VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
SOFTWARE ENGINEERING PROGRAMME

Semantic segmentation for automated annotation of satellite images

Semantinis segmentavimas automatiniam palydovinių vaizdų anotavimui

Master's Thesis

Author: Markas Mikalauskas (parašas)

Supervisor: doc. Dr. Linas Petkevičius (parašas)

Reviewer: prof. Dr. Olga Kurasova (parašas)

Vilnius – 2025

Santrauka

Šiame magistro darbe tiriami palydovinių vaizdų semantinio segmentavimo metodai su automatizuotu anotavimu, ypatingas dėmesys skiriamas šiems skirtingiems Lietuvos regionams - Kėdainių, Varėnos, Klaipėdos ir Kauno, naudojant „Segment Anything Model“ (SAM) modelį ir jo aukštos kokybės variantą (SAM-HQ). Vertinamas efektyvumas naudojant skirtingus „Vision Transformer“ (ViT) kontrolinius taškus (ViT-B, ViT-L, ViT-H). Apdorojami duomenys gauti iš Sentinel-2 palydovinių vaizdų, įskaitant debesų šalinimą panaudojant „UnCRtainTS“ įrankį, o analizei naudotas „SamGeo“ programavimo kalbos Python paketas.

Darbe modelių našumas lyginamas tiek įvesties užuominų (angl. input prompts) segmentavimo, tiek automatinio kaukių generavimo metodais, pastarajam kaip pagrindinią metriką naudojant modelio prognozuojamus „Susikirtimas padalintas iš sąjungos“ (angl. Intersection over Union) įverčius. Rezultatai rodo, kad SAM-HQ modeliai pranoksta SAM, ypač sudėtinguose ir urbanizuotuose regionuose, kur ViT-H kontrolinis taškas pasiekia didžiausią tikslumą lyginant su kitais. Vis dėlto, našumas priklauso nuo kraštovaizdžio ypatybių, kaip žemės ūkio požiūriu įvairesnis Kėdainių regionas buvo segmentuojamas patikimiau, nei tankiai miškingas Varėnos regionas. Įvesties užuominų segmentavimas pasirodė esąs mažiau patikimas nuosekliam automatiniam anotavimui. Tyrime daroma išvada, kad nors SAM-HQ siūlo tvirtą kelią link automatizuoto anotavimo, modelio ir parametrų pasirinkimas yra kritiškai svarbus ir priklausomas nuo konkretaus kraštovaizdžio. Darbe pristatoma darbo eiga ir regionams būdingos įžvalgos, taikant segmentavimo modelius Lietuvos geoerdviniams duomenims.

Raktiniai žodžiai: Semantinis segmentavimas, Automatizuotas anotavimas, Palydoviniai vaizdai, Segment Anything Model, Segment Anything Model in High Quality, Lietuvos topografija, Nuotolinis aptikimas, Gilusis mokymasis

Summary

This master's thesis investigates methodologies for the semantic segmentation of satellite images to achieve automated annotation, focusing on Lithuania's different landscapes - Kėdainiai, Varėna, Klaipėda and Kaunas regions. Research evaluates the efficacy of segmentation models, specifically the Segment Anything Model (SAM) and its high-quality variant (SAM-HQ), with various Vision Transformer (ViT) checkpoints (ViT-B, ViT-L, ViT-H). Data retrieved from Sentinel-2 satellite imagery are processed for use, including cloud removal using the UnCRtainTS tool, and analyzed using the SamGeo Python package.

The study compares model performance through both input prompt segmentation and automatic mask generation, with model's predicted Intersection over Union (IoU) scores as the primary metric for the latter. Findings indicate that SAM-HQ models generally outperform SAM with the ViT-H checkpoint often yielding the highest accuracy. However, performance varies with landscape characteristics - agriculturally diverse Kėdainiai region showed higher segmentation confidence compared to the densely forested Varėna region. Input prompt segmentation proved less reliable for consistent automatic annotation. The research concludes that while SAM-HQ offers a robust way for automated annotation, model and parameter selection are crucial and landscape dependent. The work presents workflow steps and region specific insights for applying segmentation models on Lithuanian geospatial data.

Keywords: Semantic segmentation, Automated annotation, Satellite imagery, Segment Anything Model, Segment Anything Model in High Quality, Lithuanian topography, Remote sensing, Deep learning

CONTENTS

List of Figures

List of Tables

Introduction

Satellite image segmentation is an important task in the field of remote sensing and geospatial analysis [GS22], involving the partitioning of a digital image into multiple meaningful segments. In this thesis, semantic segmentation techniques are used as a primary method for automated annotation, where annotation refers to the process of assigning class labels to pixels or regions within an image. While annotation can also refer to the manual creation of training data, the focus here is on leveraging segmentation models to automatically generate these labeled outputs for geospatial analysis.

The topography of a region, or its physical features and characteristics, plays a significant role in various applications such as agriculture, forestry, and urban planning. This thesis focuses on finding a methodology, input data and models for segmenting the diverse aspects of Lithuania's topographical landscape, which involves the use of geospatial imagery.

Lithuania is a country located in northeastern Europe, bordered by Latvia to the north, Belarus to the east, Poland to the south, and Kaliningrad Oblast (a Russian exclave) to the southwest. It has a diverse landscape that includes forests, lakes, rivers, and coastal areas of the Baltic Sea in the western side of the country. Understanding the topography of Lithuania is important for a variety of applications, including land use management, environmental monitoring, and natural resource management.

There are several challenges associated with satellite image segmentation and annotation in the context of Lithuania's topography. One challenge is the variability of the landscape, which can vary significantly depending on the location and season. This can make it difficult to accurately classify and label the different features in an image. Another challenge is the complexity of the topography, with a wide range of features that may be present in a single image. This can make it difficult to accurately segment and classify the different features. Automatic segmentation algorithms generate masks, which represent the identified regions of interest by assigning class labels to pixels. Machine learning methods have proven to be effective tools for addressing these challenges. For example, in the study [Abd20], a Convolutional Neural Network (CNN) was used to perform land cover classification on satellite images, achieving an overall accuracy of 95.2%. Similarly, in the review [YCS+22], several machine learning approaches were discussed for the task of building detection on satellite images, including decision trees, support vector machines, and neural networks. These examples demonstrate the effectiveness of machine learning methods for satellite imagery segmentation and annotation, and suggest that they will continue to be important tools in the field.

These challenges are addressed by finding an applicable one, and identifying varied satellite image data in the context of Lithuania's topographical landscape, with a focus on using open-source satellite images from Sentinel-2 to find a dataset for identifying the topographical segments and objects in the landscape. The use of machine learning and deep learning techniques [AA19] will also be explored to improve the accuracy and efficiency of these methods. Through this research, the aim is to contribute to a better understanding of Lithuania's topography by applying selected data and tools for the analysis of its satellite images.

1. Goal, objectives and expected results

1.1. Goal

The main goal of this master's thesis is to evaluate various approaches and methodologies for satellite image semantic segmentation with the objective of achieving automated annotation in the context of Lithuania's topographical landscape, identifying the optimal input data and model, which could identify and label unique objects in geospatial imagery.

1.2. Objectives

In order to achieve the main goal of this master's thesis, the following objectives will be pursued:

1. To review and analyze the existing literature on satellite image segmentation and annotation, with a focus on techniques and approaches that have been applied in the context of topographical landscape analysis;
2. To identify and collect relevant open-source satellite images from Sentinel-2 for use in the study;
3. To investigate various techniques and approaches applicable to said satellite images, employing segmentation and automated annotation techniques on those datasets;
4. To evaluate the performance of the used techniques and approaches using quantitative metrics, such as Intersection over Union because of its widespread use and effectiveness in quantifying overlap accuracy for segmentation tasks;
5. Present schemes, recommendations and guidelines how it could be applied practically.

