Landfills segmentation with OpenStreetMap, Convolutional Neural Networks and Transformers

Remote Sensing Data Project Report

El Mallah Rim and Bertille Temple
Ecole Normale Supérieure Paris Saclay
4 Av. des Sciences, 91190 Gif-sur-Yvette
elmallahrim@gmail.com and templebertille@gmail.com

April 18, 2024

Abstract

Starting from OpenStreetMap's data, this report details the development of a deep learning framework aimed at generating clean binary masks of landfills from satellite imagery. We outline a pipeline querying images from Sentinel-2 and transforming the GeoJSON polygons into masks serving as ground truth for model training. In particular, we use the UNet model and the Segment Anything Model (SAM) to assess their performance in segmenting landfill areas with an objective of contributing to the broader goal of environmental monitoring.

1 Introduction

OpenStreetMap (OSM) is a collaborative project to create a free editable map of the world. It is constructed by a community of mappers who contribute and maintain data about objects of interests as for example roads, trails or landfills all over the planet. The data is obtainable in the form of a Geojson file, which comprises lists of polygons, detailed as a list of geo-referenced vertices. Landfills are sites for permanent or long term storage of waste materials and mapping them is a crucial ecological issue for monitoring purposes. Pictures of landfills might have a high variability of representation as the materials are sometimes buried or covered, sometimes simply piled. Landfills masks are raster binary images which associate to every pixel of an image if there is a landfill or not, they can be created from the OSM polygons. While OSM landfills data represent a valuable source of information to monitor landfills, they might lack precision or might be incorrect due to the participative nature of the platform. The goal of this project is to generate landfills masks with state of the art Deep Learning(DL) methods

and to compare the hand-made masks from OSM and the masks generated by the DL models. To feed the models, we need satellites or aerial images from which masks will be generated: Sentinel2 (S2) images constitute a precious source of multi-spectral images with 13 bands and a resolution of 20 meters by pixel. In France, Very High Resolution (VHR) images from the BD Ortho maintained by the National Institute Geographical and Forestry data is also a very good source of images with a resolution of 0.20 cm by pixel. While the Ortho images have much better quality, they cover only France territory whereas S2 images has the advantage to cover all the planet. As the project is not rooted in one specific country, and that landfills monitoring is a crucial issue wordwilde, we decided to keep an international opening and to work with S2 images.

The project can be divided into 2 parts: first the images and masks will be generated form OSM polygons geojson using S2 images following a pipeline described in Section 4.2, we will consider the masks generated as the ground truth masks. Then we will evaluate the ability of various DL models to automatically segment landfills from the dataset of S2 images and OSM masks built at the previous step.

2 Related Work

This section provides an overview of the current state of research in the areas of landfills identification and segmentation, and DL models used for the task of semantic segmentation.

2.1 The different tasks related to Landfills

The approaches proposed in the literature can be grouped into different categories as mentioned in a survey about solid waste detection in Remote Sensing images [4]. Despite the common goal of detecting and monitoring solid waste disposal sites, they tackle different problems that are the following : landfill or sparse waste detection, landfill sites monitoring, landfill sites detection and monitoring, illegal dumping distribution characterization, mapping of areas with a high risk of illegal waste dumping, waste heat contamination monitoring, identification of subsurface fires within landfills and assessment of suitable landfill locations.

In our project, we specifically focus on the landfill sites detection and monitoring, with a bigger focus in creating binary masks representing the different landfills sites around the world.

2.2 Landfills in Remote Sensing

Historically, in the domain of environmental monitoring using remote sensing data, early studies employed manual interpretation of aerial images to detect waste sites. Pioneering studies such as those conducted in 1974 [5] highlighted the utility of aerial photographs for mapping the spatial distribution of waste. The evolution of these methodologies saw the integration of GIS and multispectral data, notably by researchers in

Italy, who used such data to identify stressed vegetation indicators associated with uncontrolled landfills [10].

Further advancements in the field transitioned towards machine learning approaches using Support Vector Machine(SVM) and Random Forest(RF), with notable works categorizing multispectral images for waste segmentation, achieving significant accuracy levels [9].

Recent years have seen a paradigm shift with the advent of DL techniques in remote sensing, especially in landfill detection. Rajkumar et al.[1] got their best results using a UNet-ResNet34 architecture pretrained on the ImageNet [2] and SpaceNet [3] datasets. The model was applied to annotated satellite images that we could not manage to find.

In 2023 the Segment Anything Model (SAM) proposed in [7] has achieved remarkable scores in the task of segmentation. It has not yet been applied to landfill segmentation due to its relative newness, but its impressive performance suggests it could match or surpass UNet performance.

