International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
Impact Factor (2024): 7.741

HSTNet: An Iterative Optimization Network for Semantic Segmentation of High-Resolution Remote Sensing Images

Tingkai Wang

North China electric Power University, School of Control and Computer Engineering, Beijing, 102206, China

Abstract: ***How to achieve accurate semantic segmentation of high-resolution remote sensing images is a current focal point in image semantic segmentation tasks. However, the information contained in high-resolution images is typically complex, and due to the large size of the images, they are constrained by the receptive field size of convolutional networks, making accurate semantic segmentation challenging. Significant errors exist in both local edge and overall image segmentation results. This paper presents HSTNet, a semantic segmentation network for high-resolution remote sensing images with an iterative structure. HSTNet adopts an encoder-decoder architecture similar to Unet. In HSTNet, we employ Swin-Transformer modules to learn and correlate feature tensors at different scales, aiming to capture the overall structure of high-resolution images and associate long-range geographic information across the images as much as possible. Furthermore, we devised an iterative optimization framework that progressively enhances the semantic segmentation results of the network. We observed that preliminary semantic segmentation outputs can serve as cues to facilitate the network in achieving more accurate segmentation. These initial semantic segmentation results encapsulate relationships among various semantic objects within different regions of the image, thereby reducing the cost of the network learning image features during subsequent iterations and assisting the network in achieving improved outcomes. We compared our approach with several state-of-the-art methods on the Potsdam dataset from ISPRS. The final results indicate that our method achieves outstanding performance.***

Keywords: Deep learning, Semantic segmentation, Image processing

1. Introduction

Semantic segmentation of remote sensing images aims to predict the semantic class for each pixel in the image. It has been a fundamental issue in remote sensing image processing and is an essential component of remote sensing image interpretation. With the rapid development of aerospace, remote sensing, and imaging technologies, it has become increasingly easier to acquire large quantities of high-quality, high-resolution remote sensing images. However, this also brings about a practical challenge: how to efficiently and accurately perform semantic segmentation on high-resolution remote sensing images.

Existing semantic segmentation methods primarily utilize deep learning networks to extract land cover information from images and segment pixels with different semantic meanings0. Deep learning algorithms, which hierarchically learn representative and distinctive features from data, have been introduced into the field of remote sensing and have rapidly developed. Spectral and textural information of remote sensing images are used as low-level features inputted into convolutional neural network models for pixel-based semantic segmentation, resulting in feature classification information. Compared to traditional optical remote sensing image segmentation, deep feature-based segmentation methods can leverage neural networks to implicitly establish pixel-to-semantic mapping relationships[2]. The network autonomously learns to extract target features from remote sensing images, completing the entire segmentation process without the intervention of manual feature engineering. This simultaneously enhances the accuracy of results and the generalization capability of the model[3]-[5].

Presently, the most successful and state-of-the-art semantic segmentation networks trace their origins back to a common ancestor: the Fully Convolutional Network (FCN)[6]. This seminal work replaced fully connected layers with convolutional networks to output spatial maps, which are then upsampled to generate predicted maps, thereby classifying and segmenting pixels with similar semantics in the image. Consequently, a plethora of advanced semantic segmentation networks have emerged in rapid succession.

However, when dealing with high-resolution remote sensing images, existing convolutional network methods have demonstrated their inherent limitations. The limited receptive fields of filters in convolutional networks make it difficult to effectively associate long-range relationships among pixels with similar semantics, which may lead to either over-segmentation or under-segmentation issues in images. While some methods leverage multi-scale contextual information to enhance the segmentation performance of convolutional networks, such as the stacked hourglass module proposed by Li et al. [7], which learns contextual information from different scales and extracts rich multi-scale features through intermediate supervision, and the initial module composed by Liu et al. [8], replacing common convolutional layers to provide the network with multi-scale receptive fields for acquiring multi-scale information, the collection of multi-scale information from images may lead to the loss of some local fine-grained details, thereby reducing the accuracy of image segmentation results, The inherent limitations of convolutional networks constrain the performance of semantic segmentation networks. Additionally, most existing image semantic networks often attempt to establish a direct mapping between input images and semantic segmentation results. This approach may lead the network to overlook some important underlying semantic features in the images.