1.3. Expected results

Based on the objectives outlined above, the following are the expected results of this master's thesis:

1. A comprehensive review of the existing literature on satellite image segmentation and annotation, with a focus on techniques and approaches that have been applied in the context of topographical landscape analysis;
2. Open-sourced satellite images from Sentinel-2 suitable for use in the study;
3. Identification of optimal input data and models for identifying objects in said satellite images using geospatial segmentation techniques tailored to the specific characteristics of Lithuania's topographical landscape;
4. Evaluation of results showing the performance of the used approach.

1.4. Research methods

The following steps will be followed to research satellite imagery of Lithuania's topographical landscape:

1. Review the scientific literature and conduct a thorough review of the literature on satellite imagery, topographical analysis, and relevant methods and techniques. This will help to establish the current state of knowledge on the topic and identify any gaps or areas for further investigation;

2. Collect satellite imagery data through free sources such as the ESA (European Space Agency's) Copernicus program Sentinel-2;

3. Pre-process the satellite imagery data to correct atmospheric effects or georeference the imagery to align it with a map. Then prepare the dataset as needed for analysis through the use of semantic segmentation and annotation techniques;

4. Conduct analysis of the satellite imagery dataset using various techniques and approaches, a few examples would be LuNet or other algorithms that would apply to image segmentation scenario using machine learning methods [KMC+22]. This may involve comparing different techniques and evaluating their performance based on various metrics. The main factors for consideration will be accuracy, efficiency, and robustness, applying a range of techniques and methods, such as feature extraction, pattern recognition, or spatial modeling;

5. Validate the results of the analysis to ensure that they are accurate and reliable. One way to achieve this is to compare the results to other sources of data, such as topographic maps, field observations or statistical methods to assess the robustness of the findings;

6. Present the implications of the findings, conclusions with maps and figures, what if there were any limitations or problems along the way and suggest areas where the dataset or the technique could be replaced with another one in the future.

1.5. Research direction

The focus of this thesis is on the analysis of satellite imagery of the topographical landscape of Lithuania using advanced algorithms and techniques. The research aims to identify an effective method for analyzing this type of data to extract valuable insights and information about the topographical features of the landscape. To achieve this goal, the study involves a thorough review of the literature on satellite imagery analysis, followed by a practical comparison of different algorithms and techniques for topographical analysis and evaluating segmentation models, specifically Segment Anything Model (SAM) and its high-quality variant (SAM-HQ). This provides a practical workflow and region-specific insights into their efficacy for automated annotation. The direction of the research is to contribute to the understanding of the topographical features of the Lithuanian landscape and to provide a robust, evaluated method for analyzing similar datasets in the future.
Collection of open-source satellite images:

1. European Space Agency. Sentinel-2: High-resolution optical imagery. Retrieved from https://earth.esa.int/web/sentinel/missions/sentinel-2;

2. Google. Google Earth API. Retrieved from https://developers.google.com/earth;

3. Copernicus. Retrieved from https://www.copernicus.eu;

4. Satellite imagery datasets from MIT Licensed GitHub repository. Analyzed and retrieved from: https://github.com/chrieke/awesome-satellite-imagery-datasets.

2. Literature Review

2.1. DeepLab Architectures

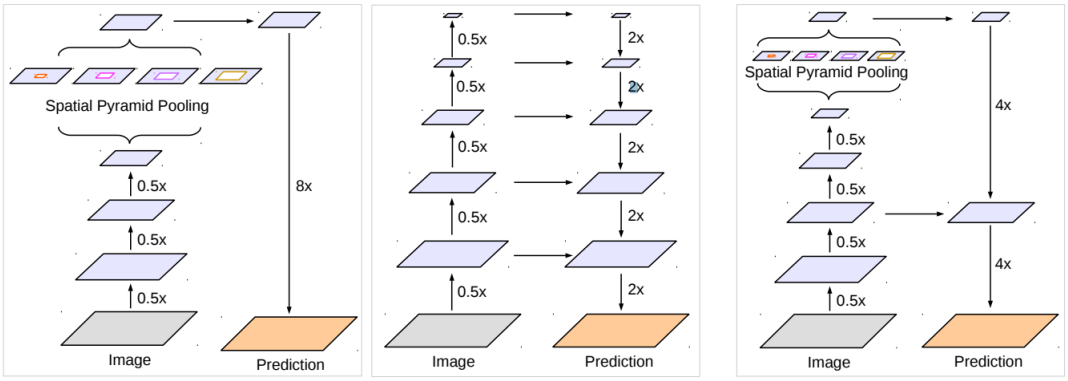
Using deep learning is one of the main ways to do tasks regarding semantic segmentation and annotation. DeepLab is a family of deep learning models known for their proficiency in semantic segmentation and one of the currently widely used architectures are DeepLabV3 and its enhanced version DeepLabV3+.

2.1.1. DeepLabV3

DeepLabV3 employs atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) to capture context information at different scales (figure 1 **a)** example). Atrous convolutions, also known as dilated convolutions, modify the effective field-of-view of convolutions without increasing computation or parameters and is used to capture multi-scale context information. The combination of these techniques allows DeepLab models to address the issue of information loss and low-resolution predictions that often occur with Fully Convolutional Neural Networks (FCNs) (figure 4). It uses a modified ResNet architecture for the encoder and incorporates skip connections in the decoder for refining segmentation output. Despite its computational intensiveness in training, DeepLabV3 surpasses models like FCNs and U-Nets in capturing extensive context and extracting features across scales. It offers a blend of high performance and architectural advancements [CPS+17].

2.1.2. DeepLabV3+

DeepLabV3+ is a further enhancement of the original model, offering a fully-convolutional solution for multi-class semantic segmentation, the proposed model as mentioned in a study by Chen et al. It uses a ResNet50 backbone and an encoder-decoder structure (figure 1 **b)** example). The model combines low-level features from the backbone with upsampled encoder features to refine segmentation results along object boundaries. It is implemented in TensorFlow and has been trained on the Crowd Instance-level Human Parsing dataset. The spatial pyramid pooling module used in DeepLabV3 and DeepLabV3+ introduces a decoder module to enhance semantic segmentation with full implementation (figure 1 **c)** example). It employs the Xception model and depthwise separable convolution, resulting in a faster and stronger encoder-decoder network. The effectiveness of the model is evidenced by high test set performance on the PASCAL VOC 2012 and Cityscapes datasets, achieved without any post-processing [CZP+18].



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Figure 1. DeepLabV3 and DeepLabV3+ [BLY+21]

Both DeepLabV3 and DeepLabV3+ are powerful tools for semantic segmentation, and their use is recommended where high precision and extensive feature extraction across scales are required. Their use of atrous convolution and ASPP allows them to capture detailed context information from different scales and achieve high performance on various datasets. They are available for public use using PyTorch's framework.

2.2. Semantic segmentation and annotation

Semantic segmentation is a process within the computer vision field that partitions an image into various segments or regions, where each segment corresponds to a specific category or class (figure 2). In the context of this thesis, the output of semantic segmentation - the assignment of a class label to each pixel is considered automated annotation. Furthermore, the main objective of semantic segmentation is to understand the image at a pixel level - each pixel gets assigned a label, that later on could be grouped into a specific class or category. This is different from image classification, where the entire image is assigned, and unlike traditional image segmentation methods that group together similar pixels based on color or texture, meaning - semantic segmentation classifies each pixel of an image based on the category it belongs to in the context of the entire image.

