UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Multi-temporal Satellite Imagery Panoptic Segmentation of Agricultural Land in Ukraine

Author:
Marian PETRUK

Supervisor:
Dr. Taras FIRMAN

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Marian PETRUK, declare that this thesis titled, "Multi-temporal Satellite Imagery Panoptic Segmentation of
Agricultural Land in Ukraine" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Multi-temporal Satellite Imagery Panoptic Segmentation of Agricultural Land in Ukraine

by Marian PETRUK

Abstract

Remote sensing of the Earth using satellites helps analyze the Earth's resources, monitor local land surface changes, and study global climate changes. In particular, farmland information helps farmers in decision-making, planning and increases productivity to achieve better agro-ecological conditions. In this work, we primarily focus on panoptic segmentation of agricultural land, a combination of two parts: 1) delineation of parcels (instance segmentation) and 2) classification of parcel crop type (semantic segmentation). Second, we explore how multi-temporal satellite imagery data compares to a single image query in segmentation performance. Third, we conduct experiments using the recent advances in Deep Learning and Computer Vision that improve the performance of such systems. Finally, we show the performance of the state-of-the-art panoptic segmentation algorithm on the agricultural land of Ukraine, where the farmland market has just opened.

Acknowledgements

I wish to thank my supervisor, Dr Taras Firman, who has guided and encouraged me during the work on the thesis. Thank to Applied Sciences Faculty at the Ukrainian Catholic University and especially to Oleksii Molchanovskyi, who strategically ensured the work was planned and continued, despite the war. Thank to UCU ML lab, and in particularly Oles Dobosevych, for the provided GPU-accelerated computing server. Finally, I would like to mention the support and feedback from my friends and family.

Contents

List of Figures

List of Tables

List of Abbreviations

SITS	**S**atellite **I**magery **T**ime **S**eries
PASTIS	**P**anoptic **A**gricultural **S**atellite **TI**me **S**eries
PaPs	**P**arcel-as-**P**oints
CRS	**C**oordinate **R**eference **S**ystems
MSI	**M**ulti-**S**pectral **I**nstrument
S2	**S**entinel-**2**
PS	**P**anoptic **S**egmentation
MLP	**M**ulti-**L**ayer **P**erceptron
DL	**D**eep **L**earning
DNNs	**D**eep **N**eural **N**etworks
CNNs	**C**onvolutional **N**eural **N**etworks
ReLU	**Re**ctified **L**inear **U**nit
GELU	**G**aussian **E**rror **L**inear **U**nit
RGB	**R**ed-**G**reen-**B**lue

For Ukraine's victory and economic prosperity

Chapter 1

Introduction

Many countries have had open farmland markets for many years, which provides economic growth, investments, more jobs, and greater productivity. In the USA (Huete and Ponce, 2010), in France (Garnot and Landrieu, 2021), in China (Frolking et al., 1999), there are already developed country-specific methods to help farmers analyze the land using satellite imagery to make better decisions on how to grow crops more optimally. However, in Ukraine, such a market opened just in 2021. While remote sensing for farmland is a known problem and research direction globally, almost no existing works use Ukrainian cropland data. Therefore we aim to use the available best practices, hopefully improve these practices, and apply the result to Ukrainian farmland.

Open access to satellite imagery data has given researchers worldwide a new way to conduct various earth observation tasks. Remote sensing using space exploration technology helps to analyze Earth resources (Schmitt et al., 2020), to monitor local changes of the land (Chen et al., 2021) and to study global (Dubovik et al., 2021), *e.g.* climate changes (McDowell et al., 2015). 20 Terabytes of new data are generated every day just through European Space Agency's Sentinel 1-3 satellites (Tarasiou and Zafeiriou, 2021). Such big amounts of data inspire us to help analyze it with cutting-edge methods.

Land cover mapping using traditional classification methods (rule-based with hand-crafted features) is ineffective and less accurate (Yu et al., 2022; Tsagkatakis et al., 2019). Deep learning (DL) has been recently introduced to classify land cover utilizing multi-scale automated feature extraction (Dubovik et al., 2021). The DL-based approach has shown significant potential for high resolution, multi-spectral, multi-temporal satellite images (Yuan et al., 2020).

In our work, we use Deep learning methods for panoptic land-cover segmentation focusing on agricultural fields (parcels). Information about crop type, field contours, field quality change in time and even prediction of how a particular crop will grow in the cropland is a must for the 21st-century agricultural business.

Chapter 2

Related Works

2.1 Remote Sensing

Satellite remote sensing is a method of surveying and analyzing information about the Earth from space. Satellite imagery is possible with satellites, which fly in the Earth's upper thermosphere and lower exosphere. For example, popular Sentinel-2 satellite orbiting at mean altitude of 786 km above Earth and sensing land surface in range of 443 nm - 2.19 µm electromagnetic spectrum (Fig. 2.1).

In Fig. 2.1, one may see the percentage of atmospheric opacity at different electromagnetic wavelengths. When opacity is at 100 % it is impossible to sense any information. There are some wavelengths at which atmospheric opacity is less than 5-10 %, and it is very convenient to observe the Earth at such ranges (Ashraf, Maah, and Yusoff, 2011). Those are wavelength that penetrates Earth's atmosphere: 1. infrared and visible light spectrum from 400 nm to ≈ 60 µm, and 2. radio spectrum from ≈ 1 cm to ≈ 20 m wavelength. Satellite Remote Sensing of the Planet Earth from space is used in many fields of science and supports the daily functionality of humans, *e.g.* predicting the weather (Dubovik et al., 2021). This instrument helps to monitor environmental pollution, climate change, ocean currents changes, sea level, land cover, exploration of mineral resources (Zhang et al., 2017a), soil moisture (Lakshmi, 2013), vegetation (Xie, Sha, and Yu, 2008), deforestation, forest fires (Lentile et al., 2006) and many more.

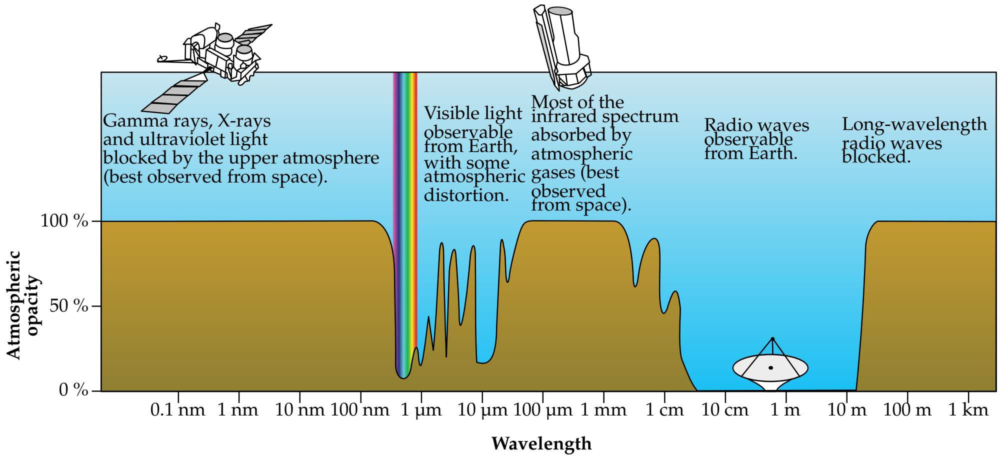


FIGURE 2.1: Electromagnetic transmittance, or opacity, of the Earth's atmosphere. (Illustration Source: Wikimedia Commons)

2.2 Panoptic Segmentation

Image Segmentation is a computer vision problem that aims to assist in image analysis or in scene understanding (Minaee et al., 2020). Panoptic segmentation (PS, Fig. 2.2) is a novel task introduced by (Kirillov et al., 2019) which rises in popularity. PS combines two segmentation problems: 1. semantic segmentation (assigning class labels to each pixel) and instance segmentation (detecting and segmenting each object instance). As the authors of the unified task mention, this type of segmentation is one crucial step toward automated real-world vision systems. Previous methods dealt with instance and semantic segmentation separately.

There are different methods to solve panoptic segmentation: Panoptic Feature Pyramids (Lin et al., 2016) or attention-aided networks (Li et al., 2018) for scene panoptic parsing. However, the growth of the task is limited due to the lack of diverse panoptically annotated datasets, as it is with image recognition or semantic segmentation tasks because data for PS is much more costly and time-consuming to annotate.

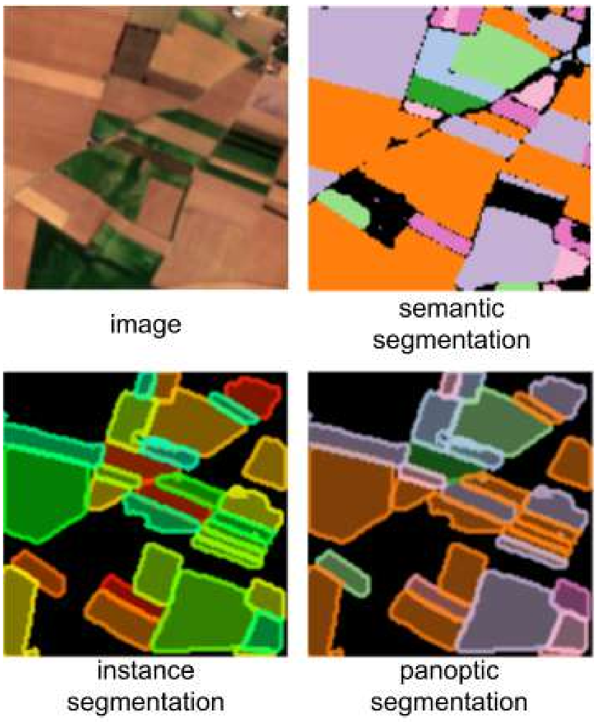


FIGURE 2.2: Vizualization of panoptic segmentation task.

2.3 Segmentation of Agricultural land

Panoptic segmentation (PS) for crop mapping has only been introduced in remote sensing research. It is an exciting topic to explore and challenging to attempt applying this method for the Ukrainian land where the farmland market has just opened, and very few works are published. Hardly any are reproducible.

To the best of our knowledge, PS of agricultural land was first introduced by Garnot and Landrieu, 2021. This work conducts panoptic segmentation of agricultural parcels on French land.

In our work, we are primarily interested in finding the edges of parcels in order to make them distinct, to be able to discriminate one crop field from the other and, second, to classify the parcel type. Such a problem can be formulated as a panoptic segmentation, outlined in Sec. 2.2. It consists of assigning a class and a unique instance identifier to each pixel.

Garnot and Landrieu, 2021 work is very related to our topic of interest. Their work argues that one should address complex temporal patterns of crops with temporal sequences of images to achieve higher segmentation accuracy. Garnot and Landrieu introduced the first end-to-end method for panoptic segmentation of Satellite Image Time Series (SITS). In Fig. 2.3 one may see the high-level visualization of the proposed method: using a sequence of satellite images (time series) output panoptic segmentation labels. Authors claim that having one image, it is impossible to accurately segment and classify parcels due to the temporal nature of the crop. In this research work, we will verify this statement.

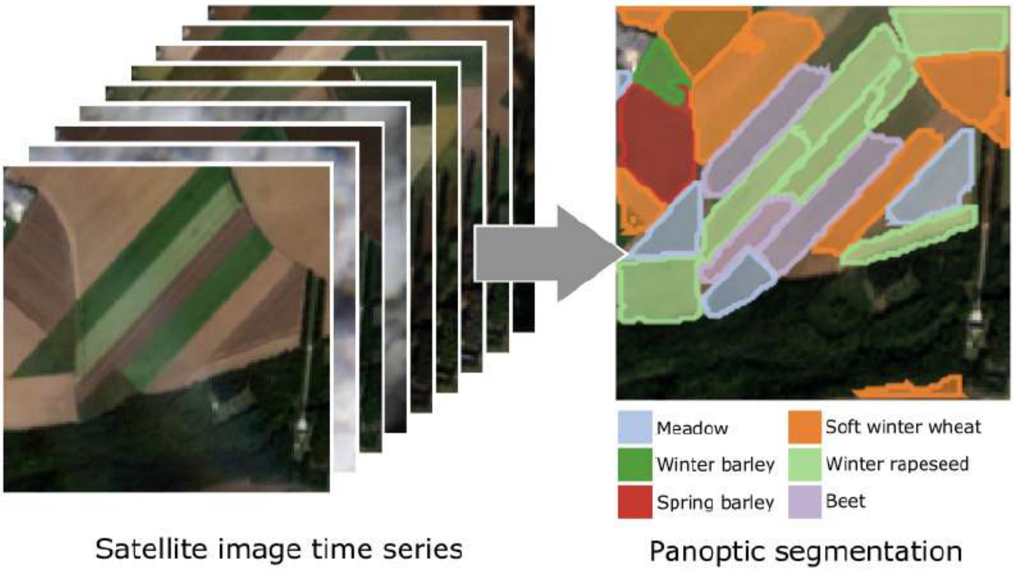


FIGURE 2.3: Diagram of Satellite Image Time Series panoptic segmentation: parcel edges and classes. (Fig. 1 in Garnot and Landrieu, 2021)

The authors developed one end-to-end network consisting of two modules: 1) a Spatio-temporal encoder, a feature extractor, and 2) a panoptic segmentation network that inputs extracted features and produces panoptic segmentation masks. Garnot and Landrieu discuss the limitation of their approach, which is the geography of the dataset they used and developed their solution. The authors mention that approach is suitable for the same meteorological context, terrain condition and farmland crop types. They also suggest that further work may be in advanced satellite image preprocessing, *e.g.* adding speckle filtering, elevation information, meteorological data. In our research work, we want to explore how we can adopt such an approach to Ukrainian terrain, its crop fields and weather context.

