Improving Road Extraction in Hyperspectral Data with Deep Learning Models

XuyingZhao[1[0009-0001-4193-3697]], ZhiboXing[2[0009-0009-1512-3963]], ZexiaoZou[3[0009-0008-9444-9816]], WuZhou[4[0009-0006-8107-4109]], ZhonghuiBian[5[0009-0003-7099-1292]] and XiaodongLi[6[0000-0002-5191-6217]]

[1] Beijing Electronic Science and Technology Institute, Fufeng Road. 7, Fengtai District, 100070, Beijing, China
yundingyi@163.com

Abstract. Extracting road networks accurately from hyperspectral data using traditional CNN models is challenging due to occlusion, changing lighting conditions, and spectral ambiguity. This study proposes a new model that combines the strengths of the U-net and Transformer architectures to capture both local and long-range features for improved road extraction accuracy and efficiency. The proposed model was evaluated on the AeroRIT hyperspectral dataset using performance metrics such as OA, MPCA, and MIOU, and compared with traditional CNN models such as U-net. The results show that the proposed model outperforms traditional models, demonstrating its potential for optimization and application in road extraction from hyperspectral data. The significance of this research lies in its potential to promote the development of the hyperspectral data analysis field, improve the accuracy of road extraction, and enable practical applications in various fields such as urban planning, traffic management, and environmental monitoring. Further research can optimize the proposed model or combine it with other methods to enhance its accuracy and efficiency for more effective road network extraction from hyperspectral data.

Keywords: Transformer-CNN, Hyperspectral Data, Road Extraction.

1 Introduction

Hyperspectral imaging provides detailed spectral information about materials in a scene, including roads. However, accurate extraction of roads from hyperspectral data is challenging due to their complex structure and similarity to other materials. Road extraction is a specific type of semantic segmentation task, and while deep learning models such as SegNet, U-net, and ResNet have shown promise, they have limitations such as loss of details [1], [2] or high computational requirements [3].To overcome these limitations, we propose a novel approach that combines Transformer and U-net architectures to capture global and local features, allowing long-range dependencies while maintaining fine-grained details. Few studies have explored this combination for hyperspectral image segmentation, making our proposed method an innovative one.

2 Related Work

2.1 Overview of Existing Hyperspectral Data Semantic Segmentation Methods

There are four main categories of methods for semantic segmentation of hyperspectral data: traditional machine learning, deep learning, hybrid and other methods.

Traditional machine learning methods, including support vector machines [4], and random forests [5]. They typically involve feature extraction and selection, followed by applying a classifier to the selected features. Deep learning methods, such as convolutional neural networks (CNNs) [6], and fully convolutional networks (FCNs) [7], have shown great promise in hyperspectral image segmentation. In addition, there are hybrid approaches [8] that combine traditional machine learning and deep learning methods. Other methods, such as clustering, graph-based methods [9], and active learning [10], have also been used for hyperspectral image segmentation.

2.2 Literature review of existing approaches for road extraction in hyperspectral data

Extracting roads from hyperspectral data can be divided into two main methods: traditional methods and deep learning-based methods [11]. Traditional methods utilize spectral features and statistical/mathematical models to extract roads [12], but have limitations in handling complex scenes and illumination changes. Deep learning-based methods, such as U-Net [13] and ResNet [14], have shown great potential for solving such problems, but require large amounts of annotated data and computational resources.

2.3 Overview of existing deep learning models for hyperspectral data analysis

Deep learning models have shown great potential in hyperspectral data analysis by automatically learning complex features from high-dimensional data.

U-Net and SegNet [15] are popular for hyperspectral image segmentation. U-Net captures both local and global features, but requires more data and computation. SegNet is computationally efficient and preserves spatial information, but may not perform as well on tasks requiring high-level context. ResNet is suitable for high-context tasks, but may not be as effective as U-Net or SegNet for hyperspectral segmentation. Transformers [16] achieve state-of-the-art results in language processing but may not be efficient for large amounts of hyperspectral data.

In our proposed method, we combine Transformer and U-net architectures to process hyperspectral images to capture global and local features, allowing long-range dependencies to be captured while maintaining fine-grained details.