It is quite a dense prediction task, since after assigning every pixel with a label, then these labels need to be categorized into different classes. Some of the class examples would be - cars, dogs, buildings or any other kinds objects that are inside an image would then be sampled into their respective different classes. However, there is a limitation - it does not differentiate between separate instances of the same object class. For instance, if the image contains two trees, all of their pixels that make the tree would be labeled as 'tree', it would not separate them into their own separate identifiable objects like tree one and tree two. This limitation could be addressed using instance segmentation, with which we would be able to count the amount of unique objects in the image. But in this case, instance segmentation will not be of any use, since we will not need to count the amount of objects in the image, in this case, simply accurately detecting objects within the presented geospatial areas of Lithuanian topography.

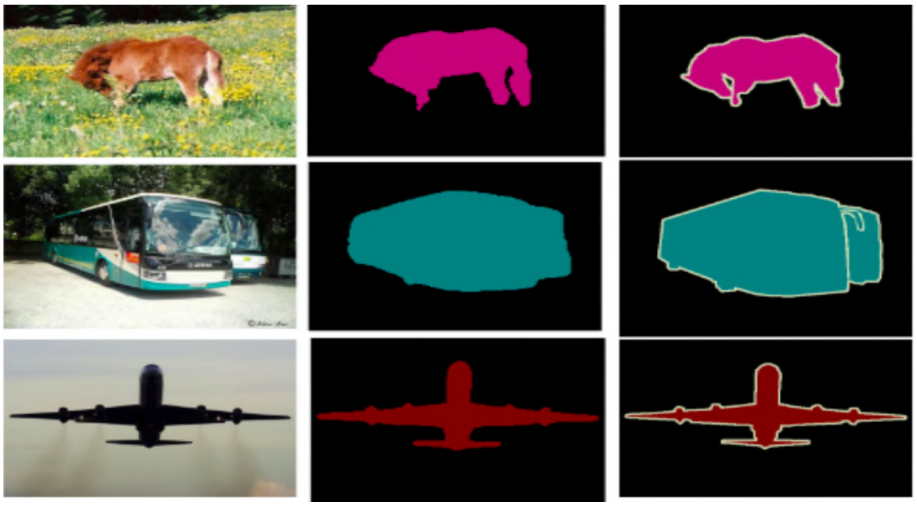


Figure 2. Semantic segmentation [GLG+18]

2.2.1. Segmentation and annotation models

The three most recent semantic segmentation models can be divided into three categories [GLG+18]:

1. Region-based segmentation;
2. FCN (Fully Convolutional Network) based segmentation;
3. Weakly supervised segmentation.

2.2.1.1. Region-based segmentation

Region-based segmentation is used for partitioning an image into multiple regions (or sets of pixels) (figure 3). It can be applied with two primary approaches - region growing, and region splitting and merging. Region growing approach is when the algorithm picks out a seed point in the image and then starts adding the neighboring pixels to the same region if they are similar based on the seeds criterion, like for example - color or intensity, then this process proceeds until there are no more pixels to add to the region set.

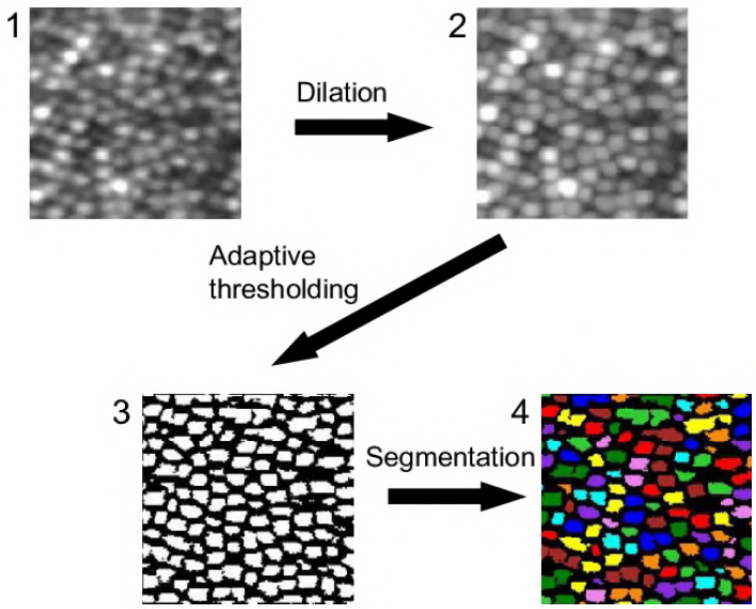


Figure 3. Region-based segmentation [MFH+17]

2.2.1.2. FCN-based segmentation

FCN-based segmentation or Fully Convolutional Network segmentation (figure 4) is an applicability extension of traditional Convolutional Neural Networks (CNN) from just an image classification method to a dense prediction model, which in abstract can be called a semantic segmentation task, where its goal is to predict the category or class of every pixel in the image, without extracting the region proposals [LSD15]. Since FCNs are composed of convolutional, pooling and upsampling layers, depending on the definition of a loss function, they can be end-to-end trainable [GLG+18]. The main issue with FCN approach is that it is propagated through several alternated convolutional and pooling layers, thus the resolution of the output feature maps is down-sampled. Which means, the direct predictions of FCNs are typically in low resolution, therefore it is quite not applicable if you want high-resolution prediction results.

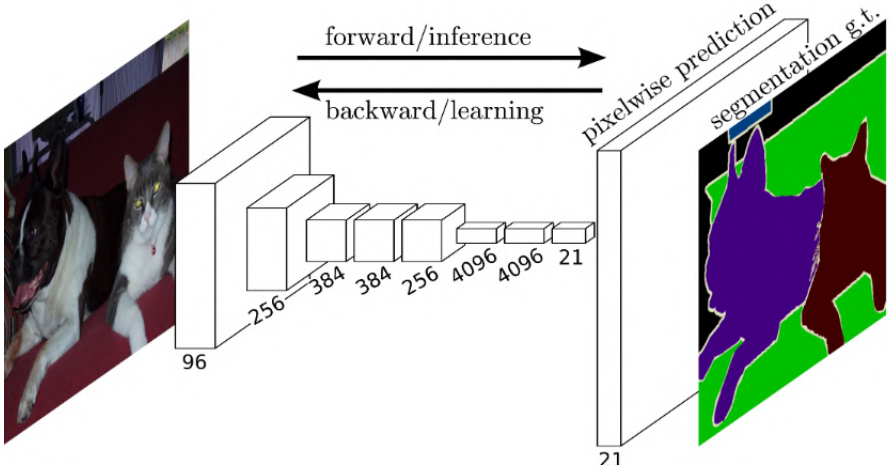


Figure 4. FCN-based segmentation [LSD15]

2.2.1.3. Weakly supervised segmentation

Weakly supervised segmentation is a method where the learning process is guided by a weaker form of supervision than fully annotated training data (figure 5). This is useful, since obtaining fully annotated data, such as pixel-wise segmentation masks for every image in the training set, can be labor-intensive and costly. The models are then trained using less detailed annotations, in a form of:

1. **Image-level labels** - each image in the training set is labeled with the classes it contains, without the source or details of those classes;
2. **Bounding boxes** - each image is annotated with bounding boxes, that indicate the location of objects without providing the exact pixel-level boundaries;
3. **Scribbles or Points** - for each identified object in the image, a scribble or a point is given as an annotation;
4. **Partial labels** - only a subset of the image is annotated at the pixel level.

