LE-BEIT: A LOCAL-ENHANCED SELF-SUPERVISED TRANSFORMER FOR SEMANTIC SEGMENTATION OF HIGH RESOLUTION REMOTE SENSING IMAGES

Yifei Huang[†] *Zideng Feng*[†] *Junli Yang*[†⋆] *Bin Wang*[‡] *Jiaying Wang*[†] *Zhenglin Xian*[†]

[†] Beijing University of Posts and Telecommunications
[‡] China Mobile Research Institute
[⋆] Corresponding Author: yangjunli@bupt.edu.cn

ABSTRACT

Semantic segmentation for remote sensing images (RSI) has been a thriving research topic for a long time. Existing supervised learning methods usually require a huge amount of labeled data. Meanwhile, large size, variation in object scales, and intricate details in RSI make it essential to capture both long-range context and local information. To address these problems, we propose Le-BEiT, a self-supervised Transformer with an improved positional encoding Local-Enhanced Positional Encoding (LePE). Self-supervised learning relieves the demanding requirement of a large amount of labeled data. The self-attention mechanism in Transformer has remarkable capability in capturing long-range context. Meanwhile, we use LePE as a substitution for Relative Positional Encoding (RPE) to represent local information more effectively. Moreover, considering the domain difference between natural images and RSI, instead of ImageNet-22K, we pre-train Le-BEiT on a very small high-resolution RSI dataset—GID. To investigate the influence of pre-training dataset size on segmentation accuracy, we furtherly conduct experiments on a larger pre-training dataset called GID-DOTA, which is 1/100 of ImageNet-22K, and have observed considerable accuracy improvements. The result of our method, which relies on a much smaller pre-trained dataset, achieves competitive accuracy compared to the counterpart on ImageNet-22K.

***Index Terms*—** Remote Sensing, Self-supervised Learning, Transformer

1. INTRODUCTION

Semantic segmentation is a fundamental and challenging task in computer vision. In recent years, semantic segmentation for remote sensing images (RSI) has been widely used in many fields, such as disaster management [1] and environmental monitoring [2].

For semantic segmentation of RSI, there are three crucial challenges: First, RSI features large size, diverse object scales, and complex details. These characteristics make it necessary to capture both long-range context and local information. Existing CNN-based algorithms lack the ability to model long-range context. Second, existing supervised learning methods require numerous labeled data, which is usually human labor-consuming. Third, most existing semantic segmentation methods for RSI use models pre-trained on ImageNet [3] to improve robustness. However, considering the domain difference between RSI and natural scene images, transfer-learning methods based on ImageNet [3] may be suboptimal if applied directly to the segmentation of RSI.

Recently, Transformer has been popularly and successfully used in computer vision [4–7], most of which outperform CNN in classification and segmentation tasks. BEiT [8] is a competitive self-supervised Transformer model, and it obtains excellent performance when fine-tuning on semantic segmentation tasks. Also, self-supervised learning is an effective solution to mitigate the demand of labeled data. Moreover, self-attention mechanism in Transformer [9] has a strong ability in capturing long-range dependency.

However, self-attention [9] mechanism itself cannot handle position information in images. To solve this problem, Positional Encoding such as Relative Positional Encoding (RPE) [10] is applied to retain positional information. In CSWin-Transformer [7], Local-Enhanced Positional Encoding (LePE) is introduced to enforce stronger local inductive bias, making it more effective in segmentation tasks for RSI due to intricate details in such images.

Therefore, we leverage LePE [7] as a substitution for RPE [10] in BEiT [8] and propose Le-BEiT, a self-supervised Transformer model with an improved positional encoding which is customized for RSI semantic segmentation. We pre-train Le-BEiT on RSI datasets instead of ImageNet-22K [3], and then fine-tune on RSI dataset Potsdam, enabling the model to better study the unique characteristics of RSI. We also use two different-size RSI datasets—GID [11] and GID-DOTA to study the influence of pre-training data size on segmentation accuracy.