This paper proposes a semantic segmentation network for

high-resolution remote sensing images with an iterative structure, leveraging Swin-Transformer modules to establish global semantic correlations of input images at different scales. This effectively enhances the accuracy of image partitioning. Further elaboration will be provided in the following paragraph.

The Transformer[9] has achieved tremendous success in both image processing and natural language processing domains, revolutionizing many tasks in these fields. It introduces attention mechanism, which enables the model to better capture relationships between different positions by allowing interactions between all positions in the input sequence. This attention mechanism enables Transformers to more effectively handle long-range dependencies, thus enhancing the model's performance and generalization capability in both natural language processing and computer vision tasks. Through self-attention mechanism, the Transformer can model and process sequential data without introducing recurrent structures, greatly accelerating both model training and inference processes. On the other hand, the existing classical image segmentation network, Unet[10], is capable of integrating low-resolution and high-resolution information. It learns and correlates image features at different scales, consequently obtaining fine-grained image segmentation results, making it particularly suitable for medical image segmentation tasks. Integrating Transformer modules at various scales within Unet can further enhance Unet's capability to incorporate long-range structural information across different scale feature maps.

In the task of image shadow detection, Patel et al.[11] utilized an image shadow removal network as a pre-processing stage to provide shadow distribution features in the image for the shadow detection network. This significantly enhanced the accuracy of shadow detection. The image shadow removal task often serves as a subsequent stage to the image shadow detection task. These two tasks are inherently correlated and similar, as both require learning from shadow-free and shadowed regions. Their work demonstrates that two networks handling related tasks can mutually provide useful information to each other, thereby promoting each other's performance.

Inspired by the aforementioned work, we propose an iterative semantic segmentation network for high-resolution remote sensing images. We introduce Swin-Transformer modules at different scales within Unet to learn and correlate long-range semantic features, aiming to enhance the network's ability to learn pixel correlations in high-resolution images. Furthermore, we employ the predicted results as input information for the second round of iteration, along with the original image, into our network for progressive optimization. This approach aims to utilize preliminary semantic segmentation results to guide the network in learning the relationships between pixels with similar semantic information at different locations. Throughout the iterative process, it reinforces the degree of association between pixels with identical semantic information and potentially corrects errors that may arise during the upsampling process, thereby progressively enhancing the quality of semantic segmentation results.

Our main contributions include:

- We design a novel semantic segmentation network, named HSTNet. This network is capable of learning and establishing long-range semantic pixel correlations within input images at different scales, addressing the limitation of existing convolutional networks in capturing global image features due to restricted receptive fields.
- We propose an iterative framework where the result of each segmentation iteration is fed back into the network to guide more accurate segmentation of the image. This structure enhances the network's ability to learn relationships between pixels with the same semantic information, progressively improving semantic segmentation results and potentially correcting errors in the previous predictions.
- Experimental results on the ISPRS Potsdam dataset indicate that further extracting multi-scale features from the encoder output and aggregating them in the decoder can enhance the segmentation performance of the network model.

2. Related Work

With the advancement of artificial intelligence and deep learning, segmentation methods based on deep features are gradually transitioning to optical remote sensing images. These methods are being enhanced based on features such as multispectral data and intra-class dissimilarity. Among these enhancements, segmentation methods based on segmentation models stand out as the most prominent. These models achieve image segmentation by training neural network classifiers for classification.