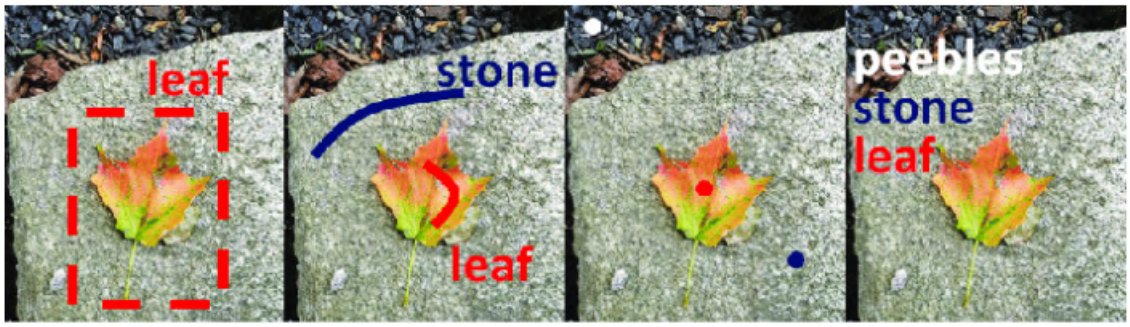


Figure 5. Weakly supervised segmentation [KOB21]

Bounding boxes (left most figure in figure 5), are annotations that define rectangular regions around specific objects or areas of interest within an image. These annotations provide precise localization information, allowing models to identify and locate objects accurately. Bounding boxes are commonly used in object detection, object localization, and object tracking tasks. By specifying the coordinates of the top-left and bottom-right corners of the bounding box, annotators can indicate the position and size of objects within an image.

Scribbles or points annotations (second from the left most figure in figure 5), on the other hand, involve marking specific areas or points of interest within an image. Scribbles are typically freehand or loosely drawn lines that approximate the boundaries of objects or regions, while points represent precise locations. These annotations are often used where the objective is to assign a label to each pixel in the image. This technique provides more detailed information about the spatial extent of objects or regions, allowing models to learn fine-grained boundaries.

Partial labels (second from the right most figure in figure 5) refer to annotations where only a subset of the available labels are assigned to an image. Instead of annotating all objects or regions in an image, annotators label only certain objects or regions of interest. This can also be useful

when annotating large datasets or when the complete annotation of all objects is not necessary for a particular task. This approach reduces annotation effort and can be leveraged in tasks such as object recognition or scene understanding, where a subset of labeled objects is sufficient for training models.

Image-level labels (right most figure in figure 5) are annotations that assign a single or multiple labels to an entire image. They provide a high-level understanding of the image content and help categorize images into different classes or categories. For example - one picture is given where some kind of a dog is present, the only label the model will have to work with will be named Dog. This technique is useful for tasks such as image classification, where the goal is to predict the most relevant label or labels for an image based on its overall content. These types of annotations offer a coarse-grained representation of the image's content and are often used to build large-scale image datasets.

2.2.2. Thoughts and concerns

Semantic segmentation and annotation are complex tasks in machine learning, it does come with its challenges, there is a need to address complex lighting and weather conditions such as cloudiness in real-world applications [MEZ+20]. Active learning methods for semantic segmentation are also gaining interest, particularly where annotations are expensive and datasets exhibit varying redundancy levels. Semantic segmentation finds applications across various sectors, including land cover mapping [SPE+20], which is relevant in the field when land mass objects are involved, like trees and other vegetation. Benchmarking through various semantic segmentation and annotation techniques with the use of mentioned DeepLabV3 architectures will be beneficial to find the best possible way to aid in satellite imagery analyzation.

2.3. Geospatial Data

A significant body of literature underscores the essential role of comprehensive, high-quality open-source satellite images for accurate Earth objects detection. Sentinel-2, stands out in this domain. This platform offers free, high-resolution multi-spectral imagery for gathering datasets, which is a great ground truth base, and is a prime choice due to its ability to capture fine-scale details necessary for semantic segmentation and annotation use in analysis Lithuanian topography.

2.3.1. Sentinel-2

Most researchers in remote sensing field are using Sentinel-2 satellite data for use in studying global topography, so considering how much research and scientific papers has been published - it is the best candidate for collecting image data for this thesis.

Sentinel-2 consists of two satellites – Sentinel-2A and Sentinel-2B, launched in 2015 and 2017 respectively. These satellites are tasked with providing multi-spectral imagery for the Earth's land surfaces, large islands, and coastal and inland waters, delivering data every five days, offering a global coverage of the Earth's land surfaces.

The data produced by Sentinel-2 is one of the most widely used datasets in remote sensing and Earth observation research. An example application would be Segarra et al. study - usage in precision agriculture, where Sentinel2-A + B twin data was used for framing remote sensing principles in the field for helping further research and development in agriculture [SBA+20]. This wide usage is mainly due to the high-resolution multi-spectral imagery it provides, with a spatial resolution of up to 10 meters, and its 13 spectral bands covering visible, near-infrared, and short-wave infrared light spectrum. These attributes make it an incredibly valuable resource for a variety of applications, including land cover mapping, vegetation health monitoring, disaster management and monitoring.

This data offers considerable advantages due to its frequency, continuity, and free availability. With a five-day revisit time at the equator and higher revisit rate at high latitudes, changes in land cover and vegetation health can be monitored consistently. This is particularly crucial for tracking seasonal changes, or for rapid response in the event of natural disasters. The continuity of data collection ensures that long-term trends can be accurately assessed, which is vital for having as much imagery as you can. One of the unique features of this programme is its open data policy, allowing researchers and public bodies worldwide free access to its datasets, in return it has democratized the use of satellite data, enabling small organizations and even individuals to conduct analyses that were previously an exclusive domain of well-funded institutions.

2.3.2. Cloud removal in satellite images

To improve the quality and usability of aforementioned satellite imagery gathered from open-sourced data like Sentinel-2 programme - removing cloud obstructions is necessary. In remote sensing, data accuracy and reliability plays a crucial part, and clouds are one of the main challenges when gathering reliable data that needs to be addressed to have accurate calculations, predictions and measures. We will have an overview of currently available methodologies that combat this issue.

2.3.2.1. Synthetic Aperture Radar

Before moving forward, we need to understand what Synthetic Aperture Radar (SAR) is. It is a radar technique that creates high-resolution images by leveraging the motion of the radar antenna (figure 6). Its application is widespread in remote sensing, mapping, and monitoring various phenomena, including forestry, volcano monitoring, civil infrastructure stability, and military surveillance. It emits electromagnetic signals towards a surface and records the echoes, combining the data from multiple positions to create a virtual antenna larger than the physical one, thereby enhancing image resolution [NBB+20].

SAR operates in different bands, like X, C, L, and P, each corresponding to specific frequencies and affecting the depth of signal penetration. The polarization mechanism of SAR reveals scattering attributes, such as rough surface, volume, and double bounce. It also has the advantage of working in longer wavelengths, enabling it to penetrate clouds and providing reliable measurements irrespective of weather conditions and darkness.

Signal processing in SAR involves matched filtering, spectral estimation techniques, and using a digital elevation model for 3D reconstruction. Different methods like multidimensional indexing, Capon method, APES method, SAMV method, and the backprojection algorithm are employed, each with its own advantages, disadvantages, and computational complexity. Its polarimetry is used to analyze scattering properties for material characterization.