Kussul et al., 2017 proposed a CNN-based approach to multi-temporal satellite land cover and crop type classification. The authors discuss that the most common approach in remote sensing prior to neural networks was a random-forest approach (Belgiu and Drăguţ, 2016). The authors compared that approach to the CNN-based and showed how CNN could improve segmentation quality. They used Landsat-8 and Sentinel-1A images to classify wheat, maize, sunflower, soybeans and sugar beet, with reported 85% accuracy in the Kyiv region, Ukraine, in 2015. The authors note that an essential factor in satellite images is the presence of clouds which needs to be addressed in the data-processing stage. However Vivien and Loic, 2021 believe that advanced deep-learning methods should learn how to tackle cloud coverage with multi-temporal SITS.

Unfortunately, the Kussul et al., 2017 did not publish the dataset or code. Therefore it is impossible to verify their approach or use the data to compare with our approach. Kussul et al. is the first work that we discovered that tries to classify land in Ukraine. However, they do only classification, *i.e.* semantic segmentation, while we aim to conduct panoptic segmentation, delineate parcels additionally, and assign identifier labels to each field.

Rußwurm et al., 2019 presented a method for early classification of crop type before the end of the vegetative period. One can augment their method into existing classification models with an additional stopping probability based on previously seen satellite data.

Zorzi, Bittner, and Fraundorfer, 2020 noticed that it is hard to keep the up-to-date cadastre maps and problematic to trace new buildings or old destruction. In addition, the authors observed that the current cadastre data have inconsistencies and errors in the form of misalignment (Fig. 2.4). They propose a method to solve this problem with DL to correct label noises and misalignments. We believe their approach may be applied to the agricultural area as well, as, in Ukraine, we also have a cadastre land map.

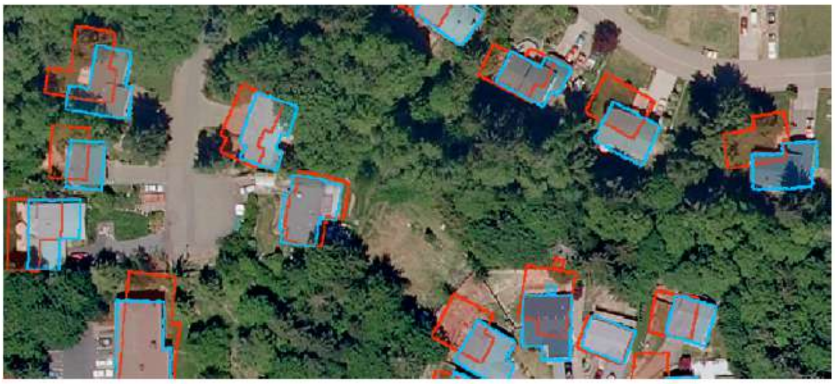


FIGURE 2.4: Results from Zorzi, Bittner, and Fraundorfer, 2020. MapRepair result. Misaligned annotations in red, corrected map in cyan.

Zhao et al., 2017 proposed a method to conduct semantic segmentation for street scene parsing using a pyramid pooling module. Their work uses multiple scales of extracted features by CNN to get local and global context information of the input image. In 2015 they achieved state-of-the-art on many public benchmark datasets. Such a pyramid module may help improve segmentation accuracy in satellite images. Even recent Multi-scale Vision transformers use the pyramid-based approach. Fan et al., 2021. We would also want to consider multi-scale, multi-level feature extraction for panoptic segmentation.

Vaze et al., 2020 compare different methods of leveraging multi-band information from satellite images with CNNs. They show that the standard selection of bands in the industry leads to worse performance than other methods. Authors compare band selection by an expert; all available bands, learning attention maps; and using Bayesian optimization to make the selection. Results show that compared to standard band selection, using all bands for CNN improves the test-time performance by 3% and using Bayesian optimization further boosts accuracy by 5.4% in total improvement. In our work, we would experiment with different bands selection.

Guérin et al., 2021 discuss that satellite images at different places were not made at the same time, and this influenced the segmentation quality. One may address this problem with multi-temporal satellite data (time-series).

Another aspect of the work is their unique motivation: "Virtual worlds in the context of digital entertainment need to be vast and realistic. In the context of the ANR project Ampli, we aim to make the task of virtual worlds authoring easier by providing a way to segment satellite images into six basic landcover classes.". Authors did their work for Ampli Anr Project, which is learning and inverse procedural modelling for authoring large virtual worlds. Recently virtual worlds became popular in the media, and big technology companies started developing such virtual worlds. It is another application of satellite imagery segmentation in the entertainment domain.

Chapter 3

Approach

In this chapter, we first define the Problem setting, second, we describe the Data sources and Data engineering stage, and finally, the modelling stage. We provide an overview of the model architecture, objective function, and metrics used.

3.1 Problem Setting

Panoptic segmentation may be considered as a combination of semantic segmentation, *i.e.* classifying each pixel on the image, and instance segmentation (delineation of all individual instances, assigning each pixel a unique identifier). In the context of satellite imagery, each semantic class is a particular type of crop on a parcel, *e.g.* corn, barley, wheat or sunflower. While instance label represents a unique parcel, marking its contour on the Earth's area of interest. Therefore, while semantic segmentation classifies only image pixels, our problem of panoptic segmentation is advanced with identifying unique objects on the image (instance segmentation).

There are several main reasons why this problem is challenging. First of all, the satellite data is multi-channel. The data which we will use is captured by the Sentinel-2 satellite. Those images have 13 spectral bands, each of which has a different resolution (See Fig. 3.1).

Second, the data is multi-temporal. Multiple works have shown that in order to obtain high accuracy, one has to use not a single satellite image but satellite image time series (SITS) for one particular region (Kussul et al., 2017; Garnot and Landrieu, 2021).

Third, a crucial aspect of the problem is the cloud cover. As we want to work with multi-temporal data, there will be clouds, and some parts will not be fully visible. Therefore, cloud occlusion is another step one must tackle in the data-prepossessing part.

3.2 Data

3.2.1 Sentinel-2 Satellite

Copernicus European Space Agency (ESA) conducts seven Satellite Missions, named Sentinel Missions, for earth observation tasks. Most popular in the research community for land surface analysis are Sentinel 1, 2 and most recently, 3. This work focuses on Sentinel 2 Mission due to its high resolution and optimal revisit period. Sentinel-2 (S2) consists of two satellites operating in a twin configuration. Each twin spacecraft carries a single Multi-Spectral Instrument (MSI) payload. MSI provides access to 13 spectral bands (from Visible, to near infrared to shortwave infrared) with varied resolution of 10m / 20m / 60m (Table 3.1). S2 has a high revisit time

Band	Used	Description	Wavelength (nm)	Spatial Resolution		
				10 m/pixel	20 m/pixel	60 m/pixel
B1		Ultra blue (Coastal and Aerosol)	443			x
B2	+	Blue	490	x		
B3	+	Green	560	x		
B4	+	Red	665	x		
B5	+	Visible and Near Infrared (VNIR), Vegetation red edge	705		x	
B6	+	Visible and Near Infrared (VNIR), Vegetation red edge	740		x	
B7	+	Visible and Near Infrared (VNIR), Vegetation red edge	783		x	
B8	+	Visible and Near Infrared (VNIR)	842	x		
B8A	+	Visible and Near Infrared (VNIR)	865		x	
B9		Short Wave Infrared (SWIR), Water vapour	940			x
B10		Short Wave Infrared (SWIR)	1375			x
B11	+	Short Wave Infrared (SWIR)	1610		x	
B12	+	Short Wave Infrared (SWIR)	2190		x	

TABLE 3.1: Spectral bands of Sentinel-2

of 5 days at the equator. The data it captures is available via Copernicus Open Access Hub (https://scihub.copernicus.eu/), which is free and open to all. Data is incapsulated into elementary granules of 100 × 100 km^2 tiles (ortho-images) that covers earth (Fig. 3.1). An example of such granule one may see in Fig. 3.2 extracted and vizualized as RGB image (10980 × 10980 px. with resolution of 10 m/pixel).

To conclude, Satellite imagery is significantly different from commonly used RGB HD pictures. They have more pixels: 10^4 vs $2 \cdot 10^3$ for width and height (W, H); 10-13 channels vs 3 RGB channels (C). Additionally, we will use temporal-dimension: stacking a 3-dimensional array in time (T), forming $T \times C \times W \times H$ tensors.

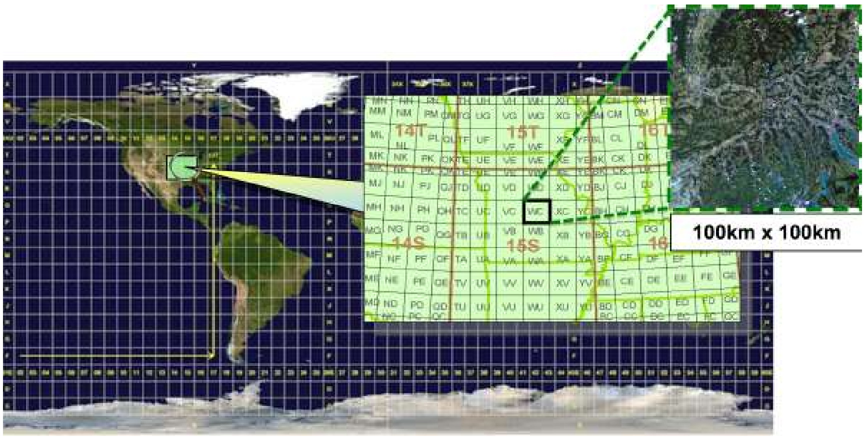


FIGURE 3.1: Granule tiling vizualization.

3.2.2 PASTIS Dataset for France

Panoptic Agricultural Satellite TIme Series (PASTIS), created by Vivien and Loic, 2021, is one of the best datasets for our problem of interest. The dataset is developed for SITS segmentation with panoptic labels of parcels for ca. 4000 km^2 of France territory. One may see in Fig. 3.3 the location of four clusters, each cluster is divided into hundreds of 128 × 128 px patches, 2433 patches in total. Each patch consists of satellite time series with varied temporal lengths, from 33 to 61 timestamps. Timestamps dates are in the range from September 2018 to November 2019. Each satellite timestamp has 10 Sentinel-2 spectral bands data. The authors selected all bands except the atmospheric bands B1, B9 and B10 (Table 3.1). Data is not filtered by cloud percentage cover because authors believe that DL algorithms should be able to learn

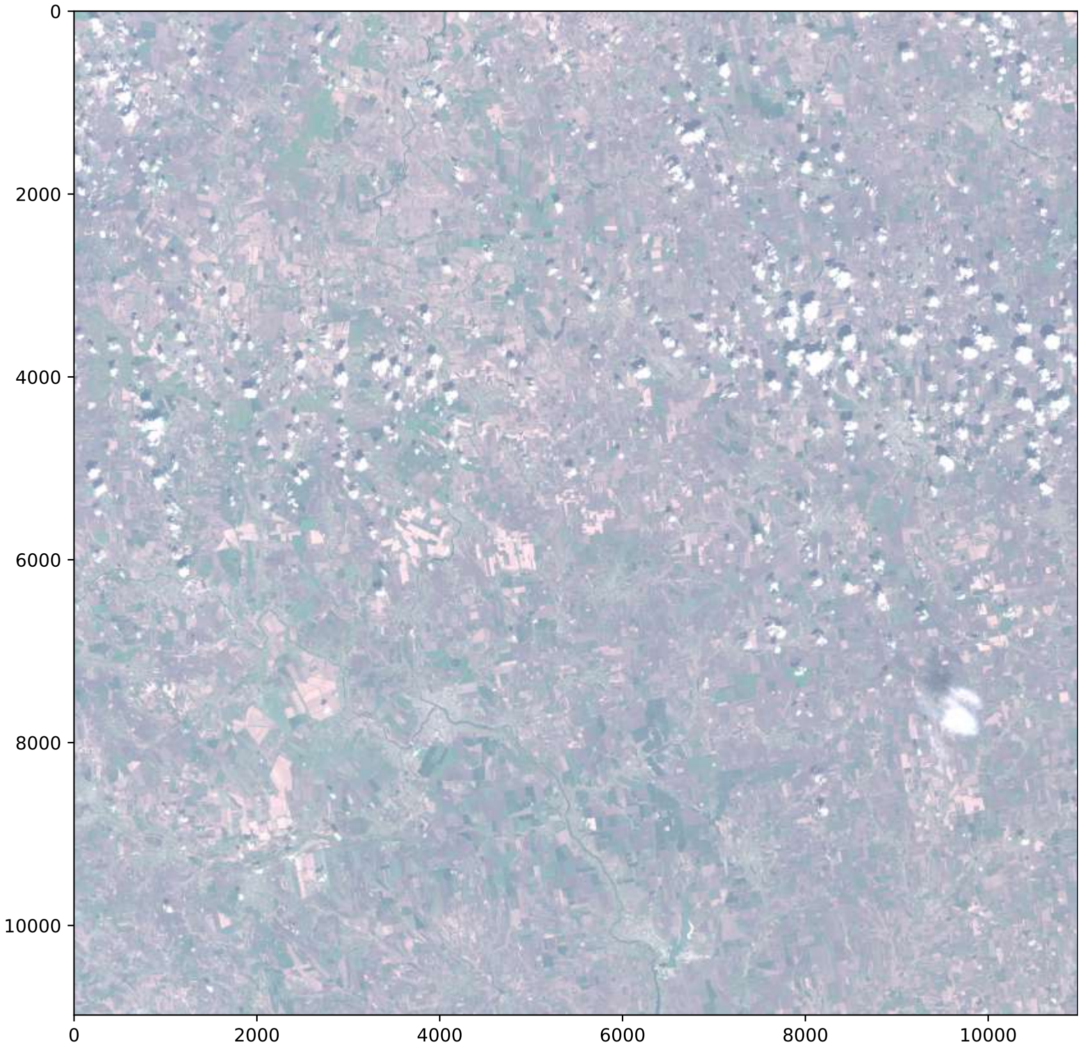


FIGURE 3.2: An example of Sentinal-2 granule.
VIzualized as RGB image (B4 + B3 + B2 band composition).
Data volume of 13 bands is 800 MB in size and covers 100 × 100 km^2.

to be robust to cloud occlusions. Each patch SITS has appropriate annotations of parcel instances and the crop type for each field (Fig. 3.4).