3 Methods

Semantic Segmentation is a computer vision process where each pixel in an image is classified into a specific category. Binary Segmentation is when we only have two categories which will be the case in our project. We will evaluate the performance of the traditional UNet model, the non-fine-tuned SAM and the fine-tuned SAM, in the context of landfills binary segmentation from satellite imagery.

3.1 UNet

UNet has emerged as a sucessfull method in [8] to segment neuronal structures in electron microscopic recordings. This CNN architecture shown in Figure 1 has symmetric down-sampling and up-sampling paths with skip connections. The up-sampling path and skip connections allow recovering the precise pixel location, essential in a task of pixel-wise classification.

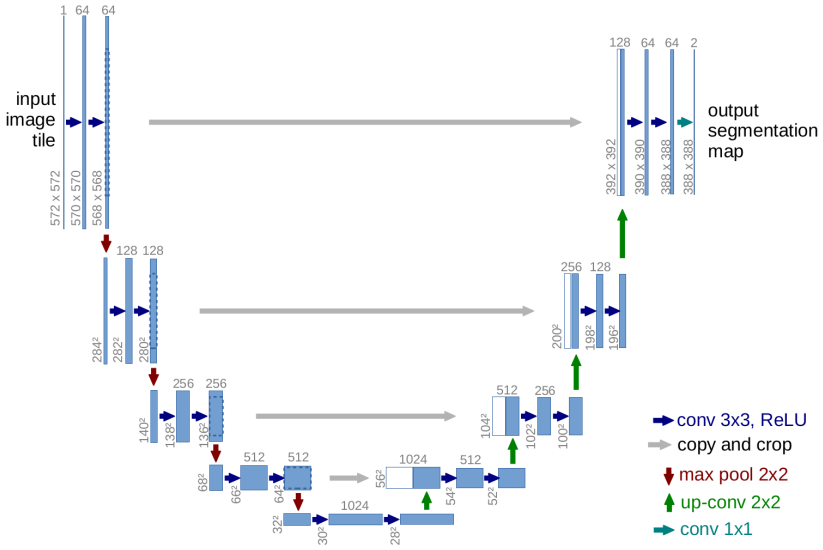

Figure 1: UNet architecture (Taken from [8]

It can be used with transfer learning by importing the pre-trained model RESNET originally developed for image recognition in [6]. Transfer learning uses the knowledge (features, weights, and biases) that a model has learned from a large and comprehensive dataset to apply it to a different but related problem.

3.2 Segment Anything Model (SAM)

By leveraging transformer-based architectures, SAM has been designed to segment an extensive range of objects without the need for extensive dataset-specific fine-tuning.

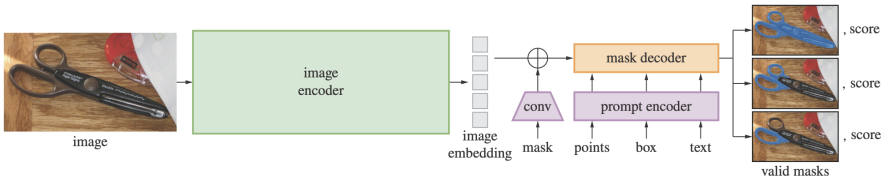

Figure 2: Segment Anything Model (SAM) overview (Taken from [7]

SAM's architecture SAM has three components, as can be seen in Figure 2: an image encoder, a prompt encoder, and a mask decoder.

1. **Image Encoder:** It is a pre-trained Vision Transformer (ViT). It processes the entire image once per instance and embeds it for subsequent segmentation tasks.

2. **Prompt Encoder:** It handles both sparse (points, boxes, text) and dense (masks) prompts. Sparse prompts are integrated using positional encodings combined with learned embeddings, while dense prompts use convolutional embeddings.

3. **Mask Decoder:** It maps the combined embeddings to segment masks using a Transformer decoder block, it incorporates self-attention and cross-attention mechanisms.

4 Experiments

The code can be found at `https://github.com/BertilleT/Landfill_segmentation`

4.1 Data location

In total 171 countries are represented in the datase but for the sake of clarity, all countries are not represented in Figure 3. Only 4.6% of the landfills are located in France, which represents 2296 polygons. The majority of the provided data are polygons located in Russia, Germany, Ukraine etc. One key question here is how landfills appearance is related to the location: for example landfills from Norway might have snow on them compared to landfills from Burkina Faso. Landfills semantic segmentation might be a task which depends of the domain. However, we decided to use all the data available, whatever the source location, as after pre-processing steps (as filtering low resolution S2 images for example), we obtained a dataset small enough. We think that reducing it even more by selecting only images from a specific country would have damage the results significantly.