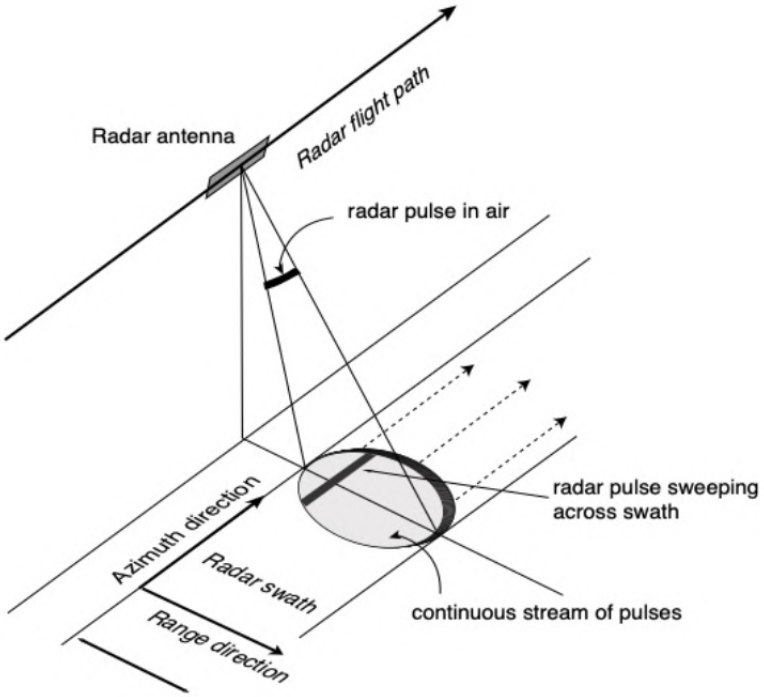


Figure 6. Synthetic Aperture Radar (https://nisar.jpl.nasa.gov/mission/get-to-know-sar)

2.3.2.2. Interferometric SAR

Interferometric SAR (InSAR) is a specific technique within SAR that extracts terrain altitude and motion information by analyzing phase differences. Techniques such as differential interferometry, Tomo-SAR (providing a 3D perspective), ultra-wideband SAR (which offers improved resolution and spectral information), and Doppler beam sharpening further enhance the capabilities of SAR. Data processing in SAR often involves Fourier transform techniques, and the resolution can be affected by coherence speckle effects [Ric07].

Understanding and interpreting SAR images necessitates comprehension of the side-looking geometry and consideration of factors like surface roughness, slopes, layover, polarization, urban areas, surface variations, and human-made objects. Nevertheless, SAR provides invaluable data for numerous applications, making it an essential tool in the field of remote sensing and Earth observation.

2.3.3. SEN12MS-CR

Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion is a paper published by Meraner et al., which was awarded the ISPRS 2020 Best Paper Award and the U.V. Helava Award. This approach is encapsulated in a deep learning model, known

as DSen2-CR or SEN12MS-CR, designed for removing clouds from optical remote sensing images [MEZ+20].

The key features of this model include the use of a deep residual neural network architecture, the fusion of SAR and optical data. The model also introduces a novel cloud-adaptive loss function, as it aims to retain the original information of the image and reduce translation artifacts. The model was trained and evaluated using the SEN12MS-CR dataset, which consists of heavily cloudy and cloud-free images. The dataset facilitated the removal of optically thick clouds and demonstrated the models ability to improve the structure and quality of the reconstructed images compared to methods that only use optical data.

The DSen2-CR model code and its PyTorch implementation are publicly available on Github. Instructions for installation and usage are also provided in the repository, and the authors recommend running the code with GPU support for better performance. It also provides an example and expected results in figure 7, where the left most image is the input cloudy image, middle one is the predicted image, and right most one is the cloud-free target image.

Figure 7. DSen2-CR using the CARL loss results [MEZ+20]

This work emphasizes the importance of cloud removal in Earth observation applications and contributes to the wider understanding of deep learning methods for cloud removal, generative adversarial networks, SAR-optical image synthesis, and remote sensing.

Another interesting work in the same area is the G-FAN model which uses a combination of ResNet and Graph Attention Network (GAT) for cloud removal in optical remote sensing imagery. It extracts non-local features to address the limitations of SAR images and employs multi-head graph-based feature aggregation modules (M-GFAM) for cloud removal, image de-blurring, and image de-noising [CZL+22].

2.3.4. SEN12MS-CR-TS

SEN12MS-CR-TS is a novel, publicly accessible remote sensing dataset designed to address the challenge of cloud removal and optical image reconstruction in the field of Earth observation. This multi-modal and multi-temporal dataset pairs radar measurements from Sentinel-1 and multi-spectral optical satellite observations from Sentinel-2, covering various globally distributed regions of interest [EXS+22]. It is an extension, pre-processed and has backwards compatibility with the

aforementioned SEN12MS-CR model, but in comparison, it is multi-modal (using both optical and radar data) and multi-temporal.

The dataset provides temporally aligned images throughout the year 2018, with 30 time samples for each of the 15,578 patch-wise observations. These temporal and spatial features help in providing diverse and comprehensive data to address cloud removal challenges effectively. Two models are proposed to be used with this dataset:

1. A multi-modal multi-temporal 3D-Convolution Neural Network (3D-CNN), which is designed to predict cloud-free images from a sequence of cloudy optical and radar images;

2. A sequence-to-sequence translation model that is intended to predict cloud-free time series from cloud-covered ones.

These models leverage the unique multi-modal and multi-temporal properties of the dataset, demonstrating the advantages of using such diverse data for reconstructing noisy or obscured images. Several models, such as the CR-TS Net and ensemble methods, have been trained and benchmarked using SEN12MS-CR-TS, showing improved performance in terms of metrics like Spectral Angle Mapper, Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). These trained models are freely available for download to facilitate and promote research in cloud removal and optical image reconstruction.

2.3.5. UnCRtainTS

UnCRtainTS is another novel model developed for the purpose of multi-temporal cloud removal in optical satellite images. It involves a combination of an attention-based architecture and a formulation for multivariate uncertainty prediction (figure 8). Its primary function is to address the challenge of cloud obstruction in remote sensing and improve the quality of optical satellite image reconstruction. With its ability to achieve state-of-the-art performance on two public cloud removal datasets, it is a significant development in the field of remote sensing technology [EGS+23].

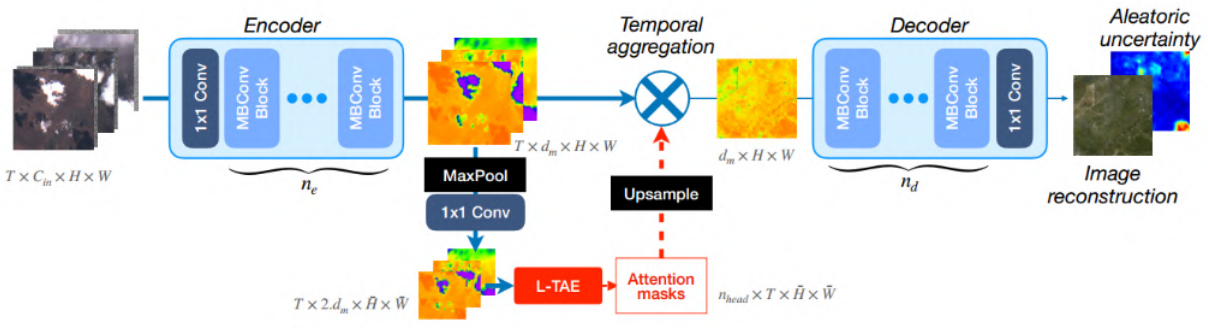


Figure 8. UnCRtainTS architecture [EGS+23]

