Semantic Segmentation of Satellite Images with Deep Learning Method

Weili Shi, Daiqing Qi

1.Introduction

Remote sensing is the process of detecting and monitoring the physical characteristics of an area by measuring its reflected and emitted radiation at a distance (typically from satellite or aircraft). Special cameras collect remotely sensed images, which help researchers "sense" things about the Earth. Some examples are:

- Cameras on satellites and airplanes take images of large areas on the Earth's surface, allowing us to see much more than we can see when standing on the ground.
- Sonar systems on ships can be used to create images of the ocean floor without needing to travel to the bottom of the ocean.
- Cameras on satellites can be used to make images of temperature changes in the oceans.

One outcome of the remote sensing is the satellite images, It is generated by processing the visible light reflected from the surface of the earth. The satellite images usually contain abundant spatial details and rich potential semantic contents for analysis.

Recently, a growing wave of deep learning technology has achieved great success in computer vision tasks such as image classification, object detection and semantic segmentation. Semantic segmentation of remotely sensed urban scene images is required in a wide range of practical applications, such as land cover mapping, urban change detection, environmental protection, and economic assessment. Driven by rapid developments in deep learning technologies. Up to now, there are two competing backbones models in computer vision: CNN and vision transformer. The convolutional neural network (CNN) has dominated semantic segmentation for many years. CNN adopts hierarchical feature representation, demonstrating strong capabilities for information extraction. However, the local property of the convolution layer limits the network from capturing the global context. Recently, as a hot topic in the domain of computer vision, Transformer has demonstrated its great potential in global information modeling, boosting many vision-related tasks such as image classification, object detection, and particularly semantic segmentation.

In our project we investigate how CNN-based and transformer-based encoders affect the performance of the semantic segmentation model on the satellite images. Specifically, we choose ResNet-18 for CNN-based encoder and Swin Transformer for transformer-based encoder.

2. Related Work

2.1. CNN-based semantic segmentation methods

The fully convolutional network (FCN) is the first effective CNN structure to address

semantic segmentation problems in an end-to-end manner. Since then, CNN-based methods have dominated the semantic segmentation task in the remote sensing field. However, the over- simplified decoder of FCN leads to a coarse-resolution segmentation, limiting the fidelity and accuracy. To address this problem, an encoder-decoder network, i.e., the UNet, was proposed for semantic segmentation, with two symmetric paths named the contracting path and the expanding path (Ronneberger et al., 2015). The contracting path extracts hierarchical features by gradually downsampling the spatial resolution of the feature maps, while the expanding path learns more contextual information by progressively restoring the spatial resolution. Subsequently, the encoder-decoder framework has become the standard structure of remote sensing image segmentation networks (Badrinarayanan et al., 2017; Chen et al., 2018a; Sun et al., 2019). Based on encoder-decoder structure, (Diakogiannis et al., 2020; Yue et al., 2019; Zhou et al., 2018) designed different skip connections to capture more abundant context, while (Liu et al., 2018; Zhao et al., 2017b; Shen et al., 2019) developed various decoders to retain semantic information. The encoder-decoder CNN-based methods, although have achieved encouraging performance, encounter bottlenecks in urban scene inter- pretation (Sherrah, 2016; Marmanis et al., 2018; Nogueira et al., 2019). To be specific, CNN-based segmentation networks with limited receptive fields can only extract local semantic features and lack the capability to model the global information from the whole image. However, within fine-resolution remotely sensed urban scene images, complicated pat- terns and human-made objects occur frequently (Kampffmeyer et al., 2016; Marcos et al., 2018; Audebert et al., 2018). It is difficult to identify these complex objects if only relying on the local information.

2.2. Global contextual information modeling

To liberate the network from the local pattern focus of CNNs, many attempts have been conducted to model global contextual information, while the most popular way is incorporating attention mechanisms into networks. For example, Wang et al. modified the dot-product self- attention mechanism and applied it to computer vision domains (Wang et al., 2018). Fu et al. appended two types of attention modules on top of a dilated FCN to adaptively integrate local features with their global dependencies (Fu et al., 2019). Huang et al. proposed a criss-cross attention block to aggregate informative global features (Huang et al., 2020). Yuan et al. developed an object context block to explore object- based global relations (Yuan et al., 2020). Attention mechanisms also improve the performance of remote sensing image segmentation networks. Yang et al. proposed an attention-fused network to fuse high-level and low-level semantic features and obtain state-of-the-art results in the semantic segmentation of fine-resolution remote sensing images (Yang et al., 2021b). Li et al. in- tegrated lightweight spatial and channel attention modules to refine semantic features adaptively for high-resolution remotely sensed image segmentation (Li et al., 2020a). Ding et al. designed a local attention block with an embedding module to capture richer contextual information (Ding et al., 2021). Li et al. developed a linear attention mechanism to reduce the computational complexity while improving performance (Li et al., 2021a). However, the above attention modules restrict the global feature representation due to over-reliance on convolutional operations. Furthermore, a single attention module cannot model the global information at multi-level semantic features in the decoder.