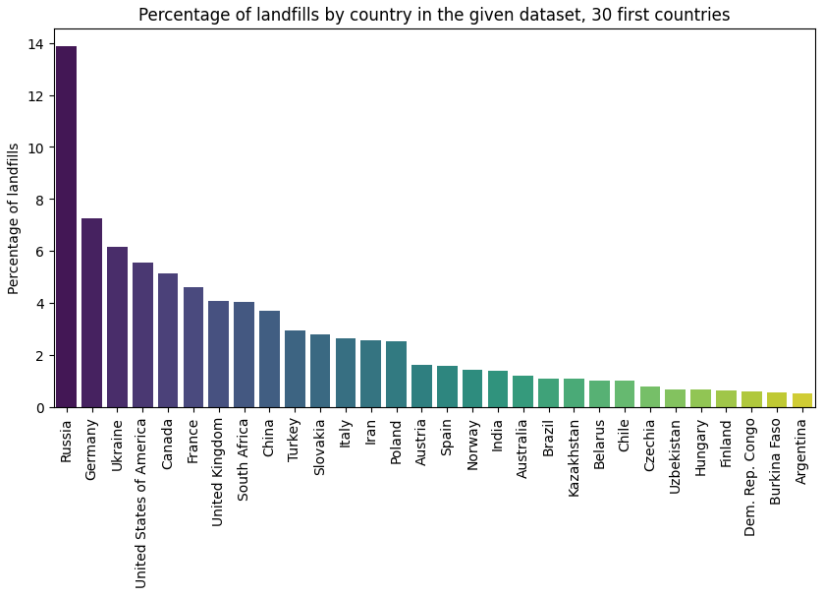

Figure 3: Location of the landfills from the geojson file provided

4.2 Pipeline to generate OSM masks

Having a geojson file with polygons correponding to landfills around the world, we needed to generate ground-truth masks in order to be able to use them fro training our models as well as comparing the results.

We will present in this section all the pipeline that creates these masks.

4.2.1 Query the tif format images

We reused the code from the TP1 as a starting point, we retrieved from it all the functions needed to query the images and had to make some decisions for some parameters :

- We opted to keep the starting and ending dates from the 01/07/2023 to 12/12/2023 to have a large panel of 6 months but still getting relatively recent images.
- We decided to take a maximum cloud covering of 10% to have good quality images for the segmentation as the clouds would alter the results.
- For the cropping of the Area of Interest (AOI), we decided the final image to be twice the size of the bounding box of the polygon representing the landfill. For example, if the bounding box of the polygon's landfill is of size (w, h), the final image will be of size $(2w, 2h)$ so the polygon represents less than 25% of the image. This would ensure to have some surroundings of the landfill so that the models will have to effectively segment 2 categories (landfill and not landfill)

All the images at this step are in the tif format as this format contains the metadata of the localization. This will help us to construct the ground truth mask corresponding to the image.

4.2.2 Creation of the datasets

Now that we have the images in the tif format as well as the coordinates (longitude, latitudes) of the points of its corresponding polygon, we want to be able to create the corresponding ground truth mask. Two main functions where necessary for this task :

- First, we wrote a function that transforms the coordinates of the points of the polygon from longitude and latitude to the pixel coordinates in the image. The function `gdal_get_longlat_of_pixel` that does the contrary was already given in the code of the TP1.
  So we defined the inverse function `gdal_get_pixel_of_longlat` which supports conversion for images hosted online by handling URLs and leverages GDAL's tools (gdaltransform) for accurate CRS-based transformations (The CRS of the input GeoTIFF is determined from the metadata in the file).

- We have the coordinates of the vertex in the image's pixel coordinates, which means that the points defining the polygon are described as lists of [x, y] pixel coordinates, with each list representing a vertex of the polygon. Then we can write the function which constructs the binary mask. It starts with a blank mask array based on the provided image dimensions and then proceeds to fill the polygon area defined by these points with ones, while the area outside remains zero, using OpenCV's `fillPoly` function.

The final step was to crop the images and masks to have square images and resize them so we can have the same resolution for all images. We decided to first remove the images for which the resolution was too low, it is the case for the small landfills as the Sentinel-2 images don't have a very good resolution (10m). We thus removed all images that have less than 20000 pixels. Then we crop the width or height when it is the smallest value and crop the mask and image to a square centering as much as possible the polygon.

We apply the final resizing when loading the pytorch dataset using cv2 so it matches the input of the model (in our case it was 256x256 for SAM as well as UNet).