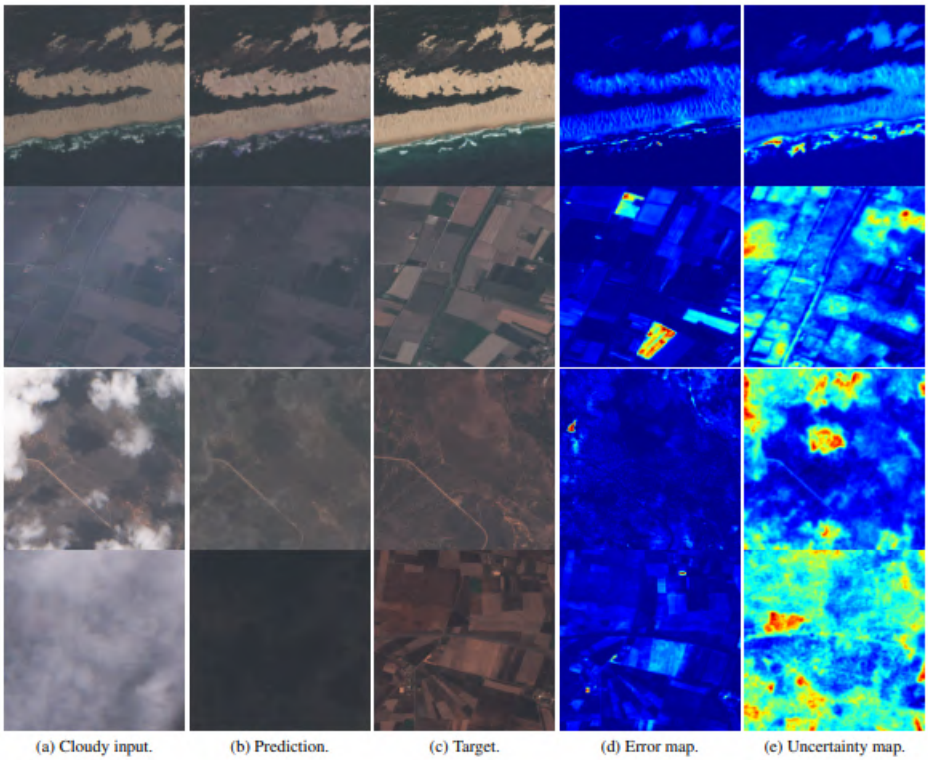


Figure 9. UnCRtainTS results [EGS+23]

One of the standout features of this model is its ability to predict uncertainties in a well-calibrated manner, which allows for precise control of the reconstruction quality. This kind of uncertainty quantification is also of importance in other imaging fields such as medical imaging and geophysical inversion techniques. By managing to effectively address the noise in MRI images and prevent information loss in data reduction, the method exhibits significant potential for advancing these technologies. In figure 9, it shows how **a)** ground truth input of a cloudy image is being **b)** predicted by the architecture, to reach the target as shown as **c)**, target result with an output of the **d)** error map and **e)** uncertainty map from the given image.

This model introduces new methodologies in the area of SAR-to-optical image translation using a hybrid cGAN, and model and data dimension reduction in geophysical inverse problems. All these capabilities make UnCRtainTS a promising and versatile tool in the broader field of image reconstruction and analysis, especially when Republic of Lithuania's area is close to the Baltic Sea and has a median cloudiness over 40% per year.

2.4. Segmentation Models

2.4.1. SamGeo tool

Bridging the gap between advanced general-purpose segmentation models and specific geospatial applications are tools designed to facilitate their use with geographic data. One such notable tool is SamGeo [WO23], a Python package developed to streamline the application of the Segment Anything Model (SAM) and its variants to geospatial raster data, such as GeoTIFF satellite images. SamGeo provides functionalities for loading geospatial data, running SAM-based segmentation,

and exporting the results in formats suitable for Geographic Information System (GIS) workflows. Its development significantly lowers the barrier to entry for utilizing these powerful segmentation capabilities in remote sensing and environmental analysis, making it a relevant instrument for practical applications like those explored in this thesis.

2.4.2. Vision Transformer Models

Vision transformer models (ViT) are used for analysis when using in conjunction with the SamGeo tool. These models represent different scales of neural network architectures, and they are mainly utilized with aforementioned SamGeo, which out of all them, three main ones stand out and were used for the following analysis:

- **ViT-B (base)** - the smallest among the three and is often used as a starting point for various applications. The model balances performance with computational efficiency, making it suitable for a broad range of applications;
- **ViT-L (large)** - larger than the base model, offers increased performance at the cost of higher computational requirements, it is designed for tasks where higher accuracy and more complex model capabilities are needed;
- **ViT-H (huge)** - the largest model out of all three, providing the highest performance among the three. It is used for the most demanding tasks, but requires significant computational resources.

The main differences among these models are their size and complexity, which impact their performance and computational requirements. Generally, the larger models achieve better accuracy at the expense of increased computational needs. These types of models can be analyzed through Segment Anything (section 2.4.3) and Segment Anything in High Quality Models (section 2.4.4) for versatile comparison through the use of SamGeo tool.

2.4.3. Segment Anything Model

SAM is a groundbreaking development model, introduced as part of company Facebook (Meta AI) Segment Anything research project. This model is used for generating object masks for all objects in an image, and it has been trained on a dataset of 11 million images and 1.1 billion masks, and has strong zero-shot performance on a variety of segmentation tasks [KMR+23]. This model utilizes Vision Transformer model checkpoints (section 2.4.2), which are built in by default in SamGeo tool, and will be used for comparison with other models. SamGeo tool leverages and is built on top of the SAM model, offering the readily available environment for semantic segmentation analysis.

2.4.4. Segment Anything in High Quality Model

Built on top of SAM foundations, SAM-HQ brings enhanced quality and precision to image segmentation tasks. SAM represents a big leap in scaling up segmentation models, allowing for powerful zero-shot capabilities and flexible prompting. Despite being trained with 1.1 billion masks,

SAM's mask prediction quality falls short in many cases, particularly when dealing with objects that have intricate structures [KYD+23]. So they have made a better prediction model where they composed a dataset of 44 thousand fine-grained masks from several different sources, which resulted in better predictability of image segmentation in comparison to SAM. Using the aforementioned Vision Transformer model checkpoints base, large and huge (section 2.4.2), they have composed a graph of how in zero-shot capabilities it is superior to SAM (figure 10).

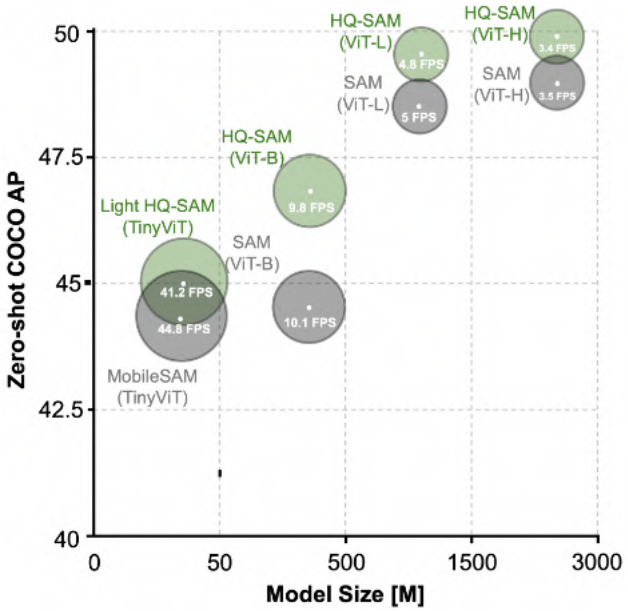


Figure 10. SAM and SAM-HQ comparison [KYD+23]

2.5. Conclusions

Semantic segmentation leading to automated annotation of satellite images, is crucial for understanding Earth's topography. The literature review revealed that semantic segmentation techniques partition an image into semantically meaningful parts, classifying each part into pre-set classes.

Deep learning methodologies like active learning strategies that optimize annotation efforts and manage diverse data distributions, specialized loss functions that promote faster and improved model convergence, and resources like SemSegLoss help address key challenges in satellite image segmentation and annotation. Deep learning's ability to learn and understand features from training data, coupled with its capacity for high-precision image processing and contextual understanding, has markedly improved accuracy and reduced manual labor. The combined application of these methodologies with established architectures like DeepLab, alongside emerging models like SAM with ViT and its variants, promises feasible and effective results in semantic segmentation and automated annotation tasks.

The Sentinel-2 program provides essential multi-spectral imagery, delivering data every five days with global coverage of Earth's land surfaces and its importance for data acquisition cannot be overstated. This extensive image database, when used with innovative datasets and models for cloud removal like UnCRtainTS is critical for resolving the challenge of cloud obstruction. This, in turn, ensures the availability of clear, reliable data for analyzing and detecting objects within

Lithuanian topography.