In 2015, the Fully Convolutional Network semantic segmentation model[6] was proposed, achieving pixel-level image semantic segmentation. It replaces the fully connected layers used for classification mapping in CNN structures with convolutional layers, and combines the information from intermediate pooling layers to generate image prediction segmentation maps. Additionally, the UNet [10] architecture exhibits more accurate segmentation performance with limited training data and has been widely applied in remote sensing image segmentation tasks. DeconvNet[12] also adopts a similar encoder-decoder architecture to upsample the image to its original size, utilizing deconvolution layers to densify sparse feature maps during upsampling instead of pooling operations. Dilated convolutions are also commonly used to alleviate the conflict between feature map size and receptive field size.

3. Methods

3.1 Net structure

We constructed HSTNet as depicted in Figure 1. While ResNet has achieved outstanding performance in image classification tasks due to its extremely deep network architecture and powerful feature representation [12], introducing numerous ResNet blocks into the network yielded limited performance improvements, and significant deficiencies persist when handling complex HRRSI segmentation tasks.

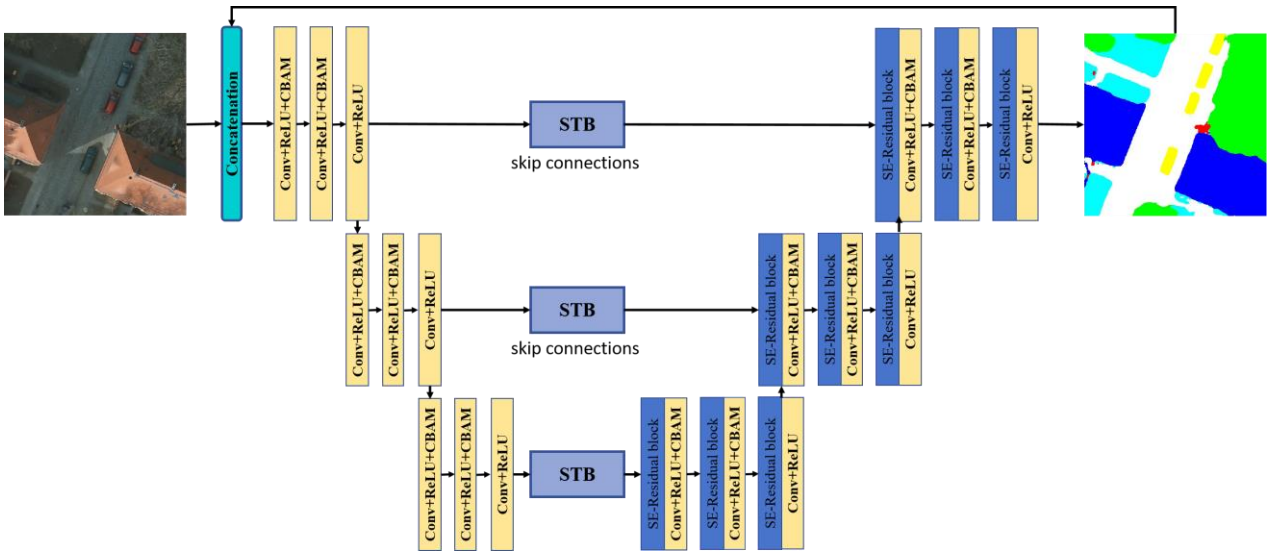

Figure 1: Our network architecture (HSTNet)

Therefore, we replaced the original fully connected layers of ResNet34 with SE-Res Blocks, placing them within the upsampling process of Unet to adaptively adjust the feature weights of input tensors. We made adjustments to the existing SE module, which consists of two convolutional layers, one pooling layer, two fully connected layers, one ReLU function, and one Sigmoid function, to further enhance the network's capability to extract image features. The primary function of the SE residual block is to enhance and optimize salient local regions in different scenes by explicitly modeling the interrelationships between convolutional feature channels. It utilizes an adaptive mechanism to learn and adjust the weights of different feature channels, thereby better guiding the convolutional network in learning local regions and enhancing features, thus aiding in removing structural information from non-target scenes. The SE residual block dynamically adjusts the weights of feature channels, enabling the network to focus more on the features that are more important for the current task, thus improving the performance and effectiveness of the model.