We summarize our contributions as follows: 1) We make the first attempt to introduce the self-supervised Transformer BEiT [8] into semantic segmentation tasks for RSI. Aimed at better handling the features of RSI, based on BEiT [8], we propose Le-BEiT with an improved positional encoding LePE [7]. 2) we pretrain Le-BEiT on GID [11] and finetune for semantic segmentation tasks on Potsdam dataset. 3) We enlarge self-supervised pre-training dataset GID [11] to GID-DOTA and has observed obvious improvements on segmentation accuracy. Moreover, the result on GID-DOTA (about 1/100 ImageNet-22K [3]) even nearly achieves the accuracy on ImageNet-22K [3]. The code is available at https://github.com/lqwrl542293/JL-Yang_CV/tree/master/Le-BEIT

2. METHOD

2.1. Method Architecture

In this paper, we propose a self-supervised Transformer Le-BEiT which targets semantic segmentation tasks for RSI. The method consists of two important processes: First, we pretrain Le-BEiT in a self-supervised manner on unlabeled RSI datasets. Second, we perform semantic segmentation tasks on labeled RSI datasets. Fig.1 shows our method architecture.

2.2. Self-supervised Pre-training

One challenge in RSI semantic segmentation is that it is extremely costly to obtain numerous reliable annotated RSI datasets. To address this issue, we employ a self-supervised learning (SSL) method—Masked Image Modeling (MIM) [8] to pre-train our model. As shown in Fig.1 (a), each image has two views during pre-training: image patches and visual tokens, which serve as the input and output representations respectively. Image patches are obtained by splitting the image into a sequence of pieces, while visual tokens [13] are the outputs of dVAE [14] tokenizer. As shown in Fig.1, after the random masking algorithm, image patches are sent into Le-BEiT encoder (illustrated in Section 2.4). For each masked position, a MIM head [8] is used to predict the corresponding visual token, instead of the raw pixels of the image patch. This SSL process helps to learn general invariant features of RSI among different representations of images.

2.3. Transfer Learning

A large number of semantic segmentation methods on natural scene images rely on pre-trained models on ImageNet [3] for transfer learning. Considering the domain difference between RSI and natural scene images, to bring out the maximum potential of transfer learning, we perform self-supervised pretraining on high-resolution RSI datasets, GID [11] and GID-DOTA, and then fine-tune on RSI datasets. Details of GID and GID-DOTA datasets are described in Section 3.1.

2.4. Le-BEiT Encoder + UperNet Decoder

It is crucial to capture both long-range context and local information in segmentation for RSI. To address these issues, we design Le-BEiT encoder based on BEiT [8] and LePE [7]. Following BEiT, ViT [15] is used as the backbone structure. As shown in Fig.1 (b), images are split into patches and flattened into patch embeddings, which serve as the input of the encoder. The encoder consists of N consecutive blocks, each of which mainly contains a multi-head self-attention (MSA) [9] and an Multilayer Perceptron (MLP) [15] block. MSA models dependencies of features without regard to their distance in RSI and thus can model the long-range context to the maximum extent. To show the characteristics of the feature maps extracted by Le-BEiT encoder, we randomly select two output channels for each stage in Fig. 1 (b, c) of base-size Le-BEiT and visualize the feature maps in Fig. 2. The feature maps of Stage N1 and N2 have finer granularity, and the receptive field of each patch is relatively small. However, the self-attention mechanism can capture long-distance dependencies across the entire image. Stage N3 and N4 have larger receptive fields, so class features can be seen more obviously.

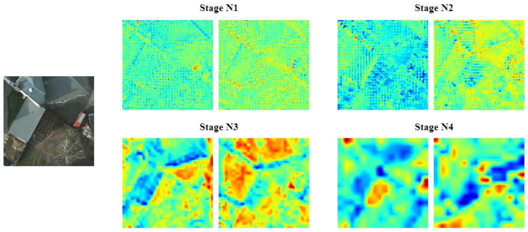

Fig. 2: Feature maps of base-size Le-BEiT.