The combined use of these reviewed methods, techniques, and tools like SamGeo contributes invaluable information for detailed, efficient, and accurate topographical understanding. Their application holds significant potential for the analysis of topographical data.

3. Key decisions

This section outlines the crucial decisions and practical applications made in the process of segmenting and annotating Lithuanian topography using satellite images. These key decisions span from data selection to the application of advanced segmentation techniques. The choice is discussed and use of Sentinel-2 imagery, emphasizing its suitability for a detailed topographical analysis of Lithuania, then move on to retrieved data to process it, detailing the steps taken to enhance image quality and prepare the data for segmentation, including spectral band selection, quality adjustments, file formatting and for some regions - cloud removal. The core methodology involves the use of SamGeo (section 2.4.1) tools, with special emphasis on models like SAM [KMR+23] and SAM-HQ [KYD+23]. How these models are applied is explored on processed data, particularly in the diverse regional landscapes and urban areas of Kėdainiai, Varėna, Klaipėda and Kaunas.

3.1. Data Source

Using the Copernicus Dataspace browser, it is possible to select any region for analysis and whilst browsing through Lithuania's topography. Regions were selected based on their diverse characteristics - Kėdainiai region features numerous small forest plots, extensive farmland, and urban areas. While Varėna region is quite dense in vegetation and forests with noticeable water bodies like Glėbo hidrographical reserve. Klaipėda and Kaunas city regions were selected for comparison of areas with high urbanization and diversity, especially with their parks and visually different urban areas, with a note for Klaipėda having access to the coastline along with Curonian Lagoon.

These regions were selected as candidates for representing the broad diversity in Lithuanian topographical landscape, since they would prove to be a challenge for the models due to how Lithuania is lush in greenery during the warmer seasons, but also diverse other aspects as well, like open fields, urban areas, water bodies and other objects from a satellite's point of view.

3.2. Retrieval and Processing

Copernicus Dataspace browser provides possibilities to select which satellite imagery would be of interest, and it provides a possibility to select data collections from specifically Sentinel2-L1C or Sentinel-2-L2A satellites, with advanced built-in tooling it is possible to do a time series variation for selecting from both satellites data collections for gathering required data. So the selection process was as follows:

- Viewing the regions and selecting singular and series of images of a few days;
- Selecting Sentinel-2-L2A as the main data collection;
- Selecting the best candidates considering default sharpness, depth and contrast;
- Create a custom visualization of data collections, using the dataspace built-in feature to select from both Sentinel-2-L1C as sub-data for retrieval;

- Since data models for both cloud removal and running data through segmentation requires to have red-green-blue imagery as initial data, a combination of *B04 (red)*, *B03 (green)* and *B02 (blue)* bands with some corrections provide natural and true color schemes.

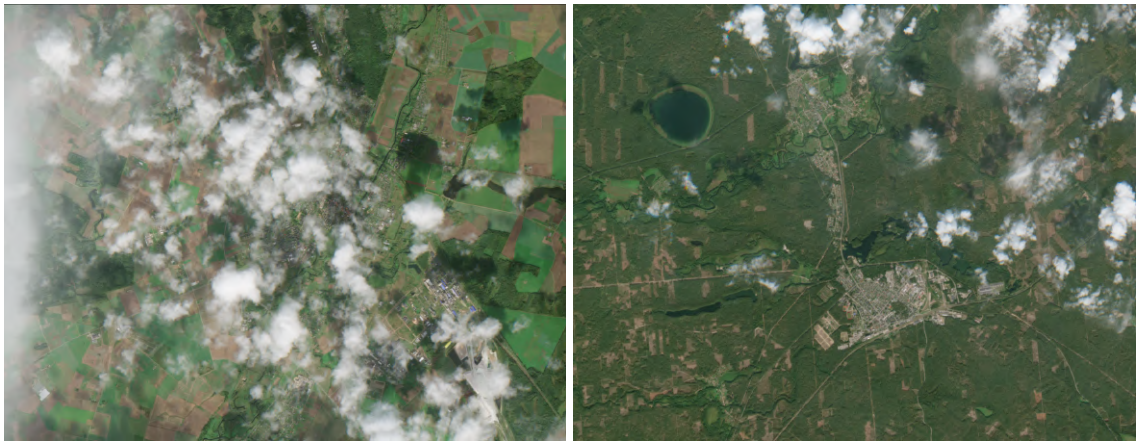
After selecting the bands, the built-in custom scripts lets adjust the finalized image parameters of gamma, saturation and reflectance of bands, with which default parameters of maximum *reflectance* of **3.0**, *saturation* of **1.2** and *gamma* of **1.8** seemed to provide with best possible results. Then, a rectangular geographical area of $100km^2$ was selected for retrieval for all four regions.

3.2.1. Processing for Segmentation

Copernicus Dataspace also provides options to download Basic, High-Resolution data or Analytical data. Analytical data format was selected as it is required by the chosen SamGeo tool (section 3.3), which requires the images to be in GeoTIFF format of 8 bits color scheme, with main layer being RGB *(Red-Green-Blue)* for analysis. These parameters were selected for downloading the data:

- Image file format of *8-bit* color scheme GeoTIFF;
- Image resolution of *1789* x *1725* pixels;
- Using compatible WGS 84 (*EPSG:4326*) coordinate system with a resolution of latitude of *0.2* seconds/pixel and longitude of *0.3* seconds/pixel;
- Selecting the aforementioned *B04*, *B03*, *B02* combination of raw data bands for RGB main layer.

For de-clouding experimentations, Kėdainiai and Varėna regions data were downloaded as shown in figure 11. Additionally, as it is rare for the country region, cloudless data was found for Klaipėda and Kaunas regions and was selected for use in experimentations as shown in figure 12.



(a) Kėdainiai (b) Varėna

Figure 11. Before de-clouding

(a) Klaipėda (b) Kaunas

Figure 12. Cloudless data

3.2.2. Cloud Removal

For the select two regions of Kėdainiai and Varėna, removing clouds from retrieved images is crucial since it impacts the results for use in satellite imagery segmentation and annotation. PatrickTUM's tool UnCRtainTS (section 2.3.5) was the owners latest model for removing clouds from images, this was selected as the primary technique to remove clouds from downloaded Copernicus data. The process went as follows:

- Retrieve multi-temporal *SEN12MS-CR-TS* European landscape dataset with a shell script;
- Clone the GitHub repository (source - https://github.com/PatrickTUM/UnCRtainTS);
- Build a *Python anaconda3* environment from included *.yaml* environment file;
- Activate the environment and install all required libraries;
- With an included python shell script, load the downloaded dataset, load satellite images for removing cloud from pre-trained model and then select an output directory.

After following these steps, the output resulting in de-clouded data was successful as can be seen in figure 13.

(a) Kėdainiai (b) Varėna

Figure 13. Declouding post-processing

3.3. Semantic Segmentation and Annotation Tool

After retrieving and processing the input data for use for segmentation, it is necessary to briefly overview workflow of the selected SamGeo (section 2.4.1) tool for use for practical analysis. Its selection was based on several factors: it is open-source, utilizes pre-trained models (thereby not requiring end-user model training), offers open access to these models and their checkpoints, is easily accessible, and does not necessitate significant computational power or resources for generating segmentations. Its main focus is on automatic semantic segmentation, which is the crucial point for this analysis. Useful features that are provided by this tool:

- Efficient and handy environment using *Jupyter Notebooks*;
- Segment *GeoTIFF* files using SAM (section 2.4.3) and SAM-HQ models (section 2.4.4);
- Interactively segment and annotate satellite images using background and foreground point markers for selecting specific area of concern;
- Possibility to tweak segmentation parameters for achieving results based on need;
- Generate and visualize segmentation results by displaying on interactive user interface elements.