Similar to Unet, we establish skip connections between the downsampling and upsampling parts at the same scale, but the skip connections are processed using a network primarily based on Swin-Transformer. The purpose of this approach is to enhance the network's ability to learn long-range semantic features. The STB module employs window-based multi-head attention and window sliding connection mechanisms to learn and enhance structural features of the entire scene. Compared to learning and enhancing structural features of the same scene through computing multi-head global attention for the entire image, this approach significantly reduces computational complexity. Due to the faster growth rate of computational complexity for global attention in images compared to natural language, the long-range feature mapping and complexity of images increase quadratically with image size. Traditional image Transformer techniques, limited by the growth rate of computational complexity, can only handle low-resolution images. Therefore, when dealing with excessively large images, traditional image Transformer backbone networks become unsuitable.

We stack the semantic segmentation results of the image with the input image and reapply them as input for iterative optimization. Introducing the semantic segmentation results from the previous iteration aids the network in more rapidly learning the correlations between pixels with the same semantic meaning, thereby enhancing the network's convergence speed and learning accuracy.

3.2 Swin-Transformer block and Skip Connection

The structure of the Swin-Transformer module we employ is illustrated in Figure 2. Firstly, the feature tensor is cropped and segmented into 8x8 small windows, with each window undergoing feature mapping. Subsequently, the feature maps of each window are normalized and fed into the WMSA module. The WMSA module, which stands for Window-based Multi-headed Self-Attention Module, computes multi-headed attention for the feature maps within each small window, thereby emphasizing the same scene structural features within each window. Next, the feature maps of adjacent windows are connected across windows to compute the feature correlations between windows. In this chapter, a total of three layers of STB modules (Swin-Transformer Blocks[13]) are stacked, downsampling the input feature maps by 4x, 8x, and 16x, respectively. This module gradually establishes the correlation weights between global features by moving windows, enhancing the network's ability to learn same-semantic features across large regions in high-resolution images.

4. Experiment

To validate the competitive performance and model generalization capability of our proposed iterative high-resolution remote sensing image semantic segmentation network, HSTNet, in high-resolution image semantic segmentation tasks, we compared it with other methods using the ISPRS Potsdam dataset ("2D semantic Labeling - Potsdam," n.d.) and selected multiple images from this dataset as test data.

Table 1: Quantitative Comparison Results

Model	MIoU(%)		mPA(%)	
	Vaihingen	Potsdam	Vaihingen	Potsdam
UNet	74.88	72.06	85.65	82.72
Deeplabv3+	77.51	71.05	88.24	81.30
PSPNet	72.24	68.76	84.73	79.24
HRNetV2	78.24	73.33	89.21	83.29
HSTNet	**85.36**	**78.52**	**94.15**	**98.67**

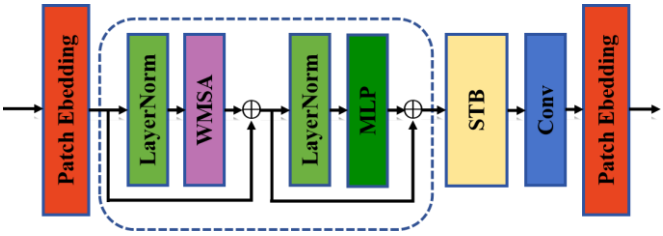

Figure 2: STB block of Skip Connection

4.1 Implementation details

We designed HSTNet based on the PyTorch 1.8.0 framework and conducted network model training on a GPU server: comprising 1 CPU (Intel Xeon E5-2640 v4) with 128GB RAM and 1 GPU (NVIDIA Tesla 3090 24GB) with 24GB VRAM. The main parameter settings include: adjusting the batch size according to different network configurations to ensure maximum memory utilization, using 40 threads, setting the initial learning rate to 1e-4, employing the ReduceLROnPlateau dynamic learning rate adjustment strategy, utilizing the AdamW optimizer (Loshchilov and Hutter, 2018), selecting the cross-entropy loss function for multi-class labeled datasets, and conducting training for 100 epochs. In our experiments, normalization parameters—mean and standard deviation for each channel—were not set to default values used in digital image semantic segmentation but were pre-calculated based on the corresponding multispectral remote sensing image dataset.

