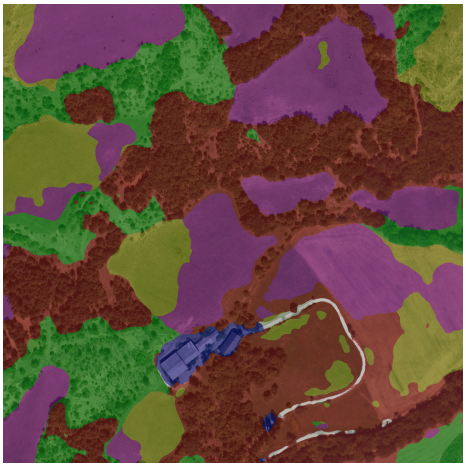
Satellite Image Semantic Segmentation

Eric Guérin[1], Killian Oechslin[1], Christian Wolf[1], Benoît Martinez[2]

[1] CNRS LIRIS - INSA Lyon
[2] Ubisoft

September 2021



Abstract In this paper, we propose a method for the automatic semantic segmentation of satellite images into six classes (sparse forest, dense forest, moor, herbaceous formation, building, and road). We rely on Swin Transformer architecture and build the dataset from IGN open data. We report quantitative and qualitative segmentation results on this dataset and discuss strengths and limitations. The dataset and the trained model are made publicly available.

1 Introduction

Virtual worlds in the context of digital entertainment need to be vast and realistic. These two factors force industries to resort to using artists massively. In

the same time, more and more geographic data such as digital satellite photography become publicly available. Unfortunately, this data is rarely segmented and cannot be used directly. In the context of the ANR project Ampli [1], we aim at making the task of virtual worlds authoring easier by providing a way to segment satellite images into six basic landcover classes. The segmentation method we use is Swin Transformer [2] (section 2) and we build the dataset from IGN public data (section 3). The obtained results are very promising (section 4) and the trained model is made publicly available together with the training dataset.

2 Swin Transformer Semantic Segmentation

Swin Transformer [2] is a general purpose computer vision backbone that has been proven very efficient and recently at the top of the state-of-the-art for image classification, object detection, and semantic segmentation. Its architecture based on Shifted WINdows makes it robust against scale variability while keeping linear efficiency with respect to the number of pixels. The Shift Windows concept consists in having a window shifted by half of its size in order to limit the self-attention computation to non-overlapping local windows while keeping possible to have cross-window connections.

In our experiments, we use an implementation [2] based on mmsegmentation [1].

3 Data Preparation and Setup

3.1 Dataset sources

To train and test the model, we used open data provided by IGN [3] which concerns French departments (Hautes-Alpes in our case). The following datasets have been used to extract the different layers:

- BD Ortho for the satellite images
- BD Foret v2 for vegetation data
- BD Topo for buildings and roads

Important: note that the data precision is 50cm per pixel. As BD Ortho is already in raster format, the only transformation we had to apply was resampling and cropping. In opposition, BD Foret and BD Topo are vector-based datasets that need to be rasterized before being used. We have used the `gdal_rasterize` command from GDAL tools to do so.

[1] `https://projet.liris.cnrs.fr/ampli/`
[2] `https://github.com/SwinTransformer/Swin-Transformer-Semantic-Segmentation`
[3] `https://geoservices.ign.fr/telechargement`

Initially, a large number of classes were present in the dataset. In BD Foret, a lot of information cannot be inferred from the satellite image (for example, difference between species). We reduced the number of classes by merging them and finally retained the following ones:

- Sparse forest
- Dense forest
- Moor
- Herbaceous formation
- Building
- Road

The purpose of the two last classes is twofold. We first wanted to avoid trapping the training into false segmentation, because buildings and roads were visually present in the satellite images and were initially assigned a vegetation class. Second, the segmentation is more precise and gives more identification of the different image elements.

3.2 Dataset preparation

Our training and test datasets are composed of tiles prepared from IGN open data. Each tile has a 1000x1000 resolution representing a 500m x 500m footprint (the resolution is 50cm per pixel). We mainly used data from the Hautes-Alpes department, and we took spatially spaced data to have as much diversity as possible and to limit the area without information (unfortunately, some places lack information). A total of 600 tiles have been used to train the model.