3.4. Practical Application

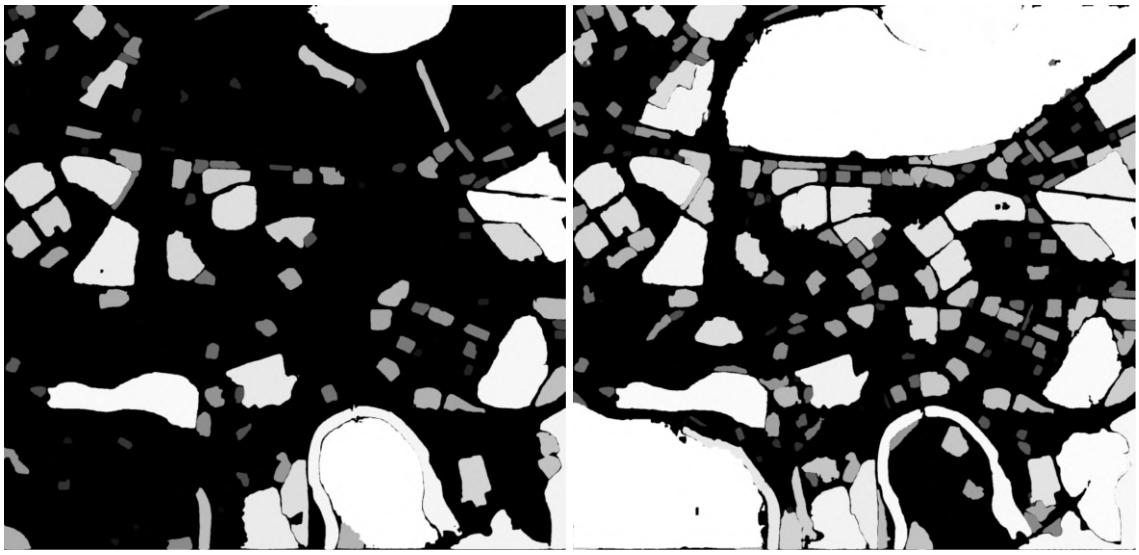
This section provides detailed description of practical analysis made on ready to use processed satellite data from Copernicus Dataspace for the four selected Lithuanian regions for semantic segmentation. Comparing the results from visual transformer models base, large and huge checkpoints (section 2.4.2), with the use of SAM and SAM-HQ models utilizing SamGeo tool (section 2.4.1). The experiment results are presented at each section using images and the performance of mentioned models and checkpoints is evaluated by comparing results from generated masks of specific

semantic segmentation methods with output statistics of resulting predicted IoU scores at the results and conclusions section.

For alignment and consistency of the analysis, these parameters were used for both SAM and SAM-HQ mask generation and annotation:

- points_per_side: 32;
- predicted_iou_threshold: 0.86;
- stability_score_threshold: 0.92;
- crop_n_layers: 1;
- crop_n_points_downscale_factor: 2;
- min_mask_region_area: 100.

The importance of using parameters for segmentation can be seen in the example when segmenting Kaunas city region, comparing the contrast of number of unique objects identified in the input data (figure 14).



(a) Without parameters (b) With parameters

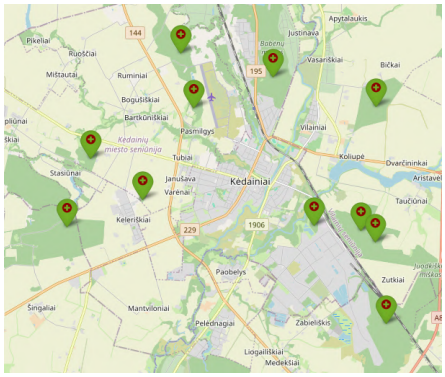
Figure 14. Comparison of parameters usage for SamGeo

3.4.1. Input prompts

Input prompts segmentation serves as initial cues for the model to distinguish target objects for segmentation (green markers) from the background (red markers for reducing noise), with the model then propagating these labels to segment the entire object based on learned visual similarities. It was selected for testing out and experimenting with SamGeo tools capabilities. Only Kėdainiai and Varėna regions were selected for these experimentations. The primary role of these input prompts is to provide crucial initial guidance to the segmentation model, enabling more targeted and accurate delineation of specific objects of interest, particularly in complex scenes or when default automatic segmentation might yield ambiguous results.

3.4.1.1. SAM

Kėdainiai municipality region was first to be used for input prompts segmentation. The target objects were larger forest bodies that are indicated as green foreground markers in figure's 15 **a)**, these included areas like Babėnai, Zutkiai, Keleriškiai and other regions, which geospatially could be labelled as forests. After the segmentation has been initialized, identical red background markers in figure's 15 **b)**, **c)** and **d)** were added so that the input annotations would be identical to each other throughout all models and their checkpoint analysis. When comparing the three, it can be seen that ViT-B model checkpoint has segmented forests most accurately in comparison to others, which was unexpected due to ViT-B checkpoint being the smallest out of the three used Visual Transformer model checkpoints.

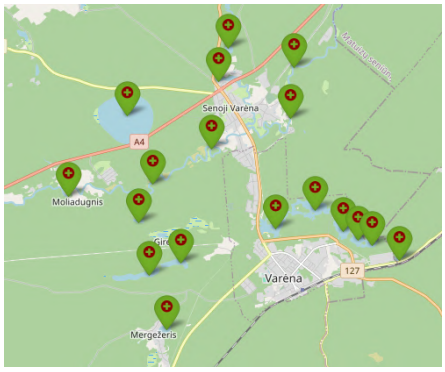


(a) Input Prompts of Forests

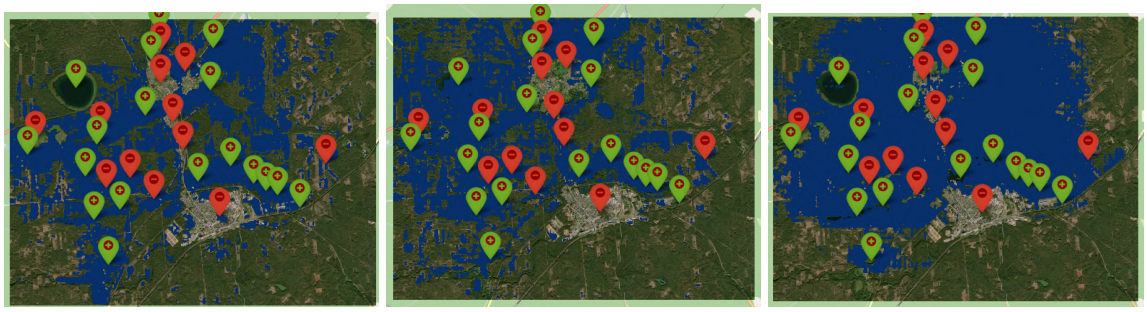
(b) ViT-B (c) ViT-L (d) ViT-H

Figure 15. Kėdainiai input prompt results

Consequently, Varėna region was the second one used for input prompts segmentation. The target objects were water entities, which include Glėbas hidrographical reserve, river Merkys and smaller ponds throughout the municipality. They are indicated as green foreground markers in figure's 16 **a)**. After the segmentation has been initialized, identical red background markers in figure's 16, **b)**, **c)**, **d)** were added so that the inputs would be identical to each other. When comparing the results, it can be seen that in ViT-L model checkpoint Glėbas hidrographical reserve was the only one that was annotated as a water entity, and in the others it was not annotated as a water body (figure 16, **c)**). By visual comparison, ViT-L model checkpoint seemed to have the least amount of noise and was more accurate targeting water entities in this region.



(a) Input Prompts of Water entities