2.3. Transformer-based semantic segmentation methods

Recently, several attempts were made to apply the Transformer for global

information extraction (Vaswani et al., 2017). Different from the CNN structure, the Transformer translates 2D image-based tasks into 1D sequence-based tasks. Due to the powerful sequence-to-sequence modelling ability, the Transformer demonstrates superior characteriza- tion of extracting global context than the above-mentioned attention- alone models and obtains state-of-the-art results on fundamental vision tasks, such as image classification (Dosovitskiy et al., 2020), object detection (Zhu et al., 2020) and semantic segmentation (Zheng et al., 2021). Driven by this, many researchers in the remote sensing field have applied the Transformer for remote sensing image scene classification (Bazi et al., 2021; Deng et al., 2021), hyperspectral image classification (He et al., 2021; Hong et al., 2021), object detection (Li et al., 2022a), change detection (Chen et al., 2021a), building and road extraction (Chen et al., 2021c; Sun et al., 2022), and especially semantic segmen- tation (Wang et al., 2021b, 2022). Most of the existing Transformers for semantic segmentation still follow the encoder-decoder framework. According to different encoder- decoder combinations, they can be divided into two categories. The first is constructed by a Transformer-based encoder and a Transformer-based decoder, namely the pure Transformer structure. Typical models include the Segmenter (Strudel et al., 2021), SegFormer (Xie et al., 2021) and SwinUNet (Cao et al., 2021). The second adopts a hybrid structure, which is composed of a Transformer-based encoder and a CNN-based decoder. Transformer-based semantic segmentation methods commonly follow the second structure. For example, the TransUNet employed the hybrid vision Transformer (Dosovitskiy et al., 2020) as the encoder for stronger feature extraction and obtains state-of-the-art re- sults in medical image segmentation (Chen et al., 2021b). The DC-Swin introduced Swin Transformer (Liu et al., 2021) as the encoder and de- signs a densely connected convolutional decoder for fine-resolution remote sensing image segmentation, surpassing the CNN-based methods by a large gap (Wang et al., 2022). (Panboonyuen et al., 2021) also selected the Swin Transformer as the encoder and utilizes various CNN-based decoders, such as UNet (Ronneberger et al., 2015), FPN (Kirillov et al., 2019) and PSP (Zhao et al., 2017a), for semantic segmentation of remotely sensed images, obtaining advanced accuracy. Despite the above advantages, the computational complexity of the Transformer-based encoder is much higher than the CNN-based encoder due to its square-complexity self-attention mechanism (Vaswani et al., 2017), which seriously affects its potential and feasibility for urban- related real-time applications. Thus, to fully harness the global context extraction ability of Transformers without resulting in high computational complexity, in this paper, we present a UNet-like Transformer with a CNN-based encoder and a Transformer-based decoder for efficient semantic segmentation of remotely sensed urban scene images. Specifically, for our UNetFormer, we select the light- weight backbone, i.e. ResNet18, as the encoder and develops an efficient global–local attention mechanism to construct Transformer blocks in the decoder. The proposed efficient global–local attention mechanism adopts a dual-branch structure, i.e. a global branch and a local branch. Such a structure allows the attention block to capture both global and local contexts, thereby surpassing the single-branch efficient attention mechanisms in Transformers that only capture global contexts (Liu et al., 2021; Zhang and Yang, 2021).

3. Method

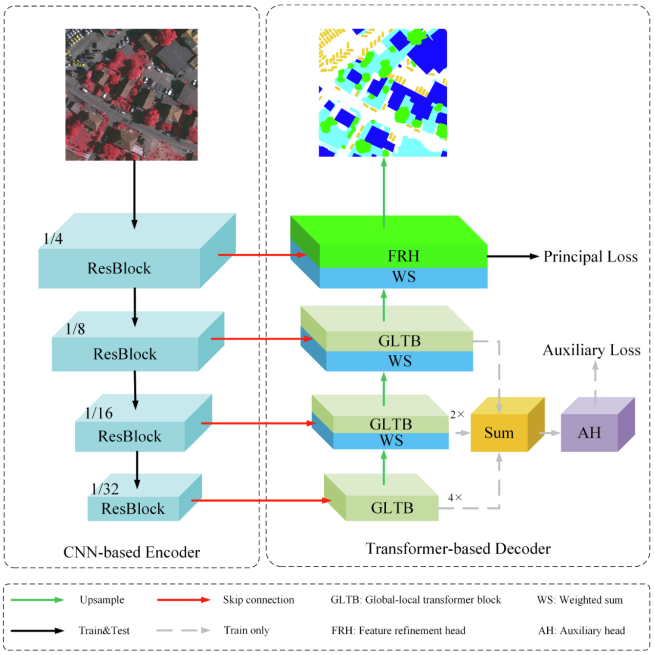

Figure 1. Illustration of the Unetformer model.