We had a total of approximately 50 000 landfills in the geojson. As querying the images of Sentinel2 took a lot of time ($\approx$ 1 hour per 300 images), we decided to do about 15% of the images. But a lot of images gave errors, mainly because there was no image with less than 10% of cloud covering in the requested period of time, so we ended up with approximately 5000 images. And after the final cropping to have images with a good resolution, the finished dataset was of 986 images.

We divided our dataset into train, validation and test with respectively 60%, 20% and 20% of the total dataset.

4.3 UNet

We implemented the UNet model with respect to the original architecture with 4 blocks to reduce the spatial dimensions of the input image while increasing the feature dimensionality (down sampling) and 4 symmetric up-sampling blocks to recover the spatial dimensions of the original.

4.4 SAM

For the experiments with SAM, we opted to first try to use the pretrained base model provided in the transformers package and secondly to finetune the model with our dataset.

As already mentioned in subsection 3.2, SAM can take as input a prompt along with the image, the different prompts are text, points or a bounding box. As the information we have are points that are the vertices of the polygon, we thought of trying to give as entry the points or the bounding box of the polygon.

The experiments we tried with the pretrained model are the following :

- Without any specific prompts, the pretrained SAM attempts to segment the entire image into multiple categories, which is not what we want for our focused task.
- When using the pretrained SAM model with the polygons' vertices as input, the results were unsatisfactory as the model interprets these input points as central elements rather than as boundary points, leading to incorrect segmentation areas.
- When using the pretrained SAM model the bounding boxes of the polygons as the input prompt, we got significantly better results. The model could accurately identify and segment the area of interest within the bounding box.

We thus decided to generate the masks with the pretrained SAM using the bounding boxes as prompt.

We also finetuned the model using also the bounding boxes as inputs. The only difference is that during the training we apply a random perturbation on the bounding boxes so the model is more robust, perturbation that we don't apply during the inference. The loss used for the training is the DiceCELoss which is a weighted average between the Dice loss and the Cross Entropy Loss.

We tried a small number of epochs (20) as well as a bigger number of epochs (100), the results obtained can be seen in Table 1.

5 Results

The following table outlines the settings for each model evaluated:

Model	Learning Rate	Optimizer	Criterion	Number of Epochs
UNet	10^{-5}	Adam	BCE	30
SAM (Fine-tuned)	10^{-5}	Adam	DiceCELoss	20/100

Table 1: Experimental settings for each model

As an evaluation metric, we used the Intersection over Union (IoU), it measures the overlap between the predicted segmentation and the ground truth, divided by the area of union between the predicted and ground truth segmentation. It can be expressed as:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

The results of our experiments, measured in terms of test IoU, are summarized in the following table:

Model	Test IoU
UNet	0.60
SAM (Pretrained)	**0.73**
SAM (Fine-tuned)	0.68

Table 2: Test IoU scores for each model

We could not find others study working with the same data as us, which makes our results not comparable with any previous work.

The best scores are obtained with SAM pre-trained, however note that the ground truth we used is not really reliable as they are made up of noisy labels derived from OSM. As a consequence, in order to check the quality of the predictions, it is more accurate to check visually the quality of the masks generated. Figure 4 shows that the best masks are obtained with UNet, because they are cleaner than the others, and seem to locate the more accurately the boundary of the landfill. OSM masks lack sometimes of precision to identify correctly the boundaries of landfill, and SAM masks have often fuzzy limits with some white pixels isolated.