Meanwhile, to compensate for the positional information of the self-attention mechanism, as well as to strengthen the local information, we use LePE [7] as a substitution for RPE [10] in BEiT [8]. Different from RPE [10] which adds the positional information inside the attention calculation, LePE [7] adds the positional encoding as a parallel module to self-attention operation on the projected values V using a depthwise convolution operator [16] ($DWConv$), as shown in formula 1, where Q, K, and V are different representations calculated from image patches [15], and d_k is dimension of patch embedding divided by number of attention heads k. This design decouples positional encoding from the self-attention calculation, and enforces stronger local inductive bias in RSI.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V + DWConv(V) \tag{1}$$

For the decoder part, we utilize UperNet decoder [12, 17] for its high efficiency and better performance in feature aggregation. As shown in Fig.1 (c), pyramidal hierarchy is employed to fuse different levels of features in encoder to achieve more detailed segmentation.

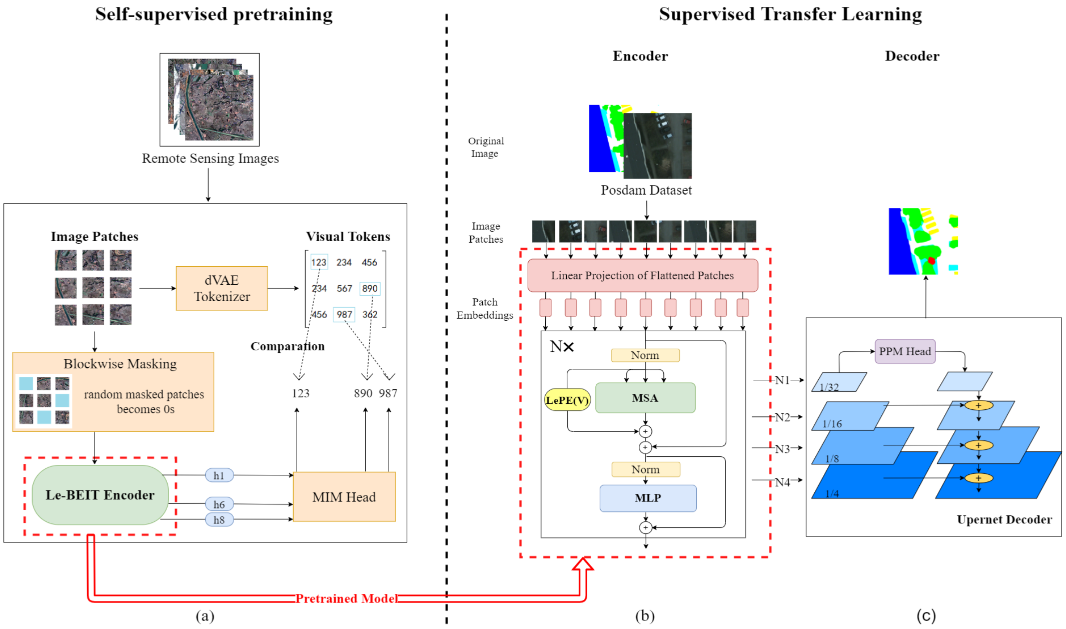

Fig. 1: Method Architecture. (a) The process of self-supervised pre-training. (b) Le-BEiT encoder. N is 12 for base size and 24 for large size. (c) UperNet [12] decoder. Block N1-N4 refers to the corresponding feature map generated by encoder Block N1-N4. N1-N4 is 3, 5, 7, 11 for base size and 7, 11, 15, 23 for large size.

3. EXPERIMENTS

3.1. Datasets

ISPRS Potsdam[1] is a 2D semantic labeling RSI dataset which contains 38 6000 × 6000 images with 5 cm resolution. We use RGB channels and apply official train-test split. Each image is divided into 144 tiles of 512 × 512.

GID [11] is a large-scale land-cover dataset. During pre-training, we only use RGB channels in its classification set which contains 150 images of size 7200 × 6800 pixels with 4m spatial resolution. Each image is split into 306 tiles of 400 × 400 pixels. GID is about 1/380 of ImageNet-22K [3] in data amount with respect to pixel numbers.

GID-DOTA In order to study the effect of pre-training data scale on segmentation accuracy, we combine four RSI datasets GID [11], DOTA v1.5 [18], NWPU-RESISC45 [19] and NWPU VHR-10 [20] in to a larger dataset—GID-DOTA, which is about 1/100 of ImageNet-22K [3].