Overall there are 18 crop types annotated with 10 m/pixel resolution, totalling 2 billion pixels. Authors made the dataset available via https://github.com/VSainteuf/pastis-benchmark.

Training-validation-test split was done by randomly splitting patches into five splits (1, 2, 3 - for training, 4 - for validation and 5 - for testing), forming five different folds to allow cross-validation. However, we are using only the first fold in our experiments due to computing time-to-train limitation. The split distribution is the following: train 1455, validation 482, and test 496 patches. The authors ensured that adjacent patches do not appear in different folds to avoid data leakage.

In our research, we would consider the PASTIS dataset our initial main dataset due to its rich annotations, large area, and European region in which we are most interested. Using the aforementioned dataset, we would also have a benchmark for comparing our experiments. As the dataset has a benchmark-leaderboard page at `https://paperswithcode.com/sota/panoptic-segmentation-on-pastis` where anyone can submit his proposed advancements.

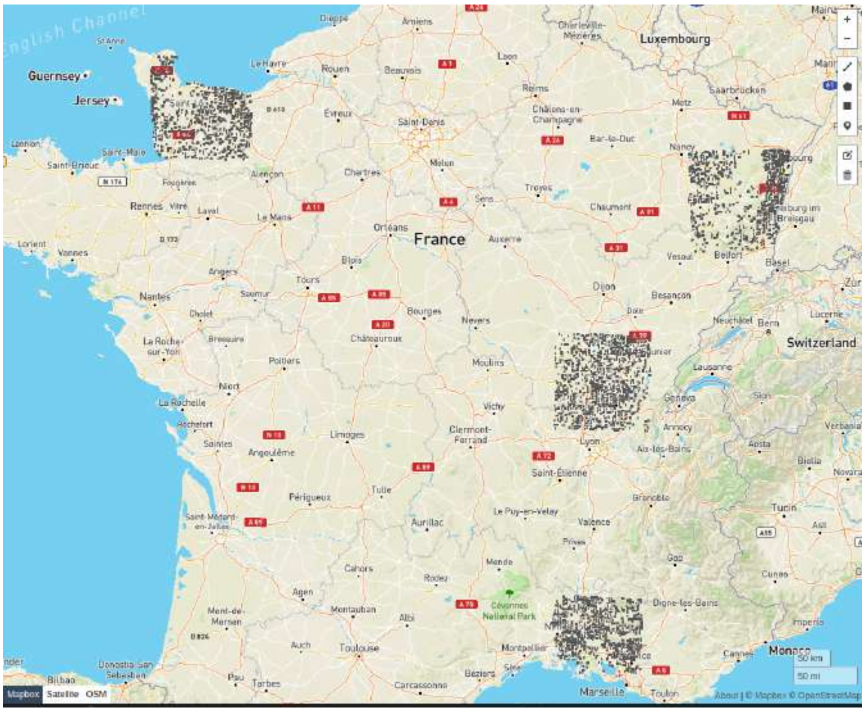


FIGURE 3.3: PASTIS dataset regions location (outlined in four black polygon-scatter plots).

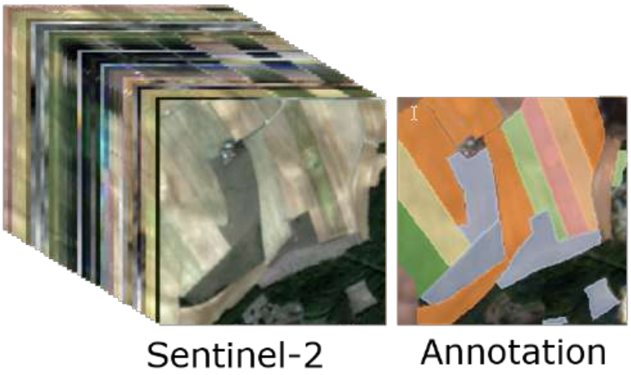


FIGURE 3.4: PASTIS Sentinel-2 satellite 10 optical bands data. (Fig. 4 in Garnot, Landrieu, and Chehata, 2021)

3.2.3 Dataset for Ukrainian territory

Since we wanted to focus on the panoptic segmentation of agricultural land in Ukraine in our research, we ought to have a dataset with similar characteristics as PASTIS for France. To the best of our knowledge, there are no such datasets with instance and semantic labels published for Ukraine. We believe that having such a dataset can open many opportunities for farmers and businesses through Data Science applications. Hopefully, this will change in nearest future, and government or research institutions will publish such a dataset. Since the land market has recently opened in Ukraine, the lack of such a dataset blocks speeding up the market growth.

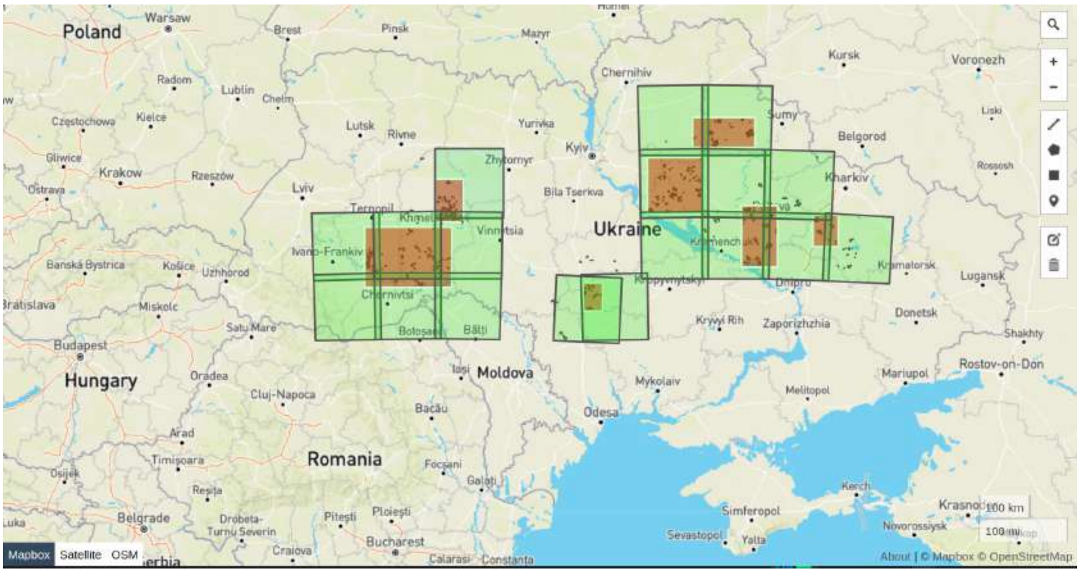


FIGURE 3.5: Data regions.
Red - bounding boxes of clusters. Green - downloaded Sentinel 2 tiles.

3.2.4 Data Engineering of Ukrainian cropland data

Thankfully, one businessman, a farmland owner, who decided to remain anonymous, provided us with some of his raw archival spreadsheets (xlsx format) and cropland polygon map (kmz format).

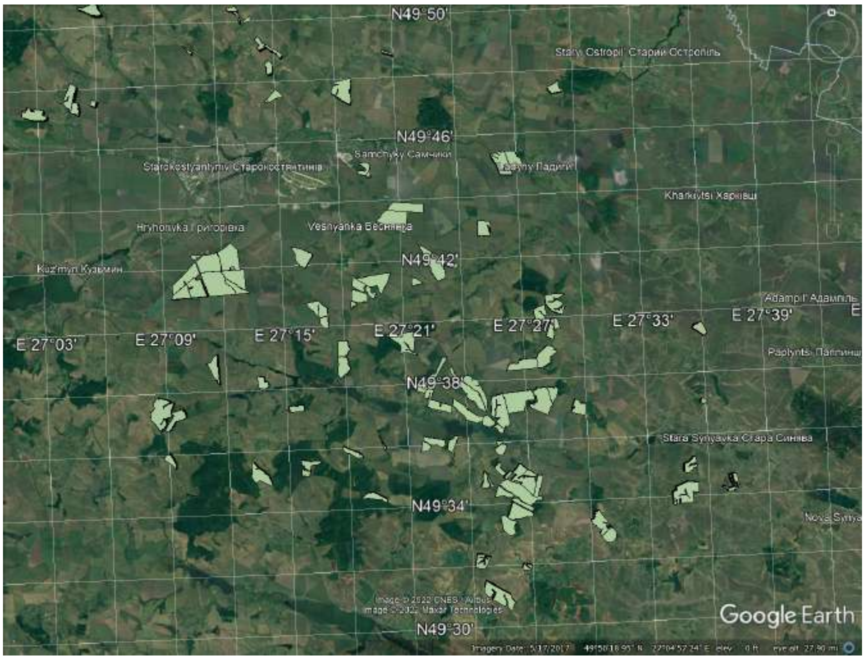


FIGURE 3.6: An example region of provided polygon map as a source of data, plotted on Google Earth satellite layer for visualization.

Therefore, following PASTIS dataset guidelines by Vivien and Loic, 2021, we processed the data to form a PASTIS-like structure for Ukrainian data. Here we describe the Data Engineering steps we performed during the processing, since the procedure requires multiple steps with attention to detail and has many nuances in the geography domain. First, for exploratory data analysis, we found that data has ca. 700 parcel polygons scattered throughout Ukraine. We sampled the parcels with crop types similar to the biological taxonomy of 18 classes in PASTIS. After mapping, we ended up with six crops: Winter rapeseed, Corn, Soybeans, Sunflower, Soft winter wheat, and Leguminous fodder. The provided data consisted of a very narrow list of crops compared to 18 in PASTIS.

When we filtered the data, we had annotations of parcel contours (instance) and crop type (semantic) information for the 2018 crop-yield year, however, in XML-like kmz format. Regarding satellite imagery, since we couldn't download all the parcel location Sentinel-2 tiles due to storage limitations and annotation processing time, we decided to group the data. So that each tile region we commit to downloading will be most optimal for us, *i.e.* will cover a maximum number of parcels within the bounding box. To achieve this, we used the DBSCAN algorithm from the Scikit-learn library by Pedregosa et al., 2011 to cluster all scattered parcels into groups. We find DBSCAN algorithm (Schubert et al., 2017; Ester et al., 1996) is appropriate here because it finds core samples of high density and expands clusters from them by measuring the distance between instances in the feature vector. Our data had parcel centre coordinates included as (latitude, longitude) (World Geodetic System 1984, used in GPS, EPSG:4326 projection) (Slater and Malys, 1998). In the algorithm, we used the haversine formula by converting Lat-Long coordinates degree into radians to compute great-circle distances between parcel centre points. Knowing that Earth's radius is $r \approx 6371$ km, we set the maximum distance between two samples, to be considered as neighbours, as 30 km ($\epsilon = \frac{\text{30 km arc length}}{\text{6371 km Earth's radius}} \approx 0.0047$ rad, central angle) and set the minimum number of samples in a neighbourhood for a point to be considered as a core point as 30 parcels. The result of the DBSCAN one can see in Fig. 3.7, where we scattered all parcels as black polygons and coloured ones that are in a specific cluster. Overall, we obtained 7 clusters with minimum density of 30 parcels in 30 km neigbourhood. For final step in clustering we found bounding boxes coordinates for each of the cluster with 3 km padding (to cover all parcels).

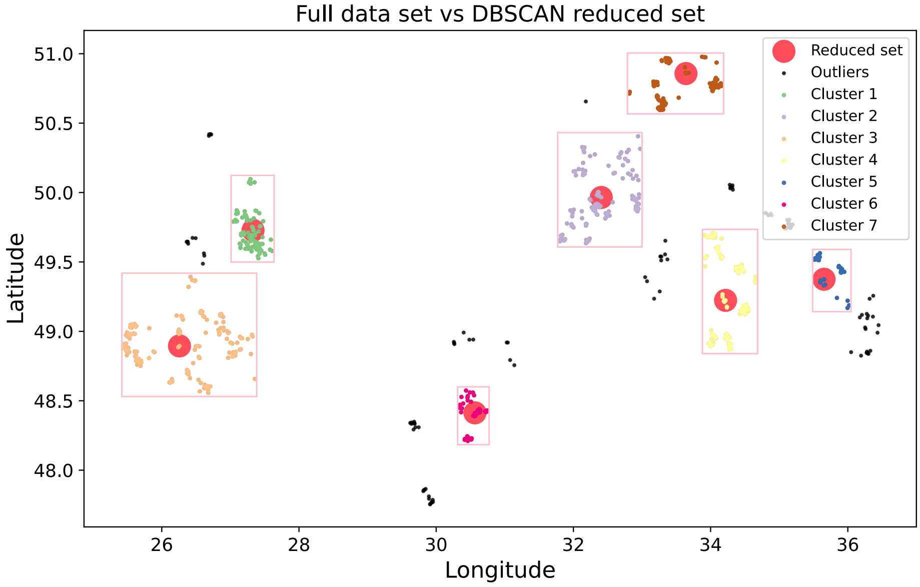


FIGURE 3.7: DBSCAN clustering result.