As illustrated in the above figure, the proposed UNetFormer is constructed using a CNN-based encoder and a Transformer-based decoder. A detailed description of each component is given in the following sections.

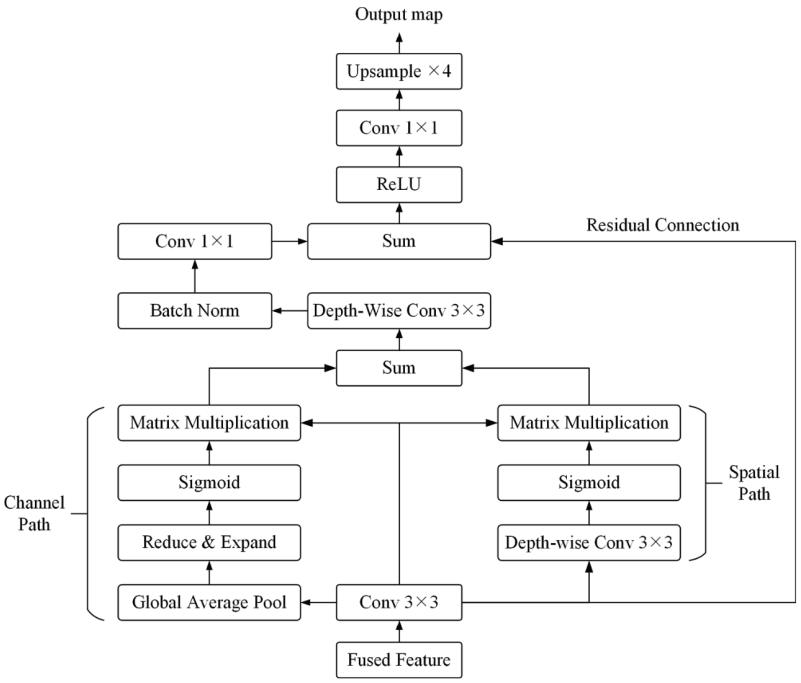

Figure 2. The illustration of the FRH

3.1. CNN-based encoder

As the ResNet18 (He et al., 2016) has demonstrated effectiveness and efficiency simultaneously in a wide range of real-time semantic seg- mentation tasks, we select

the pretrained ResNet18 as the encoder here to extract multi-scale semantic features with significantly low computational cost. ResNet18 consists of four-stage Resblocks, with each stage down-sampling the feature map with a scale factor of 2. In the proposed UNetFormer, the feature maps generated by each stage are fused with the corresponding feature maps of the decoder by a 1 × 1 convolution with the channel dimension in 64, i.e., the skip connection. Specifically, the semantic features produced by the Resblocks are aggregated with the features generated by the GLTB of the decoder using a weighted sum operation. The weighted sum operation weights the two features selectively based on their contributions to segmentation accuracy, thereby learning more generalized fusion features (Tan et al., 2020). The formulation of the weighted sum operation can be denoted as:

$$FF = \alpha \cdot RF + (1\ \alpha) \cdot GLF \quad (1)$$

where FF represents the fused feature, RF denotes the feature produced by the Resblocks, and GLF indicates the feature generated by the global–local Transformer block.

3.2. Transformer-based decoder

Complicated human-made objects occur frequently in fine-resolution remotely sensed urban images, which makes it difficult to achieve pre- cise real-time segmentation without global semantic information. To capture the global context, mainstream solutions focus on attaching a single attention block at the end of the network (Wang et al., 2018) or introducing Transformers as the encoder (Chen et al., 2021b). The former cannot capture multi-scale global features, whereas the latter significantly increases the complexity of the network and loses spatial details. In contrast, in the proposed UNetFormer, we utilize three Fig. 3. An overview of the UNetFormer.global–local Transformer blocks and a feature refinement head to build a lightweight Transformer-based decoder, as shown in Fig. 3. With such a hierarchical and lightweight design, the decoder is capable of capturing both global and local contexts at multiple scales while maintaining high efficiency.