3.2. Evaluation Metrics and Implementation Details

We choose OA and mIoU as the evaluation metrics of the model performance. During fine-tuning, we test model performance on testing set every 2000 iterations and thus we regard OA of the 2000th iteration as the *start point accuracy of segmentation* (called *start OA*). The maximum OA value in each experiment is denoted by *max OA*.

We denote original BEiT [8] with RPE [10] and Le-BEiT by BEiT-B/L and Le-BEiT-B/L, in which B/L stands for base/large size. Notice that BEiT-B/L is pretrained on ImageNet-22K [3]. We use the same pre-processing operations and loss function as BEiT [8].

3.3. Results and Analysis

3.3.1. *Comparing BEiT [8] with other SOTA Transformers*

We compare BEiT [8] which is self-supervised pre-trained on ImageNet-22K [3] with the latest Transformer methods. Segmenter [5], Segformer [4], and Swin Transformer [6] which are supervised pre-trained on ImageNet-22K [3] are used as comparation methods. Table 1 lists the *max OA* and max mIoU. BEiT-L [8] reaches 89.48% on OA, which is the SOTA among all methods, showing its competitiveness in semantic segmentation of RSI.

Table 1: Comparing BEiT with SOTA Transformers

Method	OA	mIoU
Segmenter-S [5]	88.51	73.72
Segformer-L [4]	89.29	75.49
Swin-B [6]	89.24	75.85
BEiT-B [8]	**88.89**	**74.39**
BEiT-L [8]	**89.48**	**75.79**

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

3.3.2. *Ablation Study*

Next, we focus on pre-training the improved Le-BEIT on RSI datasets. We conduct ablation study from three aspects to analyze the contributions of each component in our approach:

Increasing the number of pre-training epochs

We pre-train Le-BEIT-L for 100, 300, 500 epochs on GID [11] and compare them with the model trained from scratch. Limited by GPU computation ability, we only conduct pre-training up to 500 epochs. We give quantitive comparison in Table 2 on *start OA* and *max OA*. *Start OA* illustrates the effect of pre-training while *max OA* reveals the effect of transfer learning.

On one hand, Le-BEIT-L pre-trained for 500 epochs achieves 77.48% on *start OA*, which is 1.53% higher than that of 300 epochs, 5.05% higher than that of 100 epochs, and 12.3% higher than that of Le-BEIT-L trained from scratch. On the other hand, it achieves 87.92% on *max OA*, which is 0.15% higher than that of 300 epochs, 1.31% higher than that of 100 epochs, and 7.53% higher than that of trained from scratch. It can be seen that as the number of pre-training steps increases, *max OA* indeed obtains improvements, which is relatively smaller than that of *start OA*. These results show that increasing the number of pre-training epochs to a certain extent will bring more satisfying performance on RSI segmentation tasks.

Table 2: Ablation study on different pre-training epochs

Method	Start OA	Max OA
Le-BEIT-L (trained from scratch)	65.18	80.39
Le-BEIT-L (100 epoch)	72.43	86.61
Le-BEIT-L (300 epoch)	75.95	87.77
Le-BEIT-L (500 epoch)	77.48	87.92

Enlarging the size of pre-training datasets

To study the effect of pre-training dataset size on accuracy, we pretrain Le-BEIT on GID [11] and GID-DOTA for 500 epochs. In Table 3 we report *Start OA* and *max OA* and list results of BEIT[8] pre-trained on ImageNet-22K as well. Shifting from GID [11] to GID-DOTA, Le-BEIT-B and Le-BEIT-L obtains 1.53% and 4.21% improvements on *start OA* as well as 0.69% and 0.48% on *max OA*.

Compared to BEIT[8] pre-trained on ImageNet-22K, Le-BEIT-L even achieves 6.22% improvements on *start OA*. Also, the *max OA* of Le-BEIT-B is only 0.66% lower than the *max OA* of BEIT-B [8], noticing that the former dataset is only 1/100 of the latter. Therefore, it is very promising that if the pre-training RSI dataset is as large as ImageNet-22K [3], the performance will even outperform BEIT pretrained on ImageNet-22K [3].