The file structure of the dataset is as follows:

```
|-- data
|   |-- ign
|   |   |-- annotations
|   |   |   |-- training
|   |   |   |   |-- xxx.png
|   |   |   |   |-- yyy.png
|   |   |   |   |-- zzz.png
|   |   |   |-- validation
|   |   |-- images
|   |   |   |-- training
|   |   |   |   |-- xxx.png
|   |   |   |   |-- yyy.png
|   |   |   |   |-- zzz.png
|   |   |   |-- validation
```

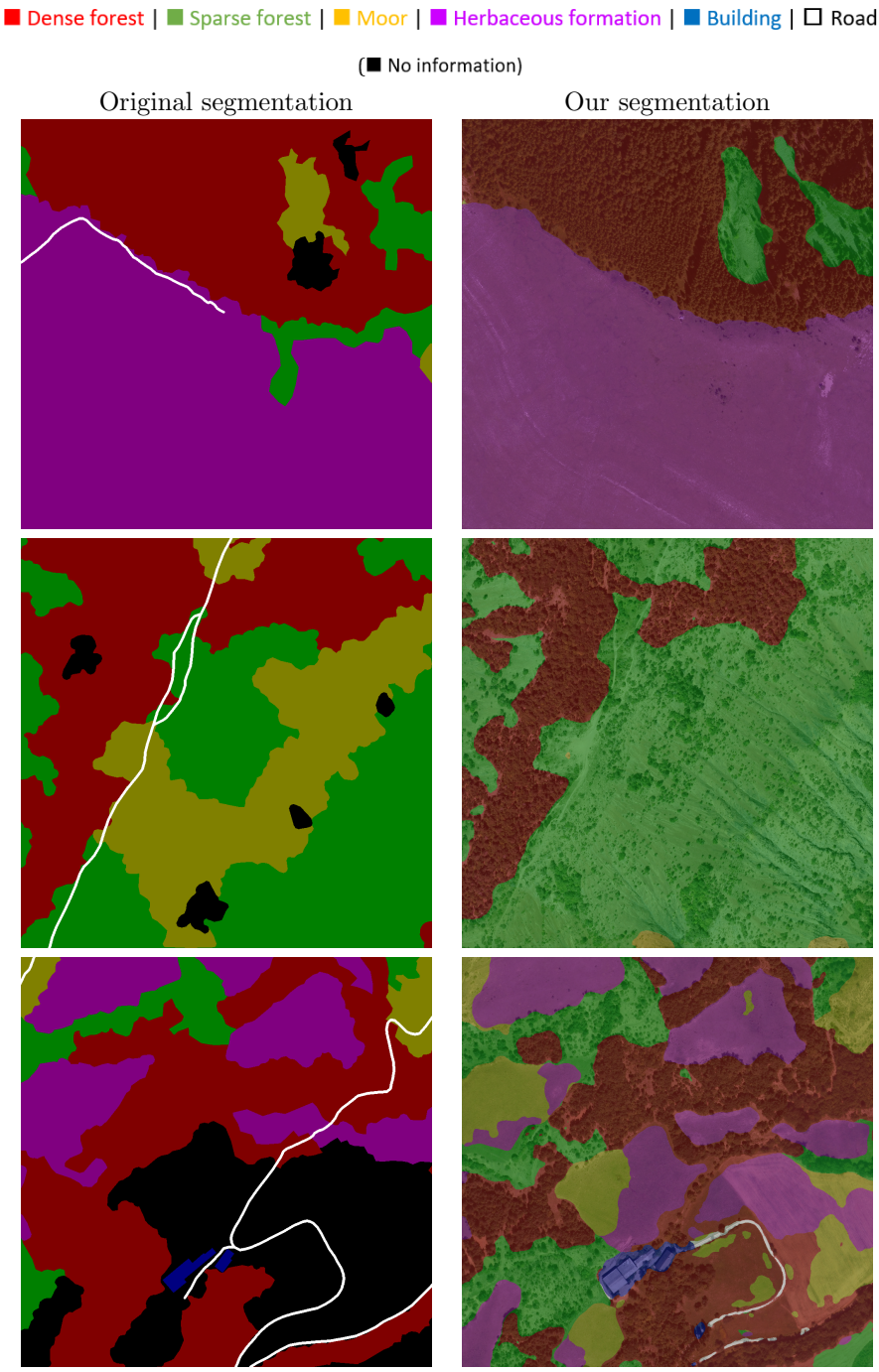


Figure 1: Main results

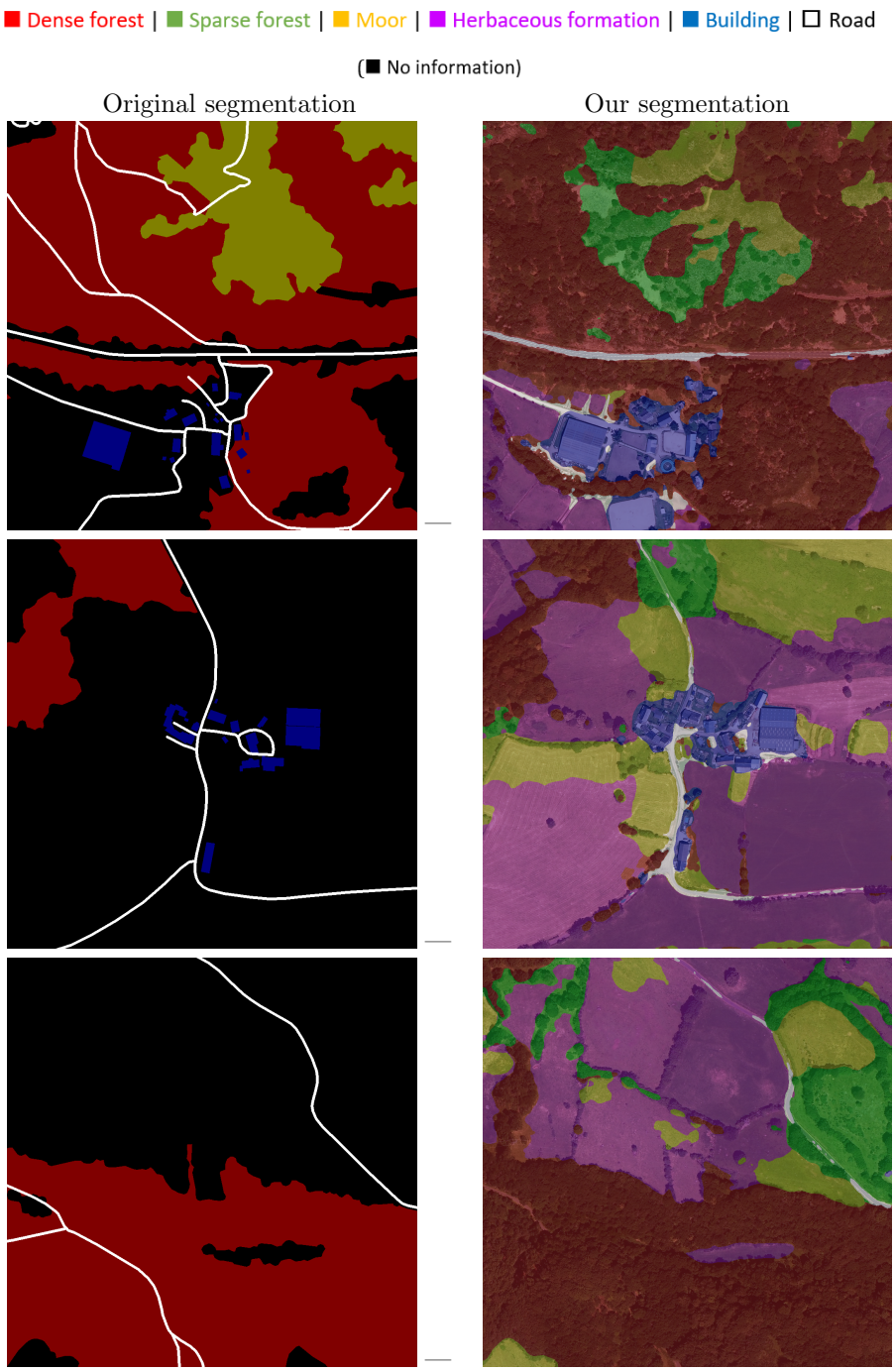


Figure 2: Cantal results

3.3 Information on the training

During the training, an ImageNet-22K pretrained model was used and we added weights on each class because the dataset was not balanced in classes distribution. The empirically chosen weights we have used are:

- Dense forest: 0.5
- Sparse forest: 1.31237
- Moor: 1.38874
- Herbaceous formation: 1.39761
- Building: 1.5
- Road: 1.47807

4 Experimental results

Backbone	Method	Crop Size	Lr Schd	mIoU	config	model
Swin-L	UPerNet	384×384	60K	54.22	config [4]	model[5]

Figure 1 shows some comparison between the original segmentation and the segmentation that has been obtained after the training (Hautes-Alpes dataset).

We have also tested the model on satellite photos from another French department to see if the trained model generalizes to other locations. We chose Cantal and a few samples of the obtained results can be seen in figure 2. These latest results show that the model is capable of producing a segmentation even if the photos are located in another department and even if there are a lot of pixels without information (in black), which is encouraging.

4.1 Limitations

As illustrated in the previous images, the results are not perfect. This is caused by the inherent limits of the data used during the training phase. The main limitations are:

1. The satellite photos and the original segmentation were not made at the same time, so the segmentation is not always accurate. For example, we can see in figure 3 a zone is segmented as "dense forest" even if there are not many trees (that is why the segmentation after training, on the right, classed it as "sparse forest").

[4] `configs/swin/config_upernet_swin_large_patch4_window12_384x384_60k_ign.py` on the repository

[5] `https://drive.google.com/file/d/1EarMOBHx6meawa6izNXJUfXRCTzhKT2M/view`

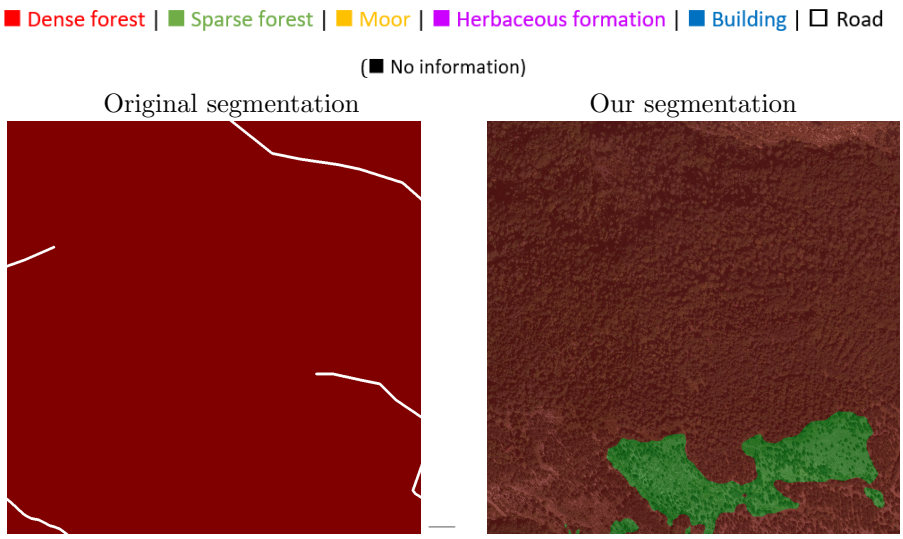


Figure 3: Example of limitation

2. Sometimes there are zones without information (represented in black) in the dataset. Fortunately, we can ignore them during the training phase, but we also lose some information, which is a problem: we thus removed the tiles that had more than 50% of unidentified pixels to try to improve the training.

3. Road segmentation is not accurate because sometimes the information is not visible in the image (hidden by trees for example), which obviously prevents it from being detected.

5 Repository

The source code, which is only a fork from the implementation of Swin Transformer can be found on github together with usage details[6].

6 Acknowledgments

This work has been funded by the ANR, project Ampli ANR-20-CE23-0001.

[6] `https://github.com/koechslin/Swin-Transformer-Semantic-Segmentation`

References

[1] MMSegmentation Contributors. Mmsegmentation, an open source semantic segmentation toolbox. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[2] 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.