Semantic Segmentation of Remote Sensing Images with Transformers and Self-Supervised Learning

Object Recognition and Computer Vision Project Report

Franki Nguimatsia and Bertille Temple
Ecole Normale Supérieure Paris Saclay
4 Av. des Sciences, 91190 Gif-sur-Yvette
franki.nguimatsiatiofack@ensae.fr and templebertille@gmail.com

January 15, 2024

Abstract

Land cover semantic segmentation is used in various Remote Sensing(RS) applications, notably in earth change detection to quantify soils artificialisation. After the publication [8] in 2015, Convolutional Neural Networks (CNNs) with U-Net architectures has imposed itself as the standard choice for this task, but the emergence of Vision Transformers (ViT) applied to segmentation in 2021 with [4] could lead to a paradigm shift. In general, ViT outperforms CNN with a sufficient large dataset and here comes the bottleneck : labeled data in semantic segmentation is costly and slow to obtain. To tackle this issue, the article under study [7] proposes a self supervised pre-training on large volumes of unlabeled data.

1. Introduction

Semantic segmentation consists in classifying each pixel of an image. After the CNNs success in 2012 with [1], the UNet model imposed itself as the reference model in semantic segmentation. Presented in [8], this model has a symmetric architecture with downsampling and upsampling paths connected by skip connections, allowing for precise localization. As introduced in [6], RESNET can be used as pretrained-weigths to benefit from transfer learning.

The success of transformers in natural language processing (NLP) in 2017 led to their adaptation in computer vision with ViT presented in [2]. The architecture of ViT for semantic segmentation was introduced in [9] with SwinUNet in 2021. Unlike traditional ViTs where tokens (image patches) have a fixed size, this architecture allows for dynamic sizing of tokens. This is achieved through the concept of shifted windows from the SwinT model introduced in [5], which helps in learning scale-invariant representations. Also, SwinUNet architecture has downsampling and upsampling paths as in UNet.

The success of transformers in NLP has not reached the field in semantic segmentation. To bridge the gap between UNet and SwinUNet, [7] introduces the combined architecture ”SwinUNet+SSL” which combines a backbone as a contracting path with Swin Transformer(SwinT) blocks, and a specific head for segmentation which plays the role of an expanding path with SwinT blocks. The backbone is pre-trained on the SEN12MS 2.1 dataset using SSL with unlabelled RS data. Following this phase, a head designed for segmentation is added. The entire model, now including this head, is further trained on a smaller, labeled dataset. While [7] states that the head for segmentation is SwinUNet, it can be seen in 1 that it is actually closer to SwinT than UNet as it contains 3 SwinT blocks aligned in an expanding path, but does not contain the symmetric contracting path typical from UNet. The contracting path have been replaced by the backbone trained with SSL. Also, while the article mentions SSL, the fact that labeled data are used to train the head of their architecture leads us to think their approach is actually closer to a weakly supervised approach than a self supervised one. That being said, to avoid confusion, we made the choice to stick to the vocabulary from the article in this report.

The aim of this project is to evaluate the benefits of combining SwinUNet with Self-Supervised Learning (SSL) by comparing it to the outcomes of using UNet and SwinUNet on two distinct datasets. Furthermore, we investigate the application of transfer learning with UNet as an alternative method to SwinUNet + SSL.

In [7], the metric used is the pixel-wise accuracy by class, the average accuracy and the mean Intersection over

Union(mIoU). In [3], the metric is IoU by class and mIoU. To have comparable results with both articles, we generated both metrics: pixel-wise accuracy and IoU by class, as long as average accuracy and mIoU. Pixel-wise accuracy is calculated as the ratio of correctly classified pixels to the total number of pixels, given by the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The Intersection over Union (IoU) is defined as the ratio of the intersection to the union of the target and predicted segments, expressed as:

$$\text{IoU} = \frac{\text{Target} \cap \text{Prediction}}{\text{Target} \cup \text{Prediction}}$$

Accuracy should be interpreted with caution in the context of an unbalanced dataset. This is because a model that by default classifies all pixels as the predominant class can misleadingly show high accuracy, despite not truly being effective in making precise predictions without errors. We find the IoU metric is more interesting since the presence of the prediction term in the denominator penalizes the overall score in the presence of numerous false negatives or false positives.