We primarily evaluate the performance of our work in terms of segmentation accuracy. To conduct an objective and fair cross-sectional comparison with other proposed network models, we employ two common evaluation metrics: MIoU (Mean Intersection over Union) and mPA (Mean Pixel Accuracy). Their respective calculation formulas are as follows:

$$\mathrm{MIoU} = \frac{1}{\mathrm{k+1}} \sum_{\mathrm{i=0}}^{k} \frac{P_{ii}}{\sum_{i=0}^{k} P_{ij} + \sum_{i=0}^{k} P_{ji} - P_{ii}} \tag{1}$$

P_{ij} represents the prediction of i as j, denoted as False Negative prediction (FN); P_{ji} represents the prediction of j as i, denoted as False Positive prediction (FP); P_{ii} represents the prediction of i as i, denoted as True Positive prediction (TP).

When computing the mPA of the predicted results, we first calculate the pixel accuracy (PA) for each class. PA represents the proportion of pixels correctly classified to the total number of pixels in that class. Specifically, for class i, the formula for pixel accuracy (PAi) is as follows:

$$PA_i = TP_i / (TP_i + FP_i) \tag{2}$$

Here, TP_i represents the number of pixels correctly classified as class i, and FP_i represents the number of pixels incorrectly classified as class i. Next, the average pixel accuracy of all classes, mPA, is calculated. The formula for mPA is as follows:

$$mPA = (PA1 + PA2 + \ldots + PAn)/n \tag{3}$$

n represents the total number of classes. By computing the average pixel accuracy for all classes, mPA provides a comprehensive assessment to gauge the overall performance of the model in image semantic segmentation tasks.

4.2 Quantitative comparison

Our experimental results, as shown in Table 1, indicate that HSTNet achieved the best MIoU results on both datasets. HSTNet outperformed HRNetV2 by 6.27% in MIoU and by 2.98% in mPA on the Vaihingen dataset. Similarly, on the Potsdam dataset, HSTNet surpassed HRNetV2 by 4.72% in MIoU and by 5.14% in mPA. Overall, HSTNet exhibits a significant advantage in accuracy performance.

4.3 Qualitative experiments

Furthermore, the semantic segmentation results of the HSTNet model on the Vaihingen and Potsdam datasets were visualized. As shown in Figure 3, during the Vaihingen dataset testing, conventional models such as HSTNet often struggle to accurately differentiate between fallow land and railway embankments, tending to overlook subtle depressions in railway embankments. This oversight can result in gaps when identifying vegetative and housing areas, revealing limitations in precise edge detection and complex landscape recognition. Similar situations arise in the Potsdam dataset, particularly in the identification of low-lying vegetation and impervious surfaces, where these models frequently omit significant details, especially when dealing with areas of complex texture or similar colors. In contrast, the HSTNet model demonstrates its unique advantages. Leveraging the potent spatial relationship encoding capability of Transformers, it effectively enhances the prediction accuracy for these challenging-to-distinguish regions, particularly in identifying areas with minute and intricate textures. HSTNet exhibits outstanding performance on both the Vaihingen and Potsdam datasets, particularly in accurately delineating edges and recognizing complex terrains, surpassing conventional models significantly, thus strongly validating the effectiveness and advancement of our approach.

