-CNN Using Range-Compressed Airborne Radar Data. IEEE Geoscience and Remote Sensing Letters, vol. 20, pp. 1-5, Art no. 3500205.
3. I. Kotaridis, M. Lazaridou, 2021. Remote Sensing Image Segmentation Advances: A Meta-Analysis. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 173, pp. 309-322
4. F.Q.L. Feng, B.A. Chen, G.Q. Li, X.C. Yao, B.B. Gao, L.C. Zhang, 2022. A Review for Sample Datasets of Remote Sensing Imagery. National Remote Sensing Bulletin, vol. 26(4), pp. 589-605.
5. J. Long, E. Shelhamer, T. Darrell, 2015. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440. SCITEPRESS.
6. O. Ronneberger, P. Fischer, T. Brox, 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III. Springer International Publishing, pp. 234-241.
7. T. Heimann, et al., 2009. Comparison and Evaluation of Methods for Liver Segmentation from CT Datasets. IEEE Transactions on Medical Imaging, vol. 28(8), pp. 1251-1265.
8. J. Wang, Z. Zheng, A. Ma, et al., 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. arXiv preprint arXiv:2110.08733. SCITEPRESS.
9. T. Sun, et al., 2019. Leveraging Crowdsourced GPS Data for Road Extraction from Aerial Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. SCITEPRESS.
10. J. Long, E. Shelhamer, T. Darrell, 2015. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440. SCITEPRESS.
11. O. Ronneberger, P. Fischer, T. Brox, 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241. SCITEPRESS.
12. V. Badrinarayanan, A. Kendall, R. Cipolla, 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39(12), pp. 2481-2495.
13. B. Baheti, et al., 2020. Eff-U-Net: A Novel Architecture for Semantic Segmentation in Unstructured Environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
14. S.S. Agaian, K. Panetta, A.M. Grigoryan, 2001. Transform-Based Image Enhancement Algorithms with Performance Measure. IEEE Transactions on Image Processing, vol. 10(3), pp. 367-382.