Comparison on different positional encoding

Lastly, we train Le-BEIT and BEIT [8] on GID [11] (denoted by BEIT-GID) for 500 epochs and keep all other parameters the same. We report *start OA*, *max OA* and Flops in Table 4. LePE[7] improves the segmentation performance of base and large BEIT-GID by 11.59% and 7.19% on *start OA* respectively, which are remarkably encouraging improvements. On *max OA*, LePE brings 1.96% and 0.82% improvements. Obvious accuracy gains with little extra computation expenses overhead prove that our method is indeed more suitable for semantic segmentation tasks for RSI.

Table 3: Results on different pretraining datasets

Method	Pre-training Data	Start OA	Max OA
Le-BEIT-B	GID	81.67	87.54
Le-BEIT-B	GID-DOTA	**83.2**	**88.23**
BEIT-B	ImageNet-22K	84.81	**88.89**
Le-BEIT-L	GID	77.48	87.92
Le-BEIT-L	GID-DOTA	**81.69**	**88.4**
BEIT-L	ImageNet-22K	75.47	89.48

Table 4: Ablation study on different positional encoding. Flops are computed with the input size (3,512,512).

Method	Start OA	Max OA	Flops
Le-BEIT-L	77.48	87.92	1115.37 GFLOPs
BEIT-GID-L	65.89	85.96	1115.11 GFLOPs
Le-BEIT-B	81.67	87.54	562.22 GFLOPs
BEIT-GID-B	74.48	86.72	562.13 GFLOPs

4. CONCLUSION

We propose a self-supervised Transformer Le-BEIT and perform self-supervised pre-training on RSI datasets. Self-supervised pre-training on RSI datasets relieves the requirements of numerous labeled data. Also, it makes transfer learning highly effective when fine-tuning on downstream RSI datasets for semantic segmentation. We employ Transformer architecture to capture long-range context and use LePE [7] to better model local information. Our work achieves a similar accuracy level compared to SOTA method with a much smaller pre-training data amount.

Future works include: 1) Furtherly increase the scale of pre-training datasets and the number of pre-training epochs. 2) Try to involve a typical method of SSL—contrastive learning for further exploration of semantic segmentation for RSI.

5. ACKNOWLEDGEMENT

This work is funded by Research Innovation Fund for College Students (No.202100008) and Teaching Reform Project (No.2019JYA05) of Beijing University of Posts and Telecommunications. This work is partially supported by MoE CMCC "Artifical Intelligence" Project No. MCM20190701, which provides JIUTIAN AI Platform.

References

[1] Christos Kyrkou and Theocharis Theocharides, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles.," in *CVPR Workshops*, 2019, pp. 517–525.

[2] Shengke Wang, Lu Liu, Liang Qu, Changyin Yu, Yujuan Sun, Feng Gao, and Junyu Dong, "Accurate ulva prolifera regions extraction of uav images with superpixel and cnns for ocean environment monitoring," *Neurocomputing*, vol. 348, pp. 158–168, 2019.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[4] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint arXiv:2105.15203*, 2021.

[5] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid, "Segmenter: Transformer for semantic segmentation," *arXiv preprint arXiv:2105.05633*, 2021.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[7] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *arXiv preprint arXiv:2107.00652*, 2021.

[8] Hangbo Bao, Li Dong, and Furu Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[10] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[11] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, pp. 111322, 2020.

[12] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 418–434.

[13] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[14] Jason Tyler Rolfe, "Discrete variational autoencoders," *arXiv preprint arXiv:1609.02200*, 2016.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[17] MMSegmentation Contributors, "Mmsegmentation : Openmmlab semantic segmentation toolbox and benchmark,
https://github.com/open-mmlab/mmsegmentation," .

[18] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[19] Gong Cheng, Junwei Han, and Xiaoqiang Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[20] Hao Su, Shunjun Wei, Min Yan, Chen Wang, Jun Shi, and Xiaoling Zhang, "Object detection and instance segmentation in remote sensing imagery based on precise mask r-cnn," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 1454–1457.