As the next step, we proceeded with downloading the Sentinel-2 archives having GeoJSON (https://geojson.org/) polygon as the bounding box of each cluster. We downloaded Sentinel-2 Ukrainian territory data for each cluster from September 2017 to November 2018. One should note that in PASTIS (3.2.2) dataset data is 2018-2019. However, as our semantic and instance annotations were only for 2018, we shifted the date range for one year while maintaining the same month range. This date selection is important in the temporal encoder of the model architecture (3.3). In total, we downloaded 800 GB of Sentinel-2 data using SentinelSat Python API (Wille et al., 2017). As bands in the data granule have different spatial resolutions, we applied bilinear interpolation (Kirkland, 2010) using Scipy (Virtanen et al., 2020) to upscale every band to 10 m/pixel resolution. Then we made a grid of non-overlapping squares of 1.28 × 1.28 km area using GeoPy by Esmukov et al., 2021 and made 128 × 128 px patches from each 100 × 100 km granule if in each specific patch there were parcels covering more than 5 ha of area. Using Rasterio, a library for geospatial raster data (Gillies et al., 2013), we reprojected and plotted annotations from GeoJSON format onto 2-dimensional arrays to form labelled instance and semantic masks. An important aspect that one should consider when working with geo-referenced data is coordinate systems projection. Since the Earth is not flat but an irregularly shaped ellipsoid (Johns, 1959), there are multiple cylindrical map projections (Snyder and Steward, 1989; Miller, 1942), with Mercator projection being the most common. It has usual for us properties, *e.g.* north direction as upward, south as downward, west leftward and east rightward. Albeit, it has a negative side-effect of showing the sizes of objects away from the equator bigger than they actually are (Fig. 3.8). There is EPSG registry created to list all known coordinate reference system (CRS). GPS uses EPSG:4326, the coordinate being a tuple of (longitude, latitude) represented in degrees, with axes of Greenwich (0° meridian) and the Equator (0° parallel) (Fig. 3.9). While tiles in Sentinel-2 may have different EPSG, which in our case were EPSG:32635 and EPSG:32636. Due to this projection difference we had to apply coordinate reprojection in our data processing pipeline from EPSG:4326 coordinates into coordinate system of the specific tile.

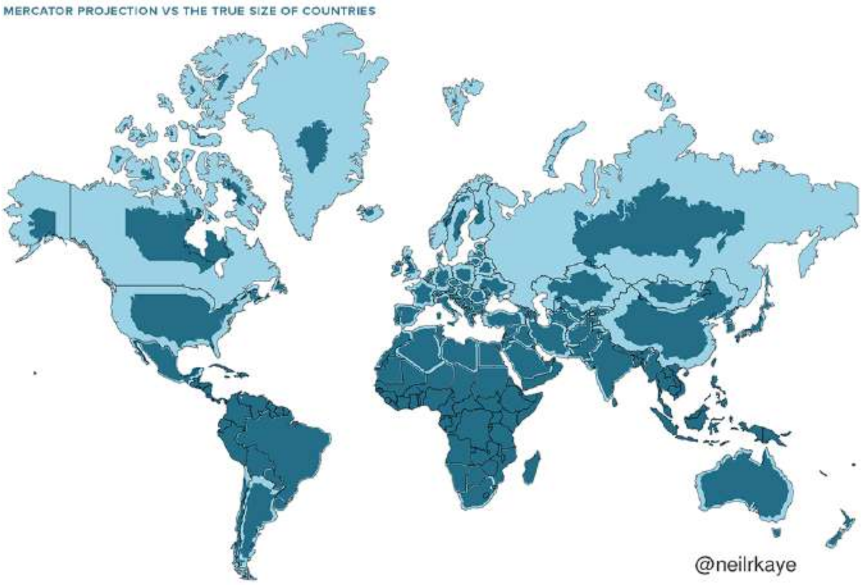


FIGURE 3.8: Mercator projection comparison to actual sizes of the countries. (Illustration author: Neil Kaye)

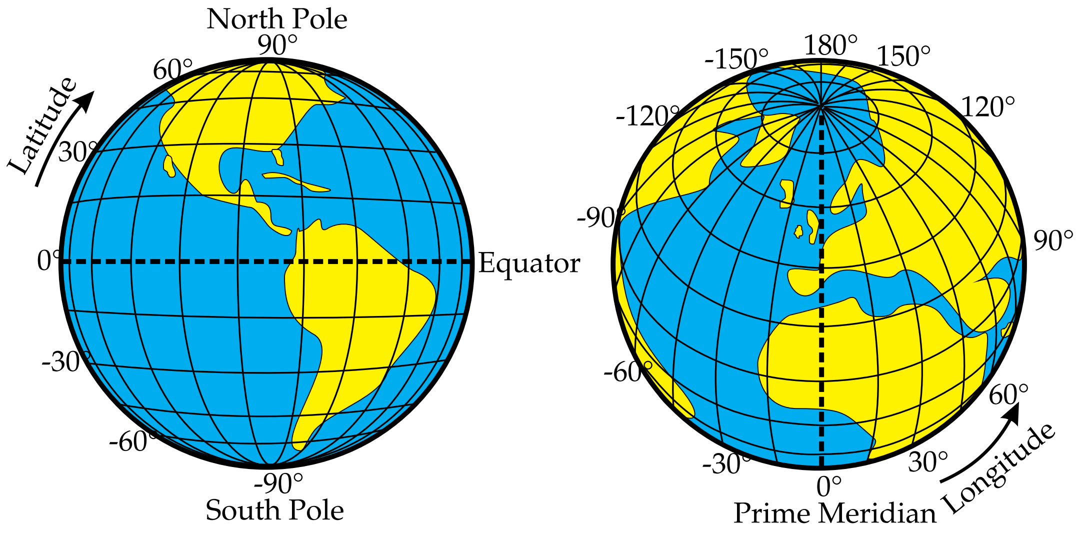
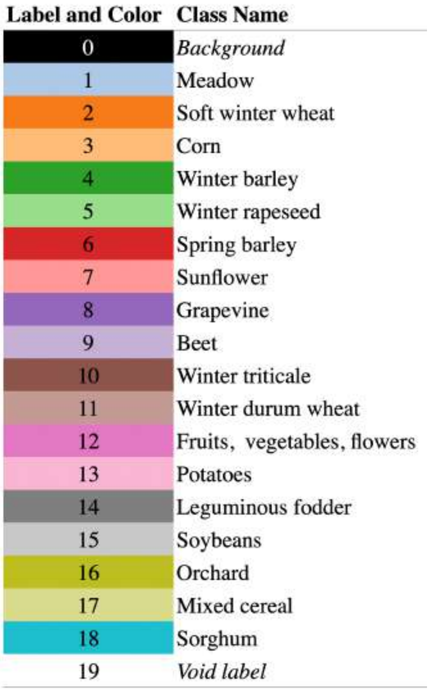


FIGURE 3.9: Latitude and Longitude of the Earth.
(Illustration Source: Wikimedia Commons)

After the "patchify" process, we formed 4-dimensional arrays of Sentinel-2 images in 10 bands with timestamps ranging from 33 to 61 scans. Each patch mapped the respective instance and semantic masks annotation aligned with satellite image EPSG projection.

N.B, However, it is important to remind the reader that the data quality of the provided Ukrainian data is not validated, *i.e.* we cannot state with 100% certainty that the provided crop type or region is entirely accurate for each parcel. Nevertheless, we proceed with such data because it is the only data source we have managed to find.



Label and Color	Class Name
0	*Background*
1	Meadow
2	Soft winter wheat
3	Corn
4	Winter barley
5	Winter rapeseed
6	Spring barley
7	Sunflower
8	Grapevine
9	Beet
10	Winter triticale
11	Winter durum wheat
12	Fruits, vegetables, flowers
13	Potatoes
14	Leguminous fodder
15	Soybeans
16	Orchard
17	Mixed cereal
18	Sorghum
19	*Void label*

FIGURE 3.10: Crop types color-mapping nomenclature from PASTIS.
Source: Fig. 2 in Vivien and Loic, 2021

For dataset generation, we use the PASTIS semantic nomenclature of crop types shown in Fig. 3.10.

3.2.5 Data processing stage hardware requirements

Since the satellite images are very high in spatial resolution (10000 × 10000 px) and consist of multiple channels (13 bands), they consume a minimum of 800 MB while in ZIP-archive. For our work, we need to have also multi-temporal SITS. Therefore we downloaded Sentinel-2 granules for ≈ 800 GB. During processing and complex Data Engineering, we downloaded to one server satellite archives and, in parallel, sent one-by-one cluster data to the processing server via a 1 Gbps Ethernet network. When more space was available after processing one cluster in the queue, we fed the following cluster to processing. For the ideal scenario, one needs at least 2-3 TB of high-speed data storage, SSD prefered, and office-grade networking with minimum Cat-5e Ethernet cables. 1 Gbps internet from ISP is preferred to work with Satellite imagery in the most efficient manner. For scaling, storage requirements are expected to scale vertically as well.

3.3 Model architecture

As a baseline for our experiments, we are using the first end-to-end Deep learning method for panoptic segmentation of SITS, named Parcel-as-Points (PaPs) developed by Garnot and Landrieu, 2021 which we introduced in Sec. 2.3. In this section, we will describe its components.

Each input patch (multi-spectral SITS) to the neural network is 4-dimensional tensor with the shape of $T \times C \times H \times W$, where T - temporal dimension (sequence from 33 to 61 timestamps), C - spectral component (10 channels, Eq. 3.1), H and W (are height and width of the patch respectively, 128 px). Following LeCun et al., 2012, before feeding the Neural network model, each channel of the input data is standardized (Eq. 3.2) to have a mean $\mu = 0$ and standard deviation $\sigma = 1$ (unit variance).

$$\bigcirc_{c=1}^{10} X = X_1 \circ \cdots \circ X_{10} \tag{3.1}$$

$$\forall X_c = \frac{X_c - \mu_c}{\sigma_c} \tag{3.2}$$

3.3.1 Feature extractor

The first module of the pipeline is the feature extractor or Spatio-temporal encoder (Fig. 3.11), Garnot and Landrieu, named this module U-TAE (U-Net with Temporal Attention Encoder). It reminds well-known in the deep-learning research community convolutional neural network (CNN), named U-Net (Ronneberger, Fischer, and Brox, 2015), for its "U"-shaped design. The feature encoder we used also has an encoder as a contractive path (downsampling, left part) and a decoder as an expansive path (upsampling, right part). There are three parts to the module: 1. Spatial encoder, 2. Temporal encoder, and 3. Spatial decoder.

Each Convolution block in the feature extractor has the following design: Conv 3 × 3 → Norm→ ReLU1 → Conv 3 × 3 → Norm→ ReLU2 → Skip connection from ReLU1 output. Where Normalization in the encoding path is Group Normalization (Wu and He, 2018) with 4 groups and Batch Normalization (Ioffe and Szegedy, 2015) in the decoder. Since each sequence consists of images from different timestamps, the samples are not identically distributed in batches. Therefore Group Normalization is used here instead of Batch normalization.

Spatial Encoder

In the feature extractor, each image of the 4-dimensional tensor is encoded by a convolutional encoder. Where from an input tensor of size $T \times C \times W \times H$ multi-level features are extracted with sizes:

1. $T \times 64 \times W \times H$
2. $T \times 64 \times \frac{W}{2} \times \frac{H}{2}$
3. $T \times 64 \times \frac{W}{4} \times \frac{H}{4}$
4. $T \times 128 \times \frac{W}{8} \times \frac{H}{8}$

N.B, As in our case each tensor also have batch dimension, *i.e.*, full tensor is $B \times T \times C \times W \times H$. And T - temporal component may have a varied length from 33 to 61, 2D-Convolutions are not dynamic to this variance. Therefore to make it static, additional preprocessing is implemented before each convolution: temporal and batch components are flattened as $B \times T = B^*$.