2. Experiments

We did three experiments. The first one focused on investigating the advantages of combining SSL with Swin-UNet when applied to Sentinel data. This involved replicating the experiments described in [7].

The second experiment aimed to determine whether the combination of SSL with SwinUNet outperforms a UNet architecture when applied to a different dataset, specifically the FLAIR1 data presented in 2.3.

In the final experiment, conducted using the FLAIR1 dataset, we compared the performances of combining SSL with SwinUNet against the use of transfer learning coupled with UNet, to evaluate which methodology offers more benefits.

We implemented an early stopping strategy for all models. Some important hyper-parameters are detailed in Table **??**. For a comprehensive view of all parameters, please refer to our code repository at LandCover_map_Transformers_SSL.

2.1. Sentinel dataset

The SEN12MS dataset, used in the 2020 IEEE GRSS Data Fusion Contest, includes 180,662 pairs of spatially aligned observations from Sentinel-1 and Sentinel-2. Sentinel-1 provides 2 channels, while Sentinel-2 offers 13 channels, both with a 10m pixel resolution. Complementing this, the DFC2020 dataset, an extension of SEN12MS, comprises 6,114 paired Sentinel-1/2 observations with masks, divided into 16% for training and 84% for validation. These datasets are hosted by the German Aerospace Center (DLR), supporting research in RS.

2.2. Sentinel results

Sentinel data			
Model	Swin-Unet	Swin-UNet + SSL	
	S1	EarlyF + Frozen	EarlyF + Fine-tuned
Forest	0.09	0.56	0.51
Shrublang	0.01	0.38	0.30
Grassland	0.03	0.05	0.09
Wetland	0.47	0.1	0.18
Croplands	0.39	0.52	0.57
Urban/Built-up	0.68	0.68	0.66
Barren	0.07	0.24	0.29
Water	0.97	0.97	0.97
Average	0.41	0.58	0.58
Average from the article	0.33	0.48	0.51

Table 1: Results in term of pixel-wise accuracy. EarlyF performs Sentinel-1/2 data fusion at the model input. The SSL backbone can be frozen or fine-tuned.

Sentinel data			
Model	Swin-Unet	Swin-UNet + SSL	
	S1	EarlyF + Frozen	EarlyF + Fine-tuned
Forest	0.09	0.43	0.43
Shrublang	0.01	0.15	0.13
Grassland	0.02	0.04	0.06
Wetland	0.04	0.04	0.06
Croplands	0.23	0.33	0.34
Urban/Built-up	0.33	0.46	0.48
Barren	0.04	0.16	0.19
Water	0.90	0.93	0.92
mIoU	0.21	0.31	0.32
mIoU from the article	0.24	0.35	0.37

Table 2: Experimental results in term of iou

The results from Table 1 indicate that training for 30 epochs as we did, yields better accuracy (up to 10% higher) compared to 200 epochs as it was done in the article, suggesting potential overfitting in the article's approach. The mIoU is slightly better in the article compared to our results in Table 2. Both Swin-UNet + SSL witn frozen backbone and fine-tuned backbone show similar performance, which is not the case in the article. Table 1 shows individual class accuracies and mIoUs for different models, with Swin-UNet + SSL models generally outperforming the Swin-Unet model. The improved performance in specific classes like Forest, Shrublang, and Urban/Built-up in the Swin-UNet + SSL models is notable.