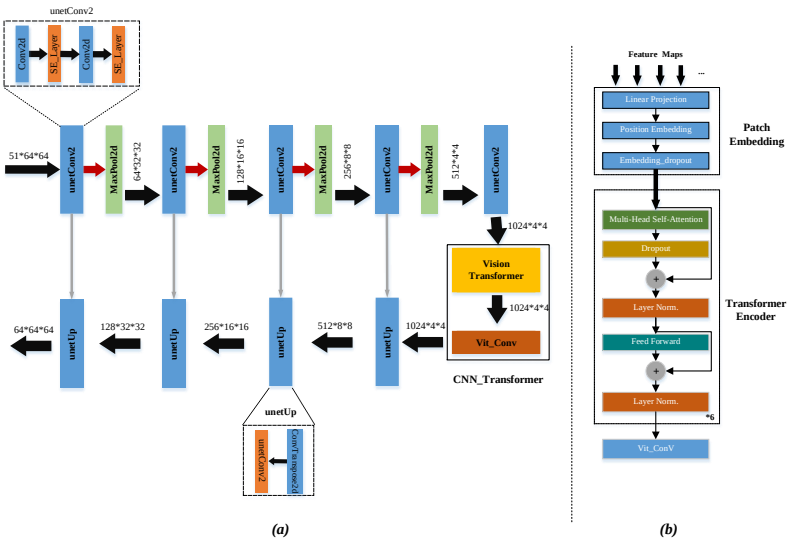

Fig. 1. A graphic description of our proposed method used in the paper.(a) CNNTrans,(b) ViT variant for our model In our proposed approach, we added a CNN-Transformer block at the center of the U-Net architecture. The CNN-Transformer block consists of a ViT model followed by a convolutional layer. The ViT model captures global context information, while the convolutional layer extracts local features.

3 Proposed Method

3.1 Description of the proposed approach for improving road extraction with deep learning models

Inspired by the Vision Transformer (ViT) model proposed by Dosovitskiy [17], for image classification, we modify the U-Net architecture by incorporating a CNN-Transformer block at the center of the network. This block utilizes the ViT encoder to extract positional embeddings from the output feature map of U-Net, which is then passed to the Transformer decoder to generate the final segmentation map. The decoder consists of multiple Transformer blocks that can perform nonlinear transformations on input features and capture long-range dependencies in feature maps. This architecture enables us to efficiently extract road features from hyperspectral data while reducing computational cost.

The CNN-Transformer block in the proposed approach can be represented by the following formula:

$$y=Concat(xcontract,CNN\text{-}Transformer(xcenter)) \quad (1)$$

where $x_{contract}$ is the feature map from the contracting path of the U-Net architecture, x_{center} is the feature map from the center of the U-Net architecture, and CNN-Transformer is the modified ViT model followed by a convolutional layer. The output

of the CNN-Transformer block is then concatenated with the corresponding features from the contracting path and fed into the expansive path.

The multi-head self-attention mechanism in the Transformer block of the ViT model can be represented by the following formula:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^o \quad (2)$$

where Q, K, and V are the query, key, and value matrices, respectively. h is the number of heads, and $head_i$ is the i-th head output. W^o is the output weight matrix used to combine the concatenated heads. In this model, using the original query (the content of the query), relative positional embeddings are used as keys, and values refer to the matrix of values obtained from the input sequence.

The feed-forward network in the Transformer block applies a nonlinear transformation to the input features and can be represented by the following formula:

$$FFN(x) = PReLU(xW_1+b_1)W_2+b_2 \quad (3)$$

where x is the input feature vector, W_1, b_1, W_2, and b_2 are the weights and biases of the two fully connected layers, and PReLU is the rectified linear unit activation function.

Overall, the proposed approach offers a promising solution to improve road extraction from hyperspectral data by leveraging the strengths of both CNN and Transformer-based networks.

3.2 Details on the Transformer and U-Net architectures used

We name the whole work CNNTrans. The input to the model is first normalized by spectral bands and resized to a fixed resolution of 64x64 pixels. The input is then subjected to four rounds of convolution and max pooling before being fed into the central convolutional layer of the U-Net architecture and output to the CNN-Transformer block.

The CNN-Transformer block processes the input image by dividing it into equal blocks, producing tensors of shape (batch_size, num_patches, output_channels). The output tensor is then reshaped to (batch_size, output-put_channels, num_patches_sqrt, num_patches_sqrt) and passed through a single convolutional layer before being used as input to the decoder part of the U-Net architecture. Here, batch_size is the number of samples in a batch, num_patches is the number of patches (or markers) extracted and processed by the transformer in the input image, output_channels is the number of channels in the output feature representation, num_patches_sqrt is the size height and width dimension, and passed upsampling outputs the final segmentation map.

