ETH zürich

IKG
Institut für Kartografie und Geoinformation
Institute of Cartography and Geoinformation

Semantic Segmentation of Historical Maps with Self-Supervised Vision Transformers

Author: Shupeng Wang,
Supervisors: Prof. Dr. Lorenz Hurni, Xue Xia, Chenjing Jiao

1 Introduction

Convolutional Neural Networks have been widely used as mainstream models for extracting meaningful information from historical maps. However, in recent years, the success of Self-Supervised Learning and Vision Transformers in the field of computer vision has sparked interest in applying these techniques also to historical maps.

In line with this trend, this project proposes an approach that combines contrastive learning and the Swin Transformer (SwinUNet SSL). The aim is to improve the accuracy and efficiency of semantic segmentation specifically for railway features in historical maps.

2 Data

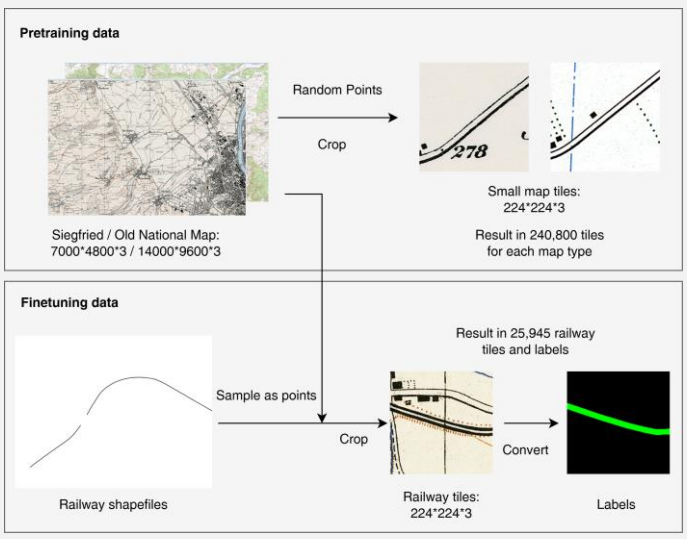

Figure 1: Data preprocessing workflow

For pretraining datasets, they are split into train and val sets in an 8:2 ratio.

For finetuning datasets, they are divided into train, val, and test sets in an 8:1:1 ratio. Smaller datasets consisting of 10%, 5%, and 2.5% of data are prepared. In addition, each dataset is also mixed with several road tiles.