2.3. FLAIR1 dataset

The FLAIR1 dataset stands for French Land cover from Aerospace ImageRy and was published by the National Institute of Geographical and Forestry Data. From the original data with 50 departments, we kept only 2 (34 and 71), which makes a dataset of 2597 images with their corresponding masks. Each image has 5 channels: R, G, B, Infrared and Elevation. Each pixel in the mask is assigned a label between 1 and 19 corresponding to the land cover class. Classes from 13 to 19 represents less than 0.15% so we group them together in the class 13 ”other”, as it was done in [3]. The sub-dataset we built has similar proportions for each class as the original dataset as it can be seen in Figures 2 and 3 in Annex.

2.4. Flair 1 Results

Qualitative results Figure 6a and 6b illustrates that masks produced by RESNETUNet are smoother than the one produced by SwinUNet+SSL, which gives the intuition that RESNETUNet might be better to classify pixels at a fine grained level.

Quantitative results The confusion matrix in 5 we obtained with SwinUNet+SSL model reveals that the model tends to predict pixels as herbaceous vegetation, which makes sense as it is an over-represented class in the training set, as depicted previously in figure 3. Also, the model tends to confuse pervious surfaces, bare soil, and plowed land. This comes from an ambiguity from the classification itself as bare soil and plowed land are both pervious surfaces. Finally, the model confuse deciduous and conifereous which can be understable as both are trees, they probably have the same scale of values for the channel elevation. Finally, by comparing our confusion matrix with the one presented in [3] and shown in 4, we observe similar results, even though the models used differ (UNet in [3]).

Flair 1 data				
Model	UNet	Swin-UNet	Swin-UNet + SSL	
			Frozen	Fine-tuned
Epoch number	5	45	45	60
building	0.82	0.83	0.82	0.78
pervious surface	0.43	0.16	0.33	0.25
impervious surface	0.47	0.82	0.78	0.78
bare soil	0.22	0	0.19	0.07
water	0.95	0.86	0.83	0.78
coniferous	0.60	0.34	0.60	0.48
deciduous	0.87	0.88	0.84	0.83
brushwood	0.34	0.50	0.38	0.36
vineyard	0.92	0.89	0.78	0.77
herbaceous vegetation	0.61	0.81	0.84	0.78
agricultural land	0.64	0.35	0.45	0.36
plowed land	0.03	0.26	0.46	0.62
other	0	0	0	0
Average	0.53	0.70	0.55	0.53

Table 3: Results in term of pixel-wise accuracy

In [3] with UNet applied to FLAIR1, the best mIoU they obtain is 0.54. The results obtained in [7] with UNet, SwinUnet, SwinUNet+SSL applied to Sentinel are written in the Table 6 and can be consulted to compare with our results obtained with FLAIR1, keeping in mind that the models are the same but not the data.

Flair 1 data					
Model	UNet	Resnet50-UNet	Swin-UNet	Swin-UNet + SSL	
				Frozen	Fine-tuned
Epoch number	5	61	45	45	60
building	0.64	0.78	0.61	0.65	0.62
pervious surface	0.21	0.40	0.14	0.23	0.19
impervious surface	0.46	0.70	0.66	0.64	0.59
bare soil	0.21	0.59	0.01	0.17	0.06
water	0.80	0.92	0.82	0.72	0.65
coniferous	0.47	0.65	0.32	0.47	0.42
deciduous	0.76	0.77	0.69	0.69	0.68
brushwood	0.27	0.52	0.34	0.26	0.26
vineyard	0.84	0.91	0.84	0.72	0.71
herbaceous vegetation	0.44	0.71	0.56	0.50	0.51
agricultural land	0.34	0.69	0.28	0.31	0.28
plowed land	0.03	0.58	0.23	0.31	0.44
other	0	0	0	0	0
mIoU	0.42	0.69	0.42	0.44	0.42