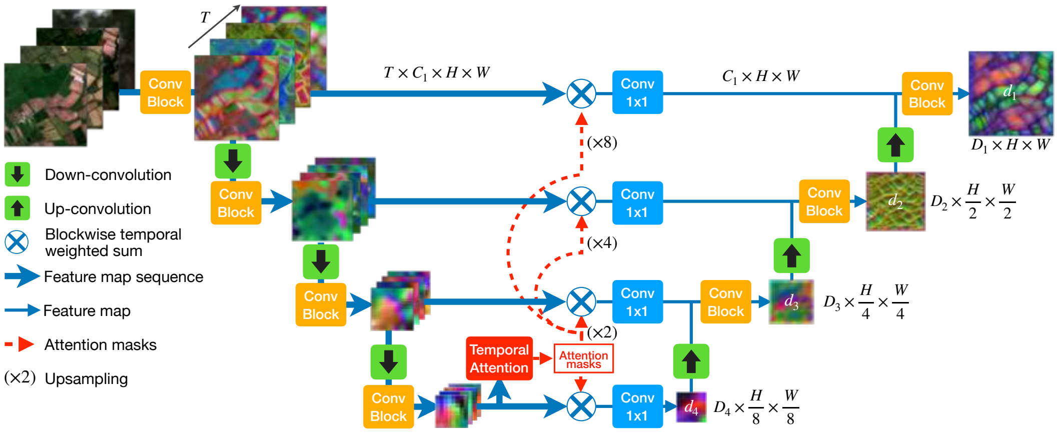


FIGURE 3.11: Feature extractor as spatio-temporal encoding.
(Source: Fig.2 in Garnot and Landrieu, 2021)

Temporal Encoder

The next step is to collapse the temporal dimension into a single representation. To achieve this, there is an attention-based method which processes temporal dimension only at the lowest feature map resolution level. Garnot and Landrieu claim that processing the higher resolution would result in a small spatial receptive field and increased memory requirements. Lightweight-Temporal Attention Encoder (L-TAE) (Garnot and Landrieu, 2020) is used in this module due to its accuracy and efficiency. L-TAE was inspired by multi-head self-attention methods (Vaswani et al., 2017). The temporal encoder with G heads, *i.e.* $G = 16$ in our case, is applied at the lowest resolution feature map extracted in 3.3.1, spatial encoder paragraph. It outputs G masks, with values $[0,1]$ and $T \times H_L \times W_L$ shape, where L denotes the lowest feature map resolution level, *i.e.* $W_L \times H_L = \frac{W}{8} \times \frac{H}{8}$ with $W = 128, H = 128$ in our case.

Then these G tensors are resized with bilinear interpolation to match the spatial resolution of all higher-level feature maps produced by the spatial encoder. Next, interpolated masks are multiplied with feature maps at each spatial level and are fed into Conv $1 \times 1 \rightarrow$ Norm→ ReLU and passed further to the Spatial decoder as skip connections from each spatial resolution level.

Spatial Decoder

In the spatial decoder part, all feature maps from the temporal encoding step are combined by concatenation with feature maps upsampled from the previous spatial resolution level. Each concatenated feature map enriched with temporal and spatial encoded information is then fed into a convolution block with the same design as described at the beginning of Sec. 3.3.1.

Finally Spatio-temporal feature extractor produces the following feature maps shapes:

1. $128 \times \frac{W}{8} \times \frac{H}{8}$
2. $64 \times \frac{W}{4} \times \frac{H}{4}$
3. $32 \times \frac{W}{2} \times \frac{H}{2}$
4. $32 \times W \times H$

These four feature maps are then passed forward to the panoptic segmentation module.

3.3.2 Panoptic Segmentation module

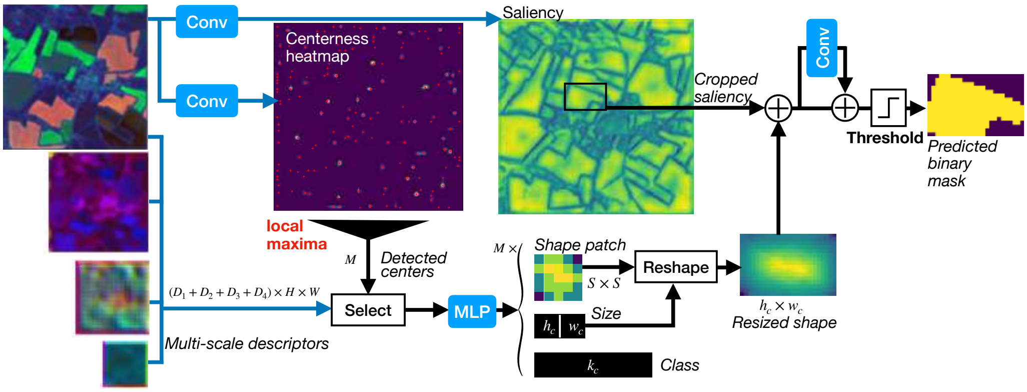


FIGURE 3.12: Panoptic segmentation module diagram.
(Source: Fig.4 in Garnot and Landrieu, 2021)

Parcel-as-Points (PaPs) module (Fig. 3.12) inspired by CenterNet (Zhou, Wang, and Krähenbühl, 2019) and CenterMask (Wang et al., 2020) methods is used here for producing pantoptic segmentation masks (as instance masks with respective class label) from multi-scale feature maps returned by the feature extractor.

Each parcel in the patch is time-invariant, *i.e.* does not change its position with time. Therefore in this approach, each particular parcel is associated with: 1. centerpoint coordinates; 2. bounding box; 3. binary instance mask in the bounding box region and semantic class information identifier ($k \in [1, 20]$, 20 crop types).

Centerness heatmap

First, a centerness heatmap is predicted, supervised by ground truth parcels' bounding boxes, which are used to find centres of all parcels in the patch. The heatmap consists of Gaussian kernels with standard deviations, taken as $\frac{1}{20}$ of the height and width of the respective parcel bounding box associated with each parcel following Eq. 6 in Garnot and Landrieu, 2021. Heatmap is produced by a convolutional layer fed with a feature map from the feature extractor at the highest resolution level. Then parcel centres are computed as local maxima of the predicted heatmap in the neighbourhood of 8 adjacent neighbours. After this operation, we have centre coordinates for parcel candidates.

The predicted heatmap is supervised with the following loss:

$$L_{\text{center}} = -\frac{1}{|P|} \sum_{\substack{i=1...H \\ j=1...W}} \begin{cases} \log(m_{i,j}) \text{ if } \hat{m}_{i,j} = 1 \\ (1-\hat{m}_{i,j})^{\beta} \log(1-m_{i,j}) \text{ else,} \end{cases} \tag{3.3}$$

where $\beta = 4$, $m_{i,j}$ - ground truth heatmap, $\hat{m}_{i,j}$ - predicted centerness heatmap and H, W - height and width of the heatmap, P - number of parcels.

Size and class estimation

Each estimated parcel centre is associated with a multi-scale feature vector constructed by concatenating pixels at the centre coordinate from all channels at each feature map scale level. This vector has a shape $(128 + 64 + 32 + 32) \times M$, where M - is the number of detected centers. We feed the feature vector into three multi-layer perceptrons (MLPs) to get three vectors: 1. size of the parcel (vector size = 2); 2. semantic class of the crop (vector size = K, $K = 20$ for us); 3. shape patch (size $S \times S$, $S = 16$ in our experiments).

Class estimation MLP is supervised with cross-entropy loss for a particular parcel p:

$$L^{p}_{\text{class}} = -k_p \log(\hat{k}_p) \tag{3.4}$$

While size estimation is supervized with a normalized L1 loss:

$$L^{p}_{\text{size}} = \frac{|h_p - \hat{h}_p|}{\hat{h}_p} + \frac{|w_p - \hat{w}_p|}{\hat{w}_p}\ , \tag{3.5}$$

where $(\hat{w}, \hat{h})$ - predicted size scalars for width and height, (w, h) - ground truth bounding box size scalars for parcel p.

Shape estimation

In this step, we combine a rough shape estimation patch with a full-resolution global saliency map to receive an instance segmentation mask.

In the Fig. 3.12, estimated parcel shape patch ($S \times S$, $S = 16$), is rescaled with predicted size for height and width. The resized shape is then added to cropped saliency feature map. Then follows residual convolution layer, which is then added as skip connection, final prediction is achieved with a sigmoid activation function. For inference, the sigmoid output is thresholded with a value of 0.4 to achieve a binary instance mask.

The supevision is provided with the following binary cross-entropy loss (BCE):

$$L^p_{\text{shape}} = \text{BCE}(\hat{l}_p, \text{crop}_p(s_p)) \,, \tag{3.6}$$

where $\hat{l}_p$ is the estimated shape, $\text{crop}_p(s_p))$ ground truth binary instance mask cropped at the predicted bounding box over the parcel center.

3.3.3 Objective function

Objective function is defined by combination of Eq. 3.3, Eq. 3.4, Eq. 3.5 and Eq. 3.6 as follows:

$$L = \lambda_{\text{center}} \times L_{\text{center}} + \frac{1}{|P'|} \sum_{p \in P'} \left(\lambda_{\text{class}} \times L^p_{\text{class}} + \lambda_{\text{size}} \times L^p_{\text{size}} + \lambda_{\text{shape}} \times L^p_{\text{shape}} \right) , \tag{3.7}$$

where P' is the subset of detected parcels, *i.e.* for which center coordinate is estimated; $\lambda_{\text{center}} = 1$, $\lambda_{\text{class}} = 1$, $\lambda_{\text{size}} = 1$, $\lambda_{\text{shape}} = 1$.

3.3.4 Metrics

Regarding quantitative measurements, we are using three metrics PQ (panoptic quality, Eq. 3.8), SQ (Segmentation Quality, Eq. 3.9), and RQ (Recognition quality, Eq. 3.10), first introduced by Kirillov et al., 2018.

Panoptic Quality is calculated for each class and averaged over all classes. This metric is as a combination of SQ and RQ.

$$\text{PQ} = \frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{3.8}$$

Segmentation quality (SQ) may be considered as an average IoU (intersection over union) of matched true positive segments.

$$SQ = \frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP|} \tag{3.9}$$

While the recognition quality (RQ) is similar to the F1-score (harmonic mean of precision and recall) classification metric.

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{3.10}$$

3.3.5 Modelling stage compute requiremnts

Model training requires GPU acceleration. In our experiments, we used one NVIDIA GeForce RTX 3090, 12 GB of GPU memory (batch size = 4 training). The single experiment takes 11 hours with the PASTIS dataset of 10 bands per SITS for 100 epochs. We note, however, that it is possible to optimize the training pipeline with distributed training using mechanisms of Pytorch Machine Learning library by Paszke et al., 2019.

For model training, we used a learning rate (LR) of 0.01 at the start and a multistep LR scheduler reducing LR twice at 60 and 80 epochs by a factor of 0.3. For gradient descent we used Adam optimization (Kingma and Ba, 2014). For monitoring training, experiment tracking, and metrics logging, we extensively use Weights & Biases by Biewald, 2020.

Chapter 4

Experiments and Results

This chapter describes the experiments we conducted and the results we obtained. Experiments are divided into two parts: 1. French Data and model architecture modifications; 2. Ukrainian Data and its specificities.

We planned experiments with two goals in mind: 1. Improve model architecture to outperform the PASTIS benchmark, and 2. Showcase panoptic segmentation on Ukrainian territory.

4.1 Experiments with French Data

4.1.1 Multi-temporal Satellite imagery vs Single timestamp

Multi-temporal satellite imagery has very recently been introduced to the problem of agricultural land segmentation, and the relevant problem of instance segmentation previously was done using only a single satellite image (Rieke, 2019). Therefore in this experiment, we wanted to verify how important it is to have multiple timestamps for a satellite patch to have a good panoptic segmentation quality.

We compared a single timestamp with multi-temporal satellite image time series. For multi-temporal, we used a sequence from 33 to 61 satellite images per patch taken from September 2018 until November 2019. While for a single timeframe, we used the 1st of June 2019.

In Table 4.1 one may see quantitative results: the first row shows numerical results from the original paper Garnot and Landrieu, 2021, the second row our reproduced metrics, and the third - only a single satellite scan per patch.

The experiment resulted in an almost 2x lower panoptic quality (PQ) measurement when using a single timestamp compared to a multi-temporal sequence, while the Segmentation quality difference is not so drastic. It may signify that in order to estimate correct semantic information, *i.e.* crop type, the model needs to see a time series of satellite scans. The reasoning here might be that different classes of crops grow differently, and in order to understand the difference, there is a need to have this crop grow change encoded into the features. Our result complies with some other findings. In Kussul et al., 2017 authors describe non-static crop nature and state that a single timestamp does not capture differences in plant phenological profiles and human interventions during harvests. When the crop has started growing, it may look similar, but the difference is more visible once it grows enough. Therefore, this task needs to work with satellite time series.

Qualitative results also show worse performance with single-date patches. One may see that fewer parcels are delineated, and some of the highlighted parcels have the wrong crop class predicted. In comparison, the multi-temporal model produces more correctly segmented parcels. Therefore our focus will be on multi-temporal satellite imagery time-series (SITS) for the rest of our experiments.

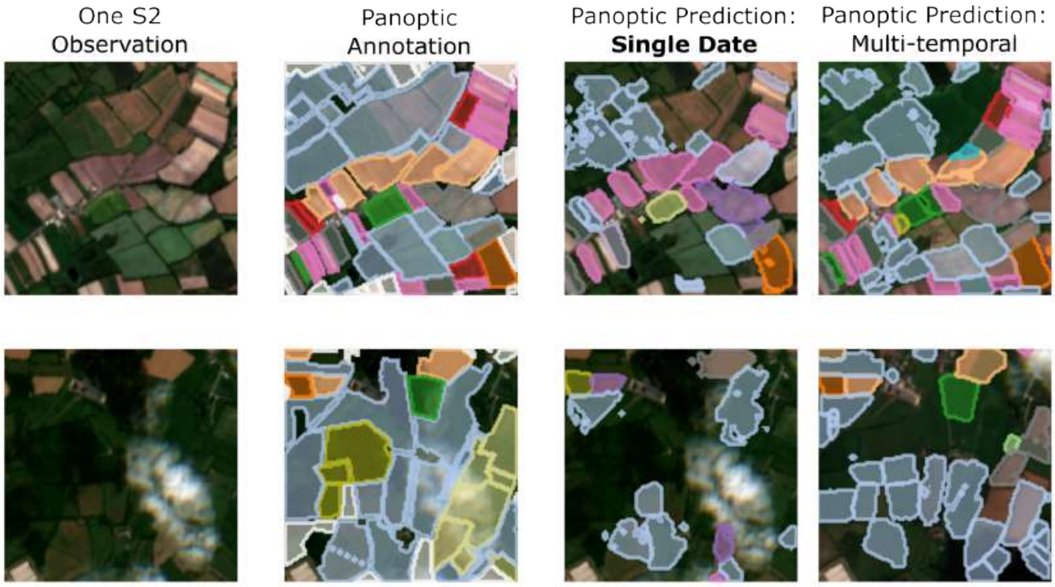


FIGURE 4.1: Qualitative results comparing single timestemp with multi-temporal SITS.