3 Methodology

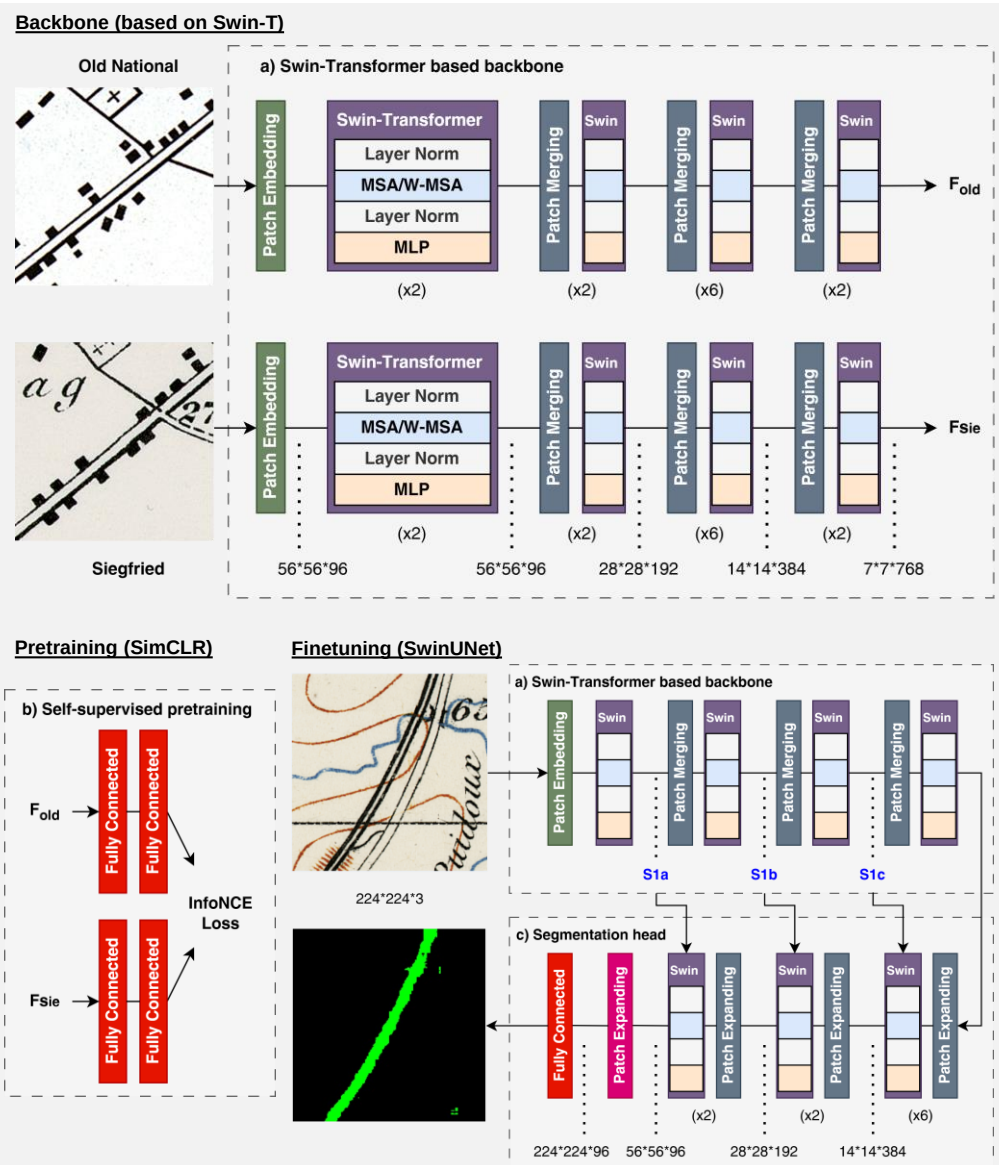

Figure 2: Model architecture

4 Results and Discussion

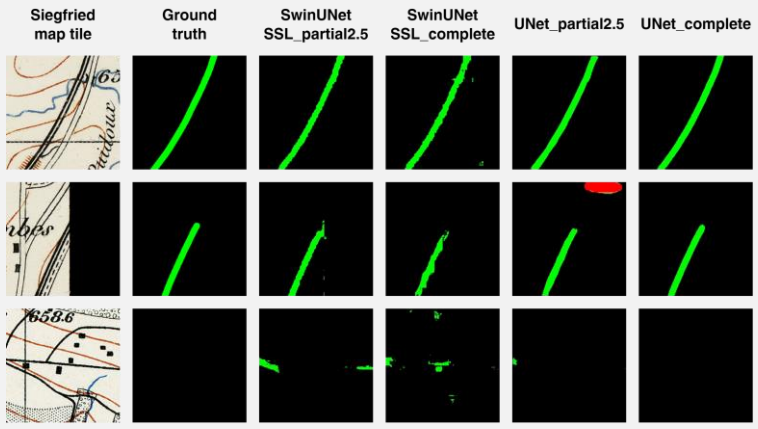

Figure 3: Model performance on the semantic segmentation of normal railways

** Experiments have also been conducted on tunnel railways and narrow railways, but only UNet_complete is able to identify pixels in those classes.

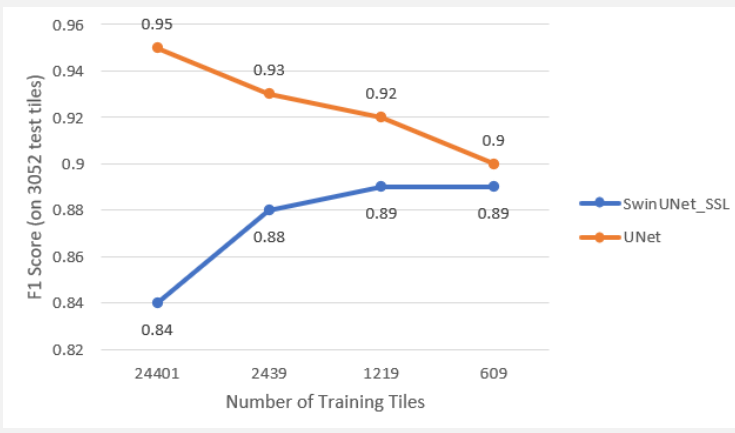

Figure 4: F1 score for normal railways

Figure 3 and 4 demonstrate that the results obtained from the SwinUNet SSL are generally not as smooth as those from the UNet model. Both models still exhibit misclassification of pixels, but when trained on a large dataset, the UNet model shows significant improvement in performance. The trend, however, is reversed for SwinUNet SSL.

5 Conclusion and Future Directions

In summary, the proposed SwinUNet SSL model in this project has demonstrated several strengths, as well as limitations.

+ **Good performance with small training dataset**: Comparable level of performance to mainstream models when data availability is limited.
+ **Efficient training speed**: Although not explicitly mentioned previously, the model has showcased the same level of training speed as mainstream models.
- **Fair performance with large training dataset**: The model's performance diminishes when trained on larger datasets.
- **Difficulty in handling imbalanced training classes**: The model encounters challenges in learning meaningful features from classes such as tunnel and narrow railways mainly due to scarcity of data.

Further directions:

- **Improve SimCLR performance**: Increasing the batch size or involving data augmentation during pretraining can introduce new positive pairs and increase the number of negative example per positive pair. This can potentially help the model generalize better and learn more robust representations.
- **Explore alternative pretraining strategies**: Other SSL models such as BEiT or Masked Autoencoders (MAE) might offer different advantages that are more suitable for Transformer-based backbones.
- **Adopt different loss functions during finetuning**: Alternative loss functions like focal loss can be employed instead of the current BCE loss to address the misclassification of pixels in imbalanced classes.

References

1. 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*. http://arxiv.org/abs/2105.05537
2. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. http://arxiv.org/abs/2002.05709
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. http://arxiv.org/abs/2103.14030
4. Scheibenreif, L. M., Hanna, J., Mommert, M., & Borth, D. (2022). *Self-supervised Vision Transformers for Land-cover Segmentation and Classification*. https://openaccess.thecvf.com/content/CVPR2022W/EarthVision/papers