Table 4: Results in term of IoU

SwinUNet without SSL gives similar performance as UNet in term of mIoU, justyfying the need for SSL to obtain an amelioration. Then, SwinUNet + SSL effectively outperforms both the baseline UNet and SwinUNet in term of mIoU as in the article. In term of average accuracy by class, Table 3 shows that the best model is SwinUnet with 0.7, but it should be taken with a grain of salt for the reasons already discussed in the paragraph about the metrics. Also, note that frozing the backbone in SwinUNet+SSL gives better result than finetuning when applied to FLAIR1 data, compared to the reverse for Sentinel. However note that the finetuning model converges much more slowly than with the frozen backbone, and that we stopped the training of the finetuning model at 60 epochs even if it did not converge. In other words, by training longer the finetuning model, we might obtain similar result as in the article. Finally, the best performance can be seen in table 4 with the model RESNETUNet which gives an impressive mIoU of 0.69. It leads us to claim that UNet combined with transfer learning seems to give better performance than SwinUNet+SSL.

3. Conclusion

While Self Supervised Learning can bring little performance to ViT in semantic segmentation, CNN combined with transfer learning remains better when applied to a subset of FLAIR1 data. This might be explainable by the fact that the architecture of ViT for semantic segmentation combined with SSL is quite recent and CNN with transfer learning in semantic segmentation benefited from more research until now. As a future experiment, we recommend conducting a proper search for hyper-parameters for each model. Additionally, we think that to really evaluate the impact of SSL, a model SwinUnet + SSL should be trained without the SSL backbone, and with the SSL backbone. To do so, the segmentation head should have its own contracting path independent from the backbone.

A. Annex

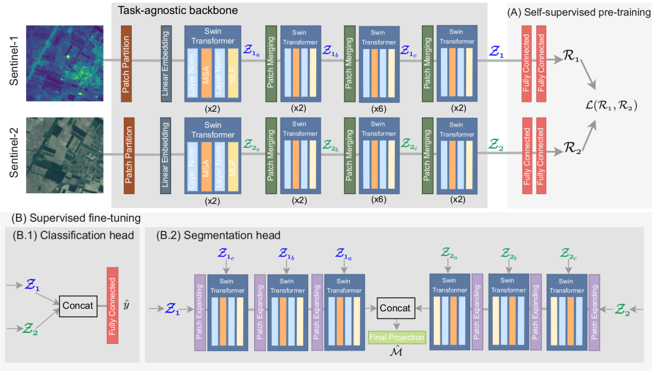


Figure 1: SwinUNet+SSL model, image taken from [7]

Parameter	UNet	SwinUNet	SwinUNet+SSL	RESNETUNet
Image Size	224	224	224	224
Batch Size	32	32	32	32
Optimizer	Adam	Adam	Adam	SGD
Learning rate	0.001	0.001	0.00001	0.02
Loss Function	Cross Entropy	Cross Entropy	Cross Entropy	Cross Entropy

Table 5: Experimental settings for models with FLAIR1

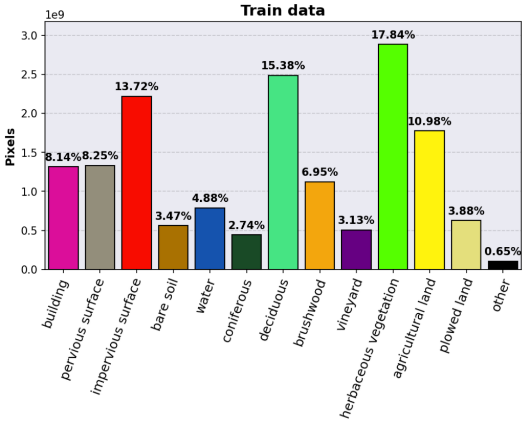


Figure 2: Class distribution from the original FLAIR1 dataset, plot taken from [3]