The ViT part in the CNN-Transformer block captures global contextual information, while the convolutional layers extract local features. Since the final output of the ViT model is not the final result, we remove the classification head of the ViT model and add a convolutional layer at the end to reduce the feature dimension and computational cost. The output of the CNN-Transformer block is concatenated with the corresponding features in the contraction path and fed into the expansion path.

The Vit_conv layer performs a single convolution operation before feeding the output of the ViT model to the decoder part of the U-Net architecture. The forward method of the CNNTransformer class processes the input image through the convolutional layers, ViT model, SignleConv layer and U-Net architecture to produce the final segmentation map.

4 Dataset and Experiments

4.1 Description of the dataset used for evaluation

To evaluate the proposed CNN-Transformer based U-Net architecture for road extraction from hyperspectral data, we use the publicly available AeroRIT dataset [18] with a spatial resolution of 1973 × 3975 pixels covering the spectral range 397nm-1003nm, The step size is 1 nm. The dataset contains 12 spectral bands, which provide important information for distinguishing different materials and surfaces in the scene. Furthermore, this dataset includes ground truth labels of road pixels, which are manually annotated and considered accurate. Due to the complexity of the scene and the presence of other objects and materials, the AeroRIT dataset is a challenging benchmark dataset for hyperspectral image analysis, making it suitable for evaluating the performance of our proposed architecture.

4.2 Description of the experiments conducted to evaluate the proposed approach

To evaluate the performance of the proposed CNN-Transformer based U-Net architecture on road extraction, we conduct a series of experiments on the AeroRIT dataset. The dataset was randomly split into 80% for training and 20% for testing, ensuring that no overlapping patches were used in the two groups.

To train the model, we used several parameters to ensure the best results. The PreLU [19] activation function is used and the mini-batch size is set to 100. The learning rate is set to 1e-4, and the cross-entropy loss function of the Adam optimizer is applied. The number of spectral bands used in the analysis will depend on the setting, with 5 options ranging from 3 (RGB) to 51 (all). Finally, the model is trained for 60 epochs on a single NVIDIA GTX 3080 GPU with 10GB of memory.

We evaluated the proposed method using three metrics: Overall Accuracy (OA) [20], Mean PerClass Accuracy (MPCA), and Mean Jaccard Index (MIOU) [21]. OA and MPCA report the percentage of correctly classified pixels, while MIOU is used to mitigate dataset bias when class representations are small. Among these metrics, we adopt MIOU as the main metric for performance evaluation, because it measures the overlap between predicted masks and ground truth masks for all classes, which is a good way to evaluate the same degree of predicted segmentation maps and ground truth labels.

We compare the performance of the proposed CNN-Transformer based U-Net architecture with the baseline U-Net architecture and other state-of-the-art road extraction methods. The experimental results are presented in the next section.

4.3 Discussion of the experimental results

Table 1. Performance of various models on the AeroRIT test set.

Model	Overall acc.(OA) ↑	MPCA↑	MIOU↑
Segnet	87.2	66.7	53.5
U-net	89.3	69.2	57.4
ResNet	89.1	68.6	58.6
CNNTrans	**90.4**	**87.4**	**74.7**

Our experimental results show that the proposed CNNTrans model outperforms other models in all aspects, among which MIOU is the most important reference index for this segmentation task. Specifically, CNNTrans has a MIOU of 74.7, which is 16.1 percentage points higher than the second-best model, Resnet, and 21.2 percentage points higher than the worst-performing model, Segnet. We also observe that all models achieve overall accuracy (OA) scores, with CNNTrans achieving the highest score of 90.4%, followed by ResNet (89.1%), U-net (89.3%) and Se-gnet (87.2%).

5 Discussion

5.1 Comparison of the proposed approach with existing methods

In this study, we compare the proposed method with several existing methods, including Segnet, U-Net, ResNet. Experimental results show that all models are able to correctly classify most of the pixels, but the proposed CNNTrans model achieves the highest overall performance by using CNN-Transformer block to consider global and local information.

Overall, our experimental results demonstrate the effectiveness of the proposed CNNTrans model for the segmentation task on this dataset. However, further research is needed to investigate the robustness and generalization ability of the proposed method to other datasets and spectral ranges.