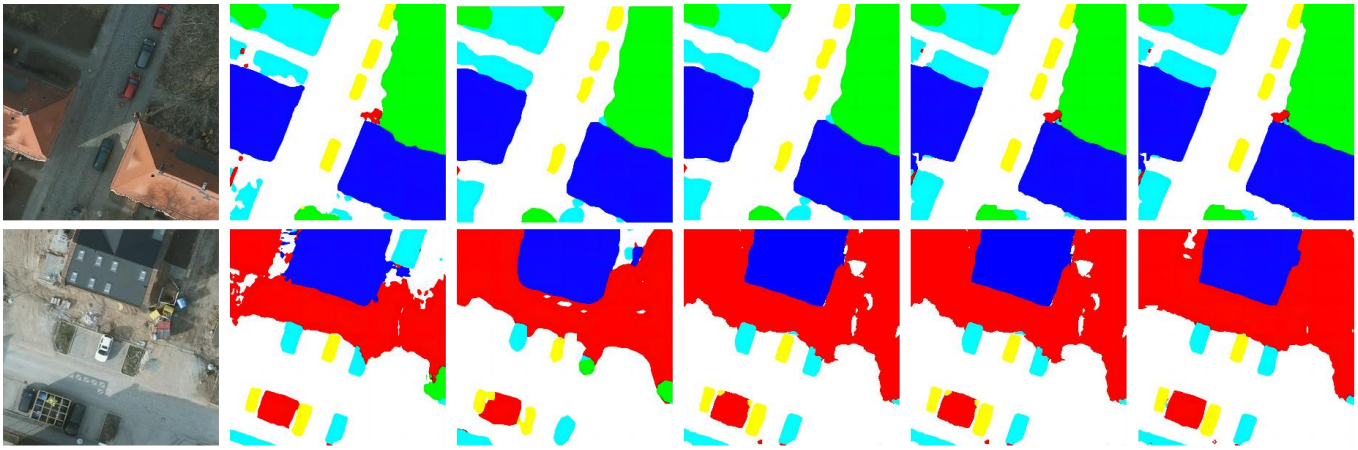

Figure 3: Qualitative Comparative Experimental Results

5. Equations

In this paper, we analyze the shortcomings of existing convolutional neural networks in handling semantic segmentation tasks on high-resolution images. Addressing these limitations, we propose a novel model called HSTNet based on Unet. We introduce the STB module into the skip-connection part of the network to enhance its capability to learn global semantic features. Additionally, we incorporate an iterative framework where the semantic segmentation results from the network are fed back as clues into our proposed model. This architecture allows the network to progressively improve the semantic segmentation results and potentially correct errors, thereby enhancing the accuracy of local information. Our HSTNet achieves excellent performance on the ISPRS Vaihingen and Potsdam datasets.

References

[1] Liu Y, Li H, Hu C, et al. Learning to aggregate multi-scale context for instance segmentation in remote sensing images[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024.
[2] Schuegraf P, Bittner K. Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN[J]. ISPRS International Journal of Geo-Information, 2019, 8(4): 191.
[3] Maggiori E, Tarabalka Y, Charpiat G, et al. High-resolution aerial image labeling with convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(12): 7092-7103.
[4] Mou L, Hua Y, Zhu X X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12416-12425.
[5] Liu Y, Fan B, Wang L, et al. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 145: 78-95.
[6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
[7] Li P, Lin Y, Schultz-Fellenz E. Contextual hourglass network for semantic segmentation of high resolution aerial imagery[J]. arXiv preprint arXiv:1810.12813, 2018.
[8] Liu Y, Minh Nguyen D, Deligiannis N, et al. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery[J]. Remote Sensing, 2017, 9(6): 522.
[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[10] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015: 234-241.
[11] Valanarasu J M J, Patel V M. Fine-context shadow detection using shadow removal[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 1705-1714.
[12] Trias-Sanz R, Stamon G, Louchet J. Using colour, texture, and hierarchial segmentation for high-resolution remote sensing[J]. ISPRS Journal of Photogrammetry and remote sensing, 2008, 63(2): 156-168.
[13] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
[14] Sun K, Zhao Y, Jiang B, et al. High-resolution representations for labeling pixels and regions[J]. arXiv preprint arXiv:1904.04514, 2019.
[15] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
[16] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.