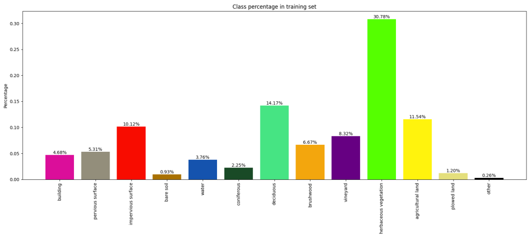


Figure 3: Class distribution of the FLAIR1 subset we built, plot made by us

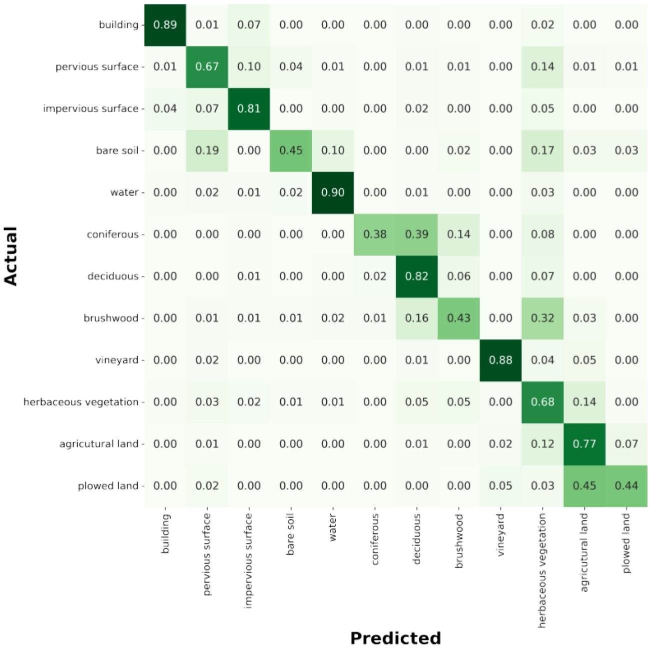


Figure 4: Confusion matrix obtained by A. Garioud et all in [3] with UNet applied to FLAIR1