5.2 Discussion of the strengths and weaknesses of the proposed approach

Compared with existing methods such as Segnet, U-net, ResNet,, our proposed new method CNNTrans for hyperspectral image segmentation using Transformer network combined with U-Net architecture shows promising results in terms of accuracy and performance.

One of the strengths of this approach is that it can efficiently process high-dimensional hyperspectral data and capture both spatial and spectral information. The combination of Transformer and U-Net architecture enables us to efficiently process input data and capture relevant features for accurate segmentation.

However, our proposed method has some limitations. One of the main limitations is its high computational cost, which may limit its practical use in real-time applications. Furthermore, the proposed method requires a large amount of labeled data for training, which may not always be feasible in some applications.

Overall, our proposed method shows promising results and has potential for further optimization and improvement for more efficient and accurate hyperspectral image segmentation.

6 Conclusion and Future Work

The main contribution of this research is to provide a new solution for hyperspectral image segmentation. The method outperforms existing methods in terms of overall accuracy, MPCA, and MIOU, indicating that it can accurately segment hyperspectral images and has the potential to advance hyperspectral imaging applications. A limitation of this approach is its relatively high computational complexity.

Future work could extend the model to include more spectral bands or higher spatial resolution data and apply it to other types of aerial or satellite imagery for object detection and segmentation tasks.

7 Acknowledgement

This work was supported by“the Fundamental Research Funds for the Central Universities”(Grant Number：328202234)

References

1. Yang, R., Yu, J., Yin, J. et al. An FA-SegNet Image Segmentation Model Based on Fuzzy Attention and Its Application in Cardiac MRI Segmentation. Int J Comput Intell Syst 15, 24 (2022).
2. N. Siddique, S. Paheding, C. P. Elkin and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," in IEEE Access, vol. 9, pp. 82031-82057, 2021, doi: 10.1109/ACCESS.2021.3086020.
3. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
4. Gualtieri, J. Anthony and Robert F. Cromp. “Support vector machines for hyperspectral remote sensing classification.” Other Conferences (1999).
5. Ham, J., Chen, Y., Crawford, M. M., & Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing, 43(3), 492-501.
6. López, J., Torres, D., Santos, S., & Atzberger, C. (2020). Spectral imagery tensor decomposition for semantic segmentation of remote sensing data through fully convolutional networks. Remote Sensing, 12(3), 517.

7. Jiao, L., Liang, M., Chen, H., Yang, S., Liu, H., & Cao, X. (2017). Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 55(10), 5585-5599.
8. Roy, Swalpa Kumar et al. “HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification.” IEEE Geoscience and Remote Sensing Letters 17 (2019): 277-281.
9. Bandyopadhyay, Debmita and Subhadip Mukherjee. “Tree species classification from hyperspectral data using graph-regularized neural networks.” ArXiv abs/2208.08675 (2022): n. pag.
10. Lenczner, Gaston, et al. "DIAL: Deep Interactive and Active Learning for Semantic Segmentation in Remote Sensing." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022): 3376-3389.
11. Y. Chen, Z. Lin, X. Zhao, G. Wang and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 6, pp. 2094-2107, June 2014, doi: 10.1109/JSTARS.2014.2329330.
12. Patil, D., Jadhav, S. (2021). Road Extraction Techniques from Remote Sensing Images: A Review. In: Raj, J.S., Iliyasu, A.M., Bestak, R., Baig, Z.A. (eds) Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore. https://doi.org/10.1007/978-981-15-9651-3_55
13. Ronneberger, Olaf et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” ArXiv abs/1505.04597 (2015): n. pag.
14. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
15. V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481-2495, 1 Dec. 2017, doi: 10.1109/TPAMI.2016.2644615.
16. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017)..
17. Dosovitskiy, Alexey et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ArXiv abs/2010.11929 (2020): n. pag.
18. A. Rangnekar, N. Mokashi, E. J. Ientilucci, C. Kanan and M. J. Hoffman, "AeroRIT: A New Scene for Hyperspectral Image Analysis," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 11, pp. 8116-8124, Nov. 2020, doi: 10.1109/TGRS.2020.2987199.
19. He, Kaiming et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” 2015 IEEE International Conference on Computer Vision (ICCV) (2015): 1026-1034.
20. Wang, Z., Wang, E. & Zhu, Y. Image segmentation evaluation: a survey of methods. Artif Intell Rev 53, 5637–5674 (2020)
21. Costa, L. D. F. (2021). Further generalizations of the Jaccard index. arXiv preprint arXiv:2110.09619.