Additionally, we reproduced the experiment from the original paper of Garnot and Landrieu, 2021 to verify reproducibility. To our surprise, in RQ and PQ, our reproduced results showed better results but worse in SQ. We will be using reproduced variant metrics for other comparisons because these numerical results might be hardware or software dependent.

Experiment	SQ	RQ	PQ
Multi-temporal, 10 bands (from paper)	**81.44**%	47.90 %	39.37 %
Multi-temporal, 10 bands (reproduced)	81.30 %	**48.09**%	**39.43**%
Single timestamp, 10 bands	77.08 %	25.52 %	20.06 %

TABLE 4.1: Mono date vs multi-temporal

4.1.2 Shape prediction MLP vs UNET-decoder

While analysing panoptic segmentation module architecture, we found that parcel shape patch is estimated with very little feature information. In panoptic segmentation module (Fig. 3.12), shape patch estimation is conducted with simple MLP. The MLP is fed with a feature vector consisting only of the pixel values of the parcel centre coordinate from all feature map channels. Whereas our feature encoder produced spatial feature maps. Therefore our hypothesis was to utilise better this spatial property of the feature maps *i.e.* use neighbourhood around parcel centre coordinate, rather than just using only a single pixel from each channel of the feature maps.

We developed a shape estimation CNN submodule to conduct the experiment mentioned above. Inspired by the effectiveness of U-NET (Ronneberger, Fischer, and Brox, 2015) for various segmentation tasks, we created a modified decoder which was fed with feature maps produced by the feature extractor (3.3.1) and which returned one channel ($S \times S$, $S = 16$ as in original architecture) shape patch.

Experiment	SQ	RQ	PQ
Original shape MLP	**81.30%**	**48.09%**	**39.43%**
U-Net decoder for shape estimator	81.05 %	47.69 %	39.07 %

Quantitative results showed that simple MLP works better for the PASTIS dataset than larger spatial feature neighbourhood fed to the U-Net CNN decoder module. This is surprising because rough parcel shape is estimated only with a single centre pixel taken from each feature channel. However final pixel-precise instance mask is produced with an additional saliency map and convolution at the final stage. Even though in this experiment numerical results are better, we believe hyperparameter tuning, *e.g.* shape patch size or neighbourhood region might produce better performance. In this experiment, we wanted to make the first test, and in the future, we might conduct experiments with all hyperparameters tested with higher computational resources.

4.1.3 Micro-design architectural change of activation function

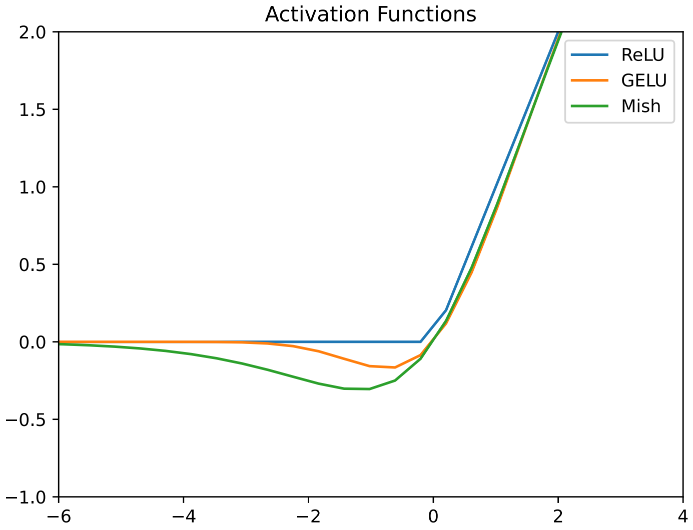


FIGURE 4.2: Vizualization of activation functions: ReLU, GELU and Mish.

Following trends in deep-learning neural-network "surgery" or network architecture design choices to increase model performance (Liu et al., 2022; Ramachandran, Zoph, and Le, 2017) we wanted to experiment with the choice of the activation function. The activation function provides non-linearity, a crucial property of a DNN. The selection of the activation function can lower or boost model performance. Therefore to test if we can outperform our benchmark with proper activation function, we compared three functions: the most common ReLU, GELU and newer one Mish. The selection of functions we made based on our experience in the domain and literature suggestions in the research community as in Zhang et al., 2021; Bochkovskiy, Wang, and Liao, 2020. One may see them plotted in Fig. 4.2.

ReLU activation function, introduced by Deng et al., 2009 is denoted as: $f(x) = max(0, x)$. This activation function is the first-to-try choice in almost every deep-learning problem. This function was introduced as a function which has a lower-to-no vanishing gradient problem. However, due to being so popular, it may not always be the best, as there are new functions which bring higher model performance.

One of the better alternatives in the research community is Gaussian Error Linear Unit (GELU) function (Hendrycks and Gimpel, 2020). It allows some negative values to be passed to other neurons rather than being zeroed as in ReLU.

Third function we wanted to experiment is Mish (Misra, 2020), defined as follows $f(x) = x \cdot \tanh(\text{softplus}(x))$, where softplus(x) $= \log(1 + e^x)$. Even though this function has similar properties with GELU, *e.g.* unbounded positive domain, and bounded negative domain. As authors of the function claim and show in their results, empirically, it provides better performance on ImageNet-1k (Deng et al., 2009), CIFAR-10 than ReLU and GELU.

The experiments show in Table 4.2, that for this model architecture and the dataset RELU outperforms GELU and Mish design choices. However, in the train-time panoptic quality measure, GELU outperformed ReLU and Mish, but with lower Segmentation quality. This means that the hard negative bound, as in ReLU, is essential for accurate parcel contour segmentation.

Experiment	test SQ	test RQ	test PQ	train PQ	train SQ
ReLU	**81.30**%	**48.09**%	**39.43**%	55.54 %	**82.86**%
GELU	80.88 %	42.33 %	34.51 %	**63.32**%	81.66 %
Mish	80.58 %	42.55 %	34.71 %	61.40 %	81.14 %

TABLE 4.2: Quantitative results with activation functions.

4.1.4 Input data channels modality

Experiment	SQ	RQ	PQ
10 bands	81.30 %	**48.09**%	**39.43**%
10 bands + elevation	80.35 %	47.39 %	38.53 %
5 bands	**81.68**%	44.97 %	37.01 %

TABLE 4.3: Quantitative results on data modalities experiments.

By using band combinations, we can extract specific information from an image. For example, there are band combinations that highlight geologic, agricultural, or vegetation features in an image. Abraham and Wloka, 2021 remind us to make a more careful selection of a band set to utilise spectres of satellite data better. In this experiment, we added elevation to the 10 bands to help model segment parcels with not-flat relief. Even though France is not a completely flat territory, we see no improvements in metrics for additional elevation channel (Table 4.3). It seems ten spectral bands are enough for panoptic segmentation.

We also tested how the model works with the first five bands, which have 10 and 20 m spatial resolution and are most sensitive to vegetation. Interestingly, having only five spectral channels, the model still shows excellent results, even outperforming segmentation quality: 81.68 % with 81.30 % if using all ten bands. It might signify that other channels are responsible for only 2.42 % panoptic quality improvement. However, to be sure of this clause, we need to make such an experiment using the other five bands because it might also be possible that information in the first five and last five bands is interchangeable.

4.2 Application of the algorithm to Ukrainian data

The second part of our experiments is the application of the method to the Ukrainian data we processed.

4.2.1 10 m/pixel

Experiment	SQ	RQ	PQ
Model trained on France, PASTIS normalization	16.43 %	**1.92**%	1.39 %
Model trained on France, Ukraine data normalization	16.73 %	1.16 %	0.84 %
Model trained on Ukraine, Ukraine data normalization	**40.51**%	1.83 %	**1.52**%

TABLE 4.4: Evaluation on Ukrainian data 10m/pixel test set.

First, we generated annotations and downloaded satellite images the same way as the French dataset (PASTIS), 10 m / pixel spatial resolution. Therefore each patch has 128 × 128 px and covers 1.28 × 1.28 km.

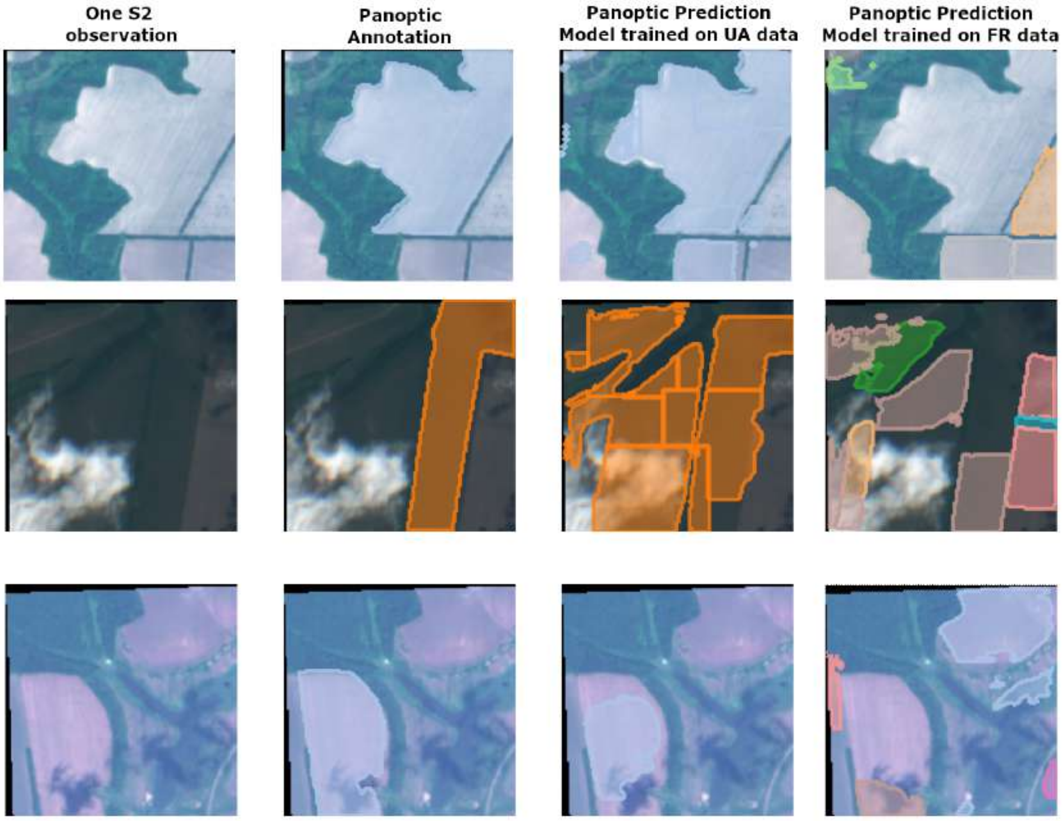

One S2 observation
Panoptic Annotation
Panoptic Prediction Model trained on UA data
Panoptic Prediction Model trained on FR data

FIGURE 4.3: Qualitative results with model train on FR (French data) and model trained on UA (Ukraine) dataset at 10m/pixel.

First, we tested how the model trained on French territory works on Ukrainian territory. Using the same data normalisation as on French patches, we receive a Segmentation quality of 16.43 %, Recognition quality of 1.92 % and panoptic quality of 1.39 %, compared to PQ of 39.43 % on PASTIS (Table 4.4). This result is far from ideal. When we computed normalisation on Ukrainian data, we saw a minute improvement in segmentation quality. But degradation in panoptic quality. Then we trained the model with weights pre-trained on PASTIS on Ukrainian data. Results improved significantly for segmentation quality, *i.e.* finding parcel contours. Recognition quality, *i.e.* parcel crop type segmentation, however, was not improved as expected.

If comparing visually, one may see in Fig. 4.3 that performance is indeed not so good. The model trained on the Ukrainian Dataset makes better guesses of the field crop type and tries to delineate parcels properly. In contrast, the model trained on French data produces more instances of more varied crop types, even ones not labelled in the ground truth.

4.2.2 Parcel area difference in PASTIS and Ukrainian Data

We hoped that the model would work correctly on our carefully generated Dataset from the data we received. As it turned out, there is a drastic difference between France and Ukraine in terms of parcel areas. When looking at the patches we generated for Ukraine and patches from PASTIS, in Fig. 4.5, it becomes evident that while in France patch may have 100+ fields inside, in Ukraine patch may have only one parcel.