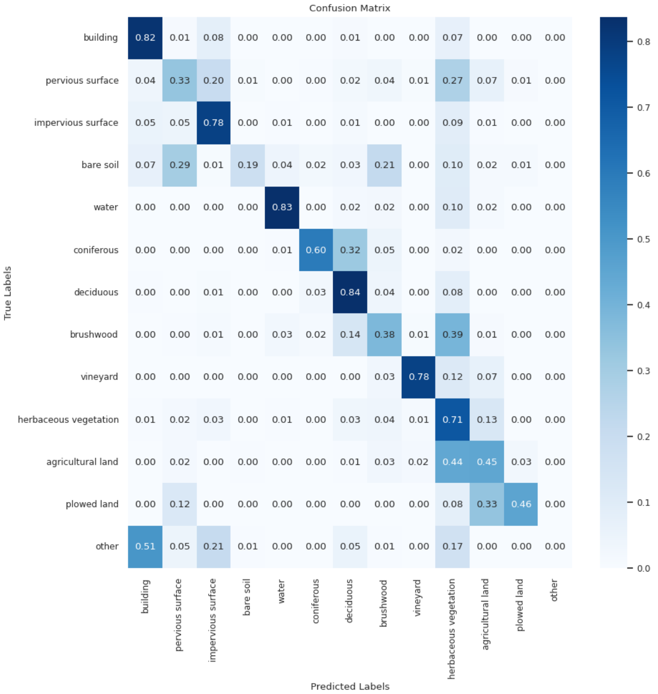
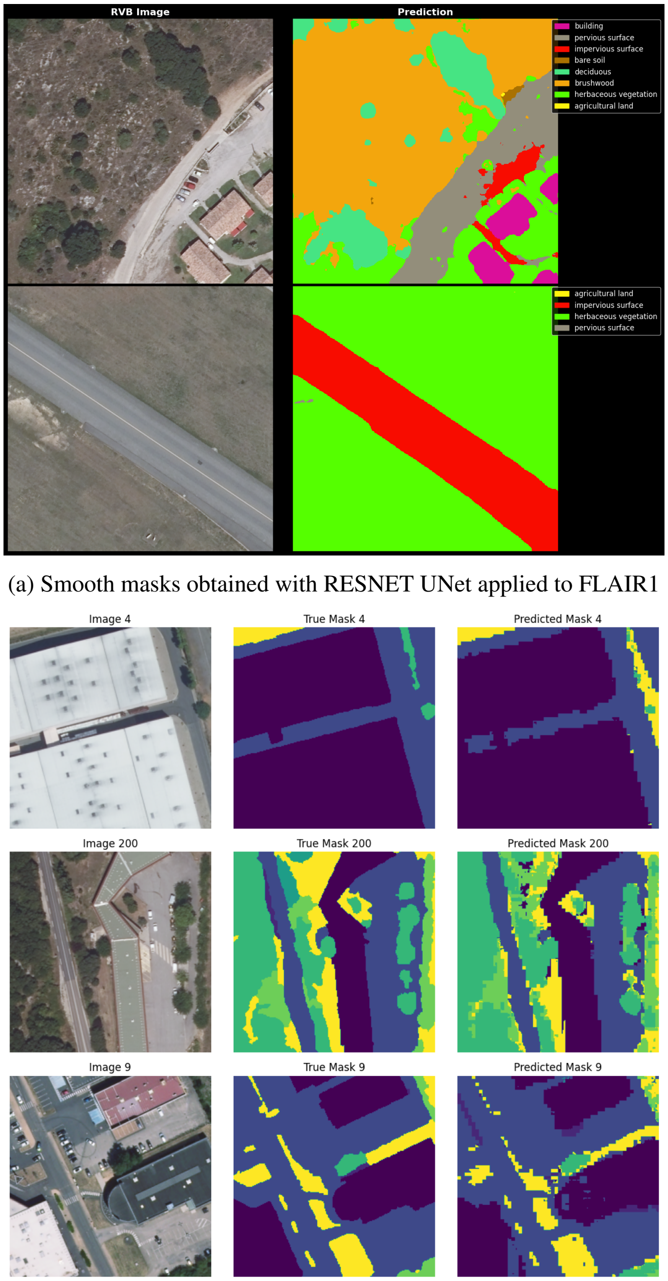


Figure 5: Confusion matrix we obtained with SwinUNet+SSL applied to FLAIR1



(a) Smooth masks obtained with RESNET UNet applied to FLAIR1

(b) Rough masks obtained with Swin UNet applied to FLAIR1

Figure 6: Masks produced by UNet vs SwinUNet

Sentinel 2 data				
Model	UNet	Swin-UNet	Swin-UNet + SSL	
			Frozen	Fine-tuned
Average accuracy	0.43	0.39	0.44	0.46
mIoU	0.31	0.28	0.32	0.35

Table 6: Results obtained by L. Scheibenreif et all in [7] with Sentinel 2

References

[1] Geoffrey E. Hinton Alex Krizhevsky, Ilya Sutskever. Imagenet classification with deep convolutional neural networks, 2012. 1

[2] Alexey Dosovitskiy et all. An image is worth 16*16 words: transformers for image recognition at scale, 2021. 1

[3] Anatol Garioud et all. Flair: French land cover from aerospace imagery, 2023. 2, 3, 4

[4] Hu Cao et all. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021. 1

[5] Ze Liu et all. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 1

[6] Shaoqing Ren Jian Sun Kaiming He, Xiangyu Zhang. Deep residual learning for image recognition, 2015. 1

[7] Michael Mommert Damian Borth Linus Scheibenreif, Joëlle Hanna. Self-supervised vision transformers for land-cover segmentation and classification, 2022. 1, 2, 3, 4, 5

[8] Thomas Brox Olaf Ronneberger, Philipp Fischer. U-net: Convolutional networks for biomedical image segmentation, 2015. 1

[9] Ivan Laptev Cordelia Schmid Robin Strudel, Ricardo Garcia. Transformer for semantic segmentation, 2021. 1