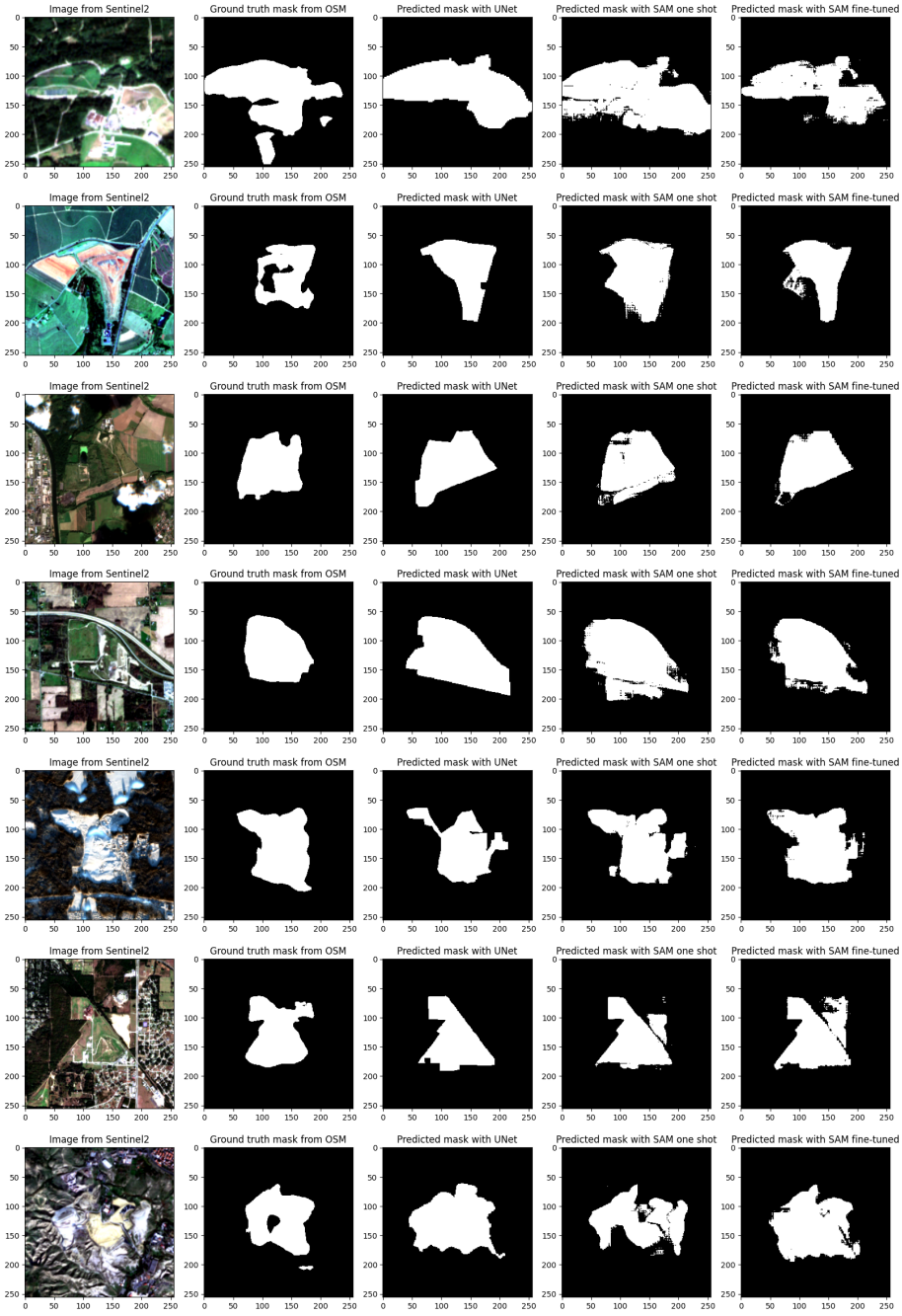

Figure 4: Comparison of masks generated from OSM, UNET and SAM

6 Conclusion

In conclusion, our work on the creation of binary masks through landfill segmentation using Convolutional Neural Networks and Transformers (particularly UNet and Segment Anything Model (SAM)) yielded promising results. The application of these models on satellite imagery from Sentinel-2, guided by OpenStreetMap data, demonstrated a good potential. However, the use of OSM data as the ground truth introduces uncertainties due to its variable accuracy and completeness and limits the evaluation of our results. This shows the need for a dataset verified by domain experts to rigorously evaluate the model performance.

Future work could be about using higher resolution imagery than the images of Sentinel-2 to enhance detection accuracy as well as using a broader dataset with expert-validated ground truths. Also, a study could be done to evaluate the benefits from transfer learning. Finally, the created masks can be used for other tasks such as change detection of these landfills contributing to more reliable environmental monitoring solutions.

References

[1] Tamas Sziranyi Andras Majdik Anupama Rajkumar. Detecting landfills using multi-spectral satellite images and deep learning methods.

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

[3] Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. Spacenet: A remote sensing dataset and challenge series, 2019.

[4] Piero Fraternali, Luca Morandini, and Sergio Luis Herrera González. Solid waste detection in remote sensing images: A survey, 2024.

[5] DONALD Garofalo and F Wobber. Solid waste and remote sensing. *Photogrammetric engineering*, 40(1):45–59, 1974.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

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

[9] Lungile Selani. *Mapping illegal dumping using a high resolution remote sensing image case study: Soweto township in South Africa*. PhD thesis, University of the Witwatersrand, Faculty of Science, School of Geography . . . , 2017.

[10] S Silvestri and M Omri. A method for the remote sensing identification of uncontrolled landfills: formulation and validation. *International Journal of Remote Sensing*, 29(4):975–989, 2008.