To validate our hypothesis, we computed the average parcel area in m^2, plotted in Fig. 4.4. From this, we see that parcels in PASTIS have a median area $10.6 \cdot 10^3$ m^2 per parcel, while in our Ukrainian dataset median parcel area is $214.15 \cdot 10^3$ m^2, 20 times larger median parcel area. We believe that this specific factor may be due to the Dataset we have and the type of business the person who gave us this Dataset conducts. However, we think this may also be due to differences between Ukraine and France in historical factors. In France, there are many small businesses around the country. In Ukraine, very few people or enterprises may hold many large-area parcels. We leave the hypothesis for historians to verify why such a difference arose in the past. On the contrary, to verify our clause about the present state, whether all parcels in Ukraine are larger than in France, we need a bigger dataset annotated for Ukrainian agricultural territory.

We decided to change the scale and experimented with 30m/pixel instead of 10m/pixel spatial resolution. The median parcel area is $23.8 \cdot 10^3$ m^2 in such case, which is only 2 times larger than in PASTIS.

4.2.3 30 m/pixel

Experiment	SQ	RQ	PQ
Model trained on France	15.99 %	5.13 %	4.39 %
Model trained with Ukrainian data	**34.17%**	**7.11%**	**6.08%**

TABLE 4.5: Evaluation on Ukrainian data 30m/pixel test set.

Here we generated a dataset with 30m/pixel spatial resolution for Ukrainian data to verify the hypothesis whether this scale will help achieve higher performance due to the large parcel are in the Ukrainian Dataset compared to French data.

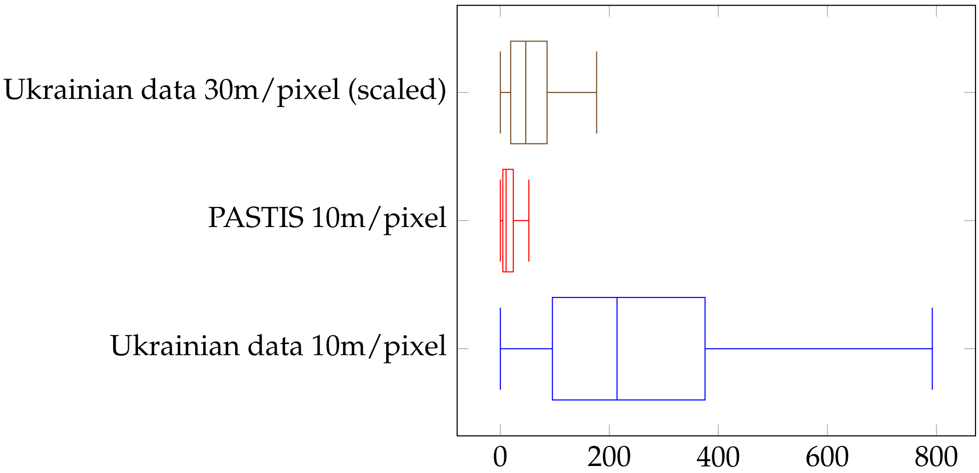
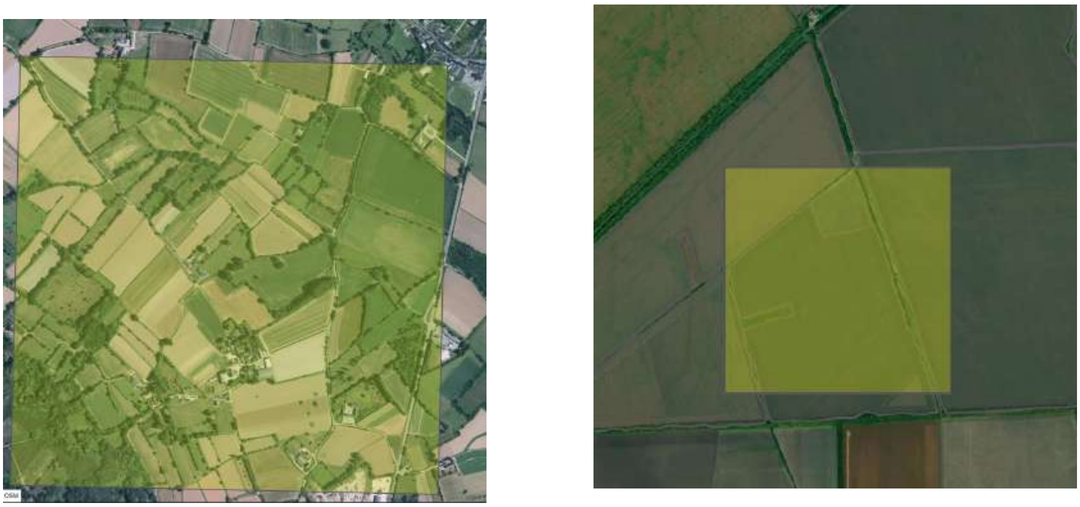


FIGURE 4.4: Parcel area in thousands (10^3) of m^2 for PASTIS dataset and for Ukrainian data. 30m/pixel (scaled) means that we used same 128 × 128 px spatial dimension as in PASTIS, due to difference with 10m/pixel Ukraine-France parcels.



(A) French land patch example. Multiple parcels in one patch.

(B) Ukrainian land patch example. Due to big parcels, one big parcel covers most of the patch area.

FIGURE 4.5: Difference between French (PASTIS) and Ukrainian Dataset. Yellow rectangles have the same physical area of 1.28 × 1.28 km on both images.

In Fig. 4.6 we show qualitative results with a model trained on French territory and a model trained on concatenated French + Ukrainian Dataset (due to the low number of training patches at 30m/pixel). The model works slightly better based on metrics in Table 4.5. Segmentation quality is 34.17%, 2 times higher than the model trained on France territory, recognition quality is 7.11%, and panoptic quality is 6.08% still not high as we would expect.

As we worked with both the French and Ukrainian Dataset, we understood how important the annotations' quality and quantity are. Regarding PASTIS, a French mapping agency created this Dataset with farmers' help, where all parcels in each patch are annotated with instance and semantic labels. While in Ukrainian Dataset that we have, on average, there are 3 parcels annotation present in each patch. Additionally, the data we have for Ukraine are limited in the number of parcels and, hence, patches. While in PASTIS, there are more than 2433 patches, with 1455 provisioned for training, in Ukrainian Dataset, we have only 51 patches, 35 for training at 30m/pixel, which is why we trained concatenated PASTIS + UA@30m Dataset for 30m/pixel experiment with Ukrainian data.

Although we aspired to have higher quantitative and qualitative results, we reached our goal to showcase the panoptic segmentation possibility for Ukraine. We found specificities in working with Ukrainian fields. Hopefully, soon, we will see a new Dataset with high panoptic quality annotations and a large area of Ukrainian land, as this will open many new research and business opportunities for the emerging market economy of Ukraine.

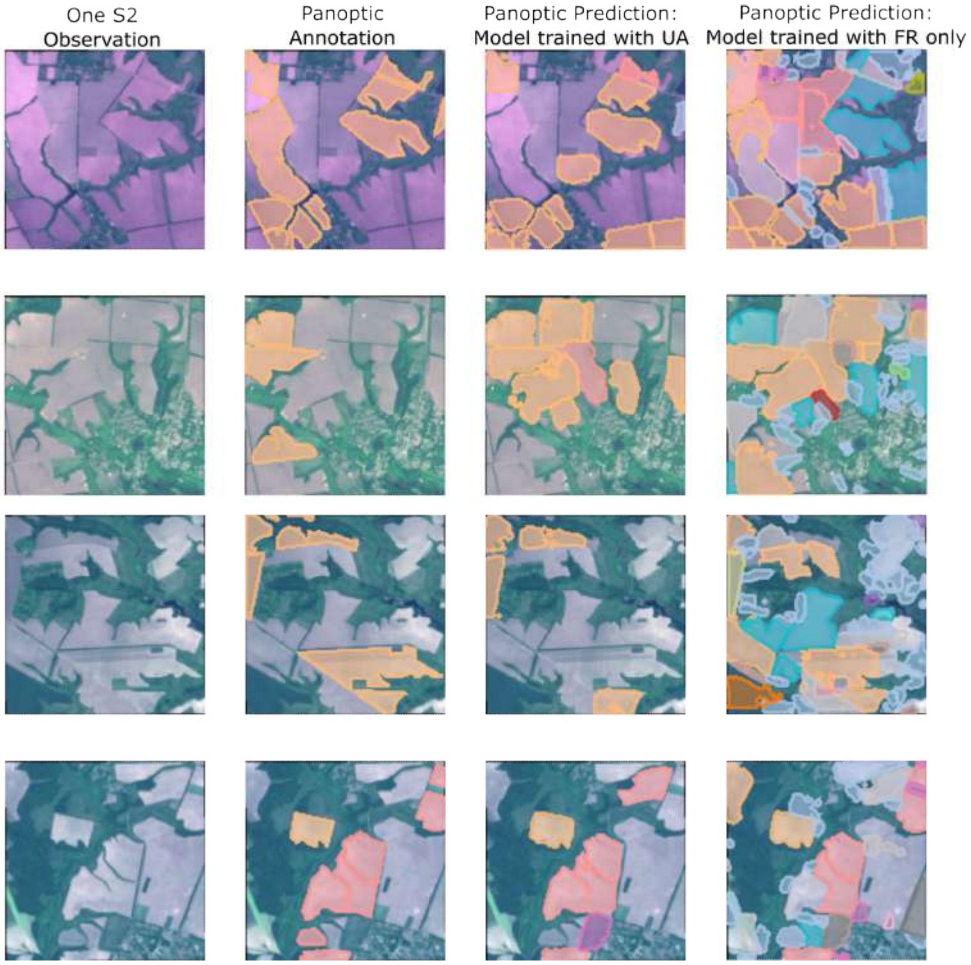


FIGURE 4.6: Qualitative results with model trained on FR (French data) and model trained on contatenated FR+UA (Ukraine) dataset at 30m/pixel.

Chapter 5

Conclusions

5.1 Discussion

Garnot and Landrieu, 2021 showed that panoptic segmentation of crop fields using multi-temporal satellite images time series is feasible. They tested their approach on French territory. We verified the reproducibility of the PASTIS dataset first, created a dataset prototype for Ukrainian territory and adopted the approach further to Ukraine territory. The geographic location between France and Ukraine is different. Still, the model transfer approach is not fully explored in the research works. Our experiments showed that to train a panoptic segmentation model, one needs to have panoptic annotations of high quality covering as much area of the given country as possible.

Furthermore, while the research community should primarily focus on improving the accuracy of panoptic segmentation for agriculture, it is equally important to explore how to make the model more robust to different input data perturbations and noise. While experimenting, we observed that while the current approach shows the feasibility of panoptic segmentation of parcels, it is at the same time sensitive to several factors: shift of the image even by 1 pixel in any direction changes the prediction; taking not enough time stamps into SITS sequence worsen the performance; very tiny and very large parcels; patches with only one parcel present. These primary aspects need to be addressed in the further exploration of this approach, making it robust and production-ready.

5.2 Conclusions

To conclude, in this written composition, we explored how multi-temporal satellite imagery data can be used for parcel panoptic segmentation. We showcased the most recent state-of-the-art method for panoptic segmentation on Ukrainian farmland data. Described how important are open annotated datasets for the performance of this algorithm. Such a project may later be advanced with parcels area calculation, computation of vegetation index and humidity levels of the farms. It may help develop a method to track crops over time and even predict how the particular plant may grow in a specific field. We believe that such a project may further be used to manage agricultural land better and help achieve good agro-ecological conditions overall. In developed countries, such data science instruments using satellite provides analytics and helps make predictions based on historical data of soil quality, make a/b testing on methods for agriculture soil moisture and melioration approaches. Assist farmers in choosing when it is better to grow a crop, predict how the crop will grow and select the best crop for the particular field. Finally, this method may help plan resources and optimize such resources for the best optimal outcome for the global climate change and food supply issues, which will make better pricing for the fields based on vegetation index quality change.

We hope our showcase of the panoptic segmentation technology, an AI-driven algorithm, can assist in rapid land recultivation after war damages to parcels (Fig. 5.1).

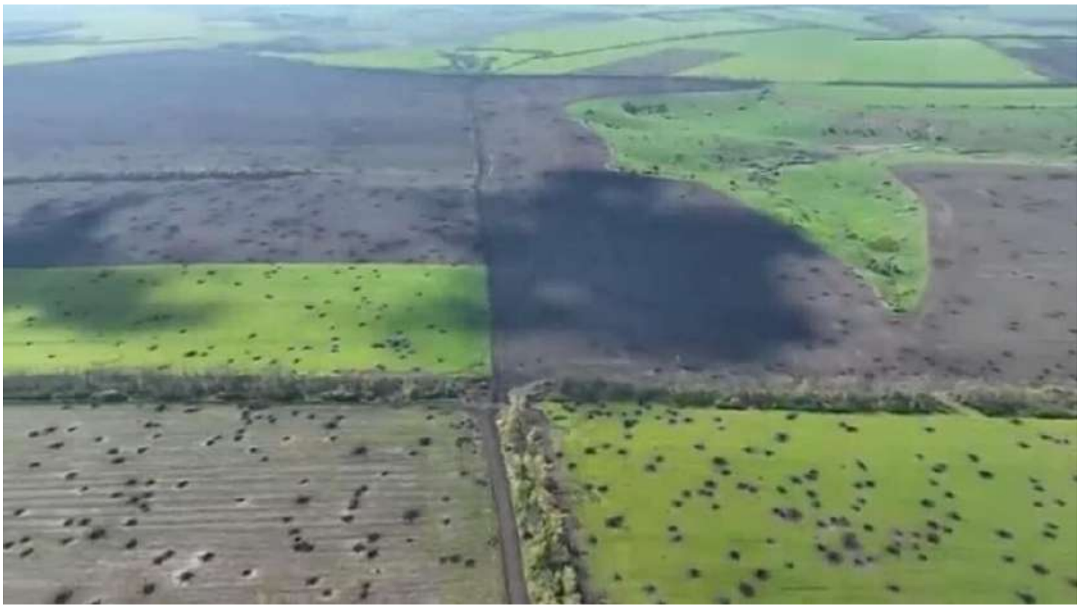


FIGURE 5.1: Drone photo of war-damaged land in Ukraine.
(Author unknown. Image Source: social media.)