3.3. Loss function

In the training phase, we employ not only the primary feature refinement head but also build an extra auxiliary head to optimize the global–local Transformer blocks, as shown in Fig. 3. This multi-head segmentation architecture has been demonstrated to be effective in previous research (Yu et al., 2020; Zhu et al., 2019). Based on the multi- head design, we apply a principal loss and an auxiliary loss to train the entire network. The principal loss L p is a combination of a dice loss L dice and a cross-entropy loss L ce.

4.Experiment

We choose Vaihingen and Potsdam as our datasets. In Potsdam, we have total 76 images with resolution of 6000*6000. Among them, 62 images are used for training and 14 images are used for validation. For Vaihingen , we have 33 images in total. 25 out of them are used for training and the rest are for validation.

Since the satellite images are too large to feed into the model, we need to slice them into smaller images patches (600*600).As shown in the figure below, we divide the

original images into several smaller images as well as their corresponding masks. In our experiment, we use one NVIDIA A100 for training. The metrics includes mIoU (mean intersection of union), mean F1 score and OA (overall accuracy). The results are shown in Table 1 and Table 2.

Figure 3. Original large images from potsdam dataset

Figure 4. smaller images from potsdam dataset for training

encoder	mIoU	F1	OA
CNN	83.3	90.6	90.5
Transformer	81.4	89.5	88.9

Table 1. The experimental results from Potsdam dataset

encoder	mIoU	F1	OA
CNN	85.1	91.8	92.6
Transformer	84.5	91.5	92.2

Table 1. The experimental results from Vaihingen dataset

The experimental results suggest that on two dataset, CNN-based and transformer-based encoders have similar performance for the same segmentation model. Specifically, CNN-based encoder can yield slightly better performance than transformer-based encoder. For instance, in Potsdam, the mIoU of CNN-based encoder is almost 2% higher than that of transformer-based encoder. The results show empirically that CNN and vision transformers are almost equivalent as the encoder. However, CNN-based encoders have slightly better performance.

We also show some visualization of the ground-truth mask and predicted masks from CNN-based model and transformer-based model.

Figure 5. Left is the testing image and right is the ground-truth mask

Figure 6. Left is the predicted mask from CNN-based encoder and right is form the transformer-based encoder

5.Conslusion

We investigate the performance of CNN-based and transformer-based encoder on the segmentation of the satellite images. The results show that the two encoders are almost equivalent and the differences between them are trivial in terms of the performance.

References

Audebert, N., Le Saux, B., Lef`evre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495.

Bazi, Y., Bashmal, L., Rahhal, M.M.A., Dayil, R.A., Ajlan, N.A., 2021. Vision transformers for remote sensing image classification. Remote Sensing 13, 516.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv: 2105.05537.

Chen, H., Qi, Z., Shi, Z., 2021a. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021b. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, K., Zou, Z., Shi, Z., 2021c. Building Extraction from Remote Sensing Images with Sparse Token Transformers. Remote Sensing 13, 4441.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062.

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

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018b. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848.

Deng, P., Xu, K., Huang, H., 2021. When CNNs meet vision transformer: A joint framework for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114.
Ding, L., Tang, H., Bruzzone, L., 2021. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 59, 426–435.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929.
Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.
Gao, L., Liu, H., Yang, M., Chen, L., Wan, Y., Xiao, Z., Qian, Y., 2021. STransFuse: Fusing Swin Transformer and Convolutional Neural Network for Remote Sensing Image Semantic Segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 10990–11003.
Griffiths, D., Boehm, J., 2019. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. ISPRS J. Photogramm. Remote Sens. 154, 70–83.
Guo, Y., Jia, X., Paull, D., 2018. Effective Sequential Classifier Training for SVM-Based Multitemporal Remote Sensing Image Classification. IEEE Trans. Image Process. 27, 3036–3048.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
He, X., Chen, Y., Lin, Z., 2021. Spatial-spectral transformer for hyperspectral image classification. Remote Sensing 13, 498.
Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J., 2021. SpectralFormer: Rethinking hyperspectral image classification with transformers. IEEE Transactions on Geoscience and Remote Sensing.
Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K., Sclaroff, S., 2020. Real-time semantic segmentation with fast attention. IEEE Rob. Autom. Lett. 6, 263–270.
Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T.S., 2020. CCNet: Criss-Cross Attention for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Kampffmeyer, M., Salberg, A.-B., Jenssen, R., 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 1–9.
Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens. 145, 60–77.
Kirillov, A., Girshick, R., He, K., Doll´ar, P., 2019. Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6399–6408.