Chapter 6

Future Advancements

- Hopefully government or businesses of Ukraine will pro-bono publish annotated dataset of Ukrainian farmland for research purposes as French Mapping Agency did for the French PASTIS dataset. Such labelled data is needed to develop accurate and robust algorithms for the Ukrainian agriculture sector.
- In Ukraine, there is a cadastre `https://bit.ly/cadastre_ukraine` map which also suffers from annotation defects as in Zorzi, Bittner, and Fraundorfer, 2020, which might be improved using satellite imagery segmentation.
- With the rising popularity of deep-learning methods for satellite imagery tasks, we anticipate the appearance of new benchmarks and pre-trained models for more than 3 RGB channels as it is now with ImageNet pre-trained feature extractors. Therefore we recommend experimenting with more advanced feature extractors from other Computer vision problems pre-trained on big datasets, *e.g.* AlphaNet by Wang et al., 2021, to then apply for panoptic segmentation as a downstream task.
- Since there are data of similar nature, *i.e.* multi-temporal and multi-channel, in other fields of science, we believe this panoptic segmentation approach may be applied to other fields as well, such as Biomedical imaging or Astrophysics.
- Experiment with recent data augmentation techniques such as RandAugment (Cubuk et al., 2019), MixUp (Zhang et al., 2017b) or CutMix (Yun et al., 2019) to make the model more robust.

Bibliography

Abraham, Joshua and Calden Wloka (2021). *Edge Detection for Satellite Images without Deep Networks*. arXiv: 2105.12633 [cs.CV].

Ashraf, Muhammad, Mohd Maah, and Ismail Yusoff (2011). "Introduction to Remote Sensing of Biomass". In: ISBN: 978-953-307-490-0. DOI: 10.5772/16462.

Belgiu, Mariana and Lucian Drăguţ (2016). "Random forest in remote sensing: A review of applications and future directions". In: *ISPRS journal of photogrammetry and remote sensing* 114, pp. 24–31.

Biewald, Lukas (2020). *Experiment Tracking with Weights and Biases*. Software available from wandb.com. URL: https://www.wandb.com/.

Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao (2020). "Yolov4: Optimal speed and accuracy of object detection". In: *arXiv preprint arXiv:2004.10934*.

Chen, Jie et al. (2021). "DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14, 1194–1206. ISSN: 2151-1535. DOI: 10.1109/jstars.2020.3037893. URL: http://dx.doi.org/10.1109/JSTARS.2020.3037893.

Cubuk, Ekin D. et al. (2019). *RandAugment: Practical automated data augmentation with a reduced search space*. DOI: 10.48550/ARXIV.1909.13719. URL: https://arxiv.org/abs/1909.13719.

Deng, Jia et al. (2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

Dubovik, Oleg et al. (2021). "Grand Challenges in Satellite Remote Sensing". In: *Frontiers in Remote Sensing* 2, p. 1. ISSN: 2673-6187. DOI: 10.3389/frsen.2021.619818. URL: https://www.frontiersin.org/article/10.3389/frsen.2021.619818.

Esmukov, Tigas et al. (2021). *Geocoding library for Python.* Version 2.2.0. URL: https://github.com/geopy/geopy.

Ester, Martin et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, 226–231.

Fan, Haoqi et al. (2021). *Multiscale Vision Transformers*. arXiv: 2104.11227 [cs.CV].

Frolking, Steve et al. (1999). "Agricultural land-use in China: a comparison of area estimates from ground-based census and satellite-borne remote sensing". In: *Global Ecology and Biogeography* 8.5, pp. 407–416.

Garnot, Vivien Sainte Fare and Loic Landrieu (2020). *Lightweight Temporal Self-Attention for Classifying Satellite Image Time Series*. DOI: 10.48550/ARXIV.2007.00586. URL: https://arxiv.org/abs/2007.00586.

Garnot, Vivien Sainte Fare and Loic Landrieu (2021). "Panoptic Segmentation of Satellite Image Time Series With Convolutional Temporal Attention Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4872–4881.

Garnot, Vivien Sainte Fare, Loic Landrieu, and Nesrine Chehata (2021). *Multi-Modal Temporal Attention Models for Crop Mapping from Satellite Time Series*. arXiv: 2112.07558 [cs.CV].

Gillies, Sean et al. (2013). *Rasterio: geospatial raster I/O for Python programmers*. Mapbox. URL: https://github.com/rasterio/rasterio.

Guérin, Eric et al. (2021). *Satellite Image Semantic Segmentation*. arXiv: 2110.05812 [cs.CV].

Hendrycks, Dan and Kevin Gimpel (2020). *Gaussian Error Linear Units (GELUs)*. arXiv: 1606.08415 [cs.LG].

Huete, A. R. and G. Ponce (2010). "SATELLITE OBSERVED SHIFTS IN SEASONALITY AND VEGETATION -RAINFALL RELATIONSHIPS IN THE SOUTHWEST USA". In.

Ioffe, Sergey and Christian Szegedy (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. DOI: 10.48550/ARXIV.1502.03167. URL: https://arxiv.org/abs/1502.03167.

Johns, R. K. C. (Dec. 1959). "The Figure of the Earth". In: 53, p. 257.

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. DOI: 10.48550/ARXIV.1412.6980. URL: https://arxiv.org/abs/1412.6980.

Kirillov, Alexander et al. (2018). *Panoptic Segmentation*. DOI: 10.48550/ARXIV.1801.00868. URL: https://arxiv.org/abs/1801.00868.

— (2019). *Panoptic Segmentation*. arXiv: 1801.00868 [cs.CV].

Kirkland, Earl J. (2010). "Bilinear Interpolation". In: *Advanced Computing in Electron Microscopy*. Boston, MA: Springer US, pp. 261–263. ISBN: 978-1-4419-6533-2. DOI: 10.1007/978-1-4419-6533-2_12. URL: https://doi.org/10.1007/978-1-4419-6533-2_12.

Kussul, Nataliia et al. (2017). "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data". In: *IEEE Geoscience and Remote Sensing Letters* 14.5, pp. 778–782. DOI: 10.1109/LGRS.2017.2681128.

Lakshmi, Venkat (2013). "Remote sensing of soil moisture". In: *International Scholarly Research Notices* 2013.

LeCun, Yann A. et al. (2012). "Efficient BackProp". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–48. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_3. URL: https://doi.org/10.1007/978-3-642-35289-8_3.

Lentile, Leigh B et al. (2006). "Remote sensing techniques to assess active fire characteristics and post-fire effects". In: *International Journal of Wildland Fire* 15.3, pp. 319–345.

Li, Yanwei et al. (2018). *Attention-guided Unified Network for Panoptic Segmentation*. DOI: 10.48550/ARXIV.1812.03904. URL: https://arxiv.org/abs/1812.03904.

Lin, Tsung-Yi et al. (2016). *Feature Pyramid Networks for Object Detection*. DOI: 10.48550/ARXIV.1612.03144. URL: https://arxiv.org/abs/1612.03144.

Liu, Zhuang et al. (2022). *A ConvNet for the 2020s*. arXiv: 2201.03545 [cs.CV].

McDowell, Nate G. et al. (2015). "Global satellite monitoring of climate-induced vegetation disturbances". In: *Trends in Plant Science* 20.2, pp. 114–123.

Miller, O. M. (1942). "Notes on Cylindrical World Map Projections". In: *Geographical Review* 32.3, pp. 424–430. ISSN: 00167428. URL: http://www.jstor.org/stable/210384 (visited on 06/01/2022).

Minaee, Shervin et al. (2020). *Image Segmentation Using Deep Learning: A Survey*. DOI: 10.48550/ARXIV.2001.05566. URL: https://arxiv.org/abs/2001.05566.

Misra, Diganta (2020). *Mish: A Self Regularized Non-Monotonic Activation Function*. arXiv: 1908.08681 [cs.LG].

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Ramachandran, Prajit, Barret Zoph, and Quoc V. Le (2017). *Searching for Activation Functions*. arXiv: 1710.05941 [cs.NE].

Rieke, Christoph (2019). *Deep Learning for Instance Segmentation of Agricultural Fields*. https://github.com/chrieke/InstanceSegmentation_Sentinel2.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. DOI: 10.48550/ARXIV.1505.04597. URL: https://arxiv.org/abs/1505.04597.

Rußwurm, Marc et al. (2019). *Early Classification for Agricultural Monitoring from Satellite Time Series*. arXiv: 1908.10283 [cs.LG].

Schmitt, Michael et al. (2020). *Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping – Challenges and Opportunities*. arXiv: 2002.08254 [cs.CV].

Schubert, Erich et al. (2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". In: *ACM Trans. Database Syst.* 42.3. ISSN: 0362-5915. DOI: 10.1145/3068335. URL: https://doi.org/10.1145/3068335.

Slater, James A. and Stephen Malys (1998). "WGS 84 — Past, Present and Future". In: *Advances in Positioning and Reference Frames*. Ed. by Fritz K. Brunner. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–7. ISBN: 978-3-662-03714-0.

Snyder, John Parr and Harry Steward (1989). *Bibliography of map projections*. 1856. US Government Printing Office.

Tarasiou, Michail and Stefanos Zafeiriou (2021). *DeepSatData: Building large scale datasets of satellite images for training machine learning models*. arXiv: 2104.13824 [cs.CV].

Tsagkatakis, Grigorios et al. (2019). "Survey of Deep-Learning Approaches for Remote Sensing Observation Enhancement". In: *Sensors* 19.18. ISSN: 1424-8220. DOI: 10.3390/s19183929. URL: https://www.mdpi.com/1424-8220/19/18/3929.

Vaswani, Ashish et al. (2017). *Attention Is All You Need*. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

Vaze, Sagar et al. (2020). *Optimal Use of Multi-spectral Satellite Data with Convolutional Neural Networks*. arXiv: 2009.07000 [cs.CV].

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Vivien, Sainte Fare Garnot and Landrieu Loic (2021). *PASTIS - Panoptic Segmentation of Satellite image TIme Series*. Version 1.0. DOI: 10.5281/zenodo.5012942. URL: https://doi.org/10.5281/zenodo.5012942.

Wang, Dilin et al. (2021). *AlphaNet: Improved Training of Supernets with Alpha-Divergence*. arXiv: 2102.07954 [cs.CV].

Wang, Yuqing et al. (2020). *CenterMask: single shot instance segmentation with point representation*. DOI: 10.48550/ARXIV.2004.04446. URL: https://arxiv.org/abs/2004.04446.

Wille, Marcel et al. (Aug. 2017). *Sentinelsat (v0.12): Utility to search and download Copernicus Sentinel satellite images*. DOI: 10.5281/zenodo.841118.

Wu, Yuxin and Kaiming He (2018). *Group Normalization*. DOI: 10.48550/ARXIV.1803.08494. URL: https://arxiv.org/abs/1803.08494.

Xie, Yichun, Zongyao Sha, and Mei Yu (2008). "Remote sensing imagery in vegetation mapping: a review". In: *Journal of plant ecology* 1.1, pp. 9–23.

Yu, Kunyong et al. (2022). "Comparison of Classical Methods and Mask R-CNN for Automatic Tree Detection and Mapping Using UAV Imagery". In: *Remote Sensing* 14.2. ISSN: 2072-4292. DOI: 10.3390/rs14020295. URL: https://www.mdpi.com/2072-4292/14/2/295.

Yuan, Qiangqiang et al. (May 2020). "Deep learning in environmental remote sensing: Achievements and challenges". In: *Remote Sensing of Environment* 241, p. 111716. DOI: 10.1016/j.rse.2020.111716.

Yun, Sangdoo et al. (2019). *CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features*. DOI: 10.48550/ARXIV.1905.04899. URL: https://arxiv.org/abs/1905.04899.

Zhang, Chengye et al. (2017a). "Advancing the PROSPECT-5 Model to Simulate the Spectral Reflectance of Copper-Stressed Leaves". In: *Remote Sensing* 9.11. ISSN: 2072-4292. DOI: 10.3390/rs9111191. URL: https://www.mdpi.com/2072-4292/9/11/1191.

Zhang, Hongyi et al. (2017b). *mixup: Beyond Empirical Risk Minimization*. DOI: 10.48550/ARXIV.1710.09412. URL: https://arxiv.org/abs/1710.09412.

Zhang, Xia et al. (2021). "A Study on Different Functionalities and Performances among Different Activation Functions across Different ANNs for Image Classification". In: *Journal of Physics: Conference Series*. Vol. 1732. 1. IOP Publishing, p. 012026.

Zhao, Hengshuang et al. (2017). *Pyramid Scene Parsing Network*. arXiv: 1612.01105 [cs.CV].

Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl (2019). *Objects as Points*. DOI: 10.48550/ARXIV.1904.07850. URL: https://arxiv.org/abs/1904.07850.

Zorzi, Stefano, Ksenia Bittner, and Friedrich Fraundorfer (2020). *Map-Repair: Deep Cadastre Maps Alignment and Temporal Inconsistencies Fix in Satellite Images*. arXiv: 2007.12470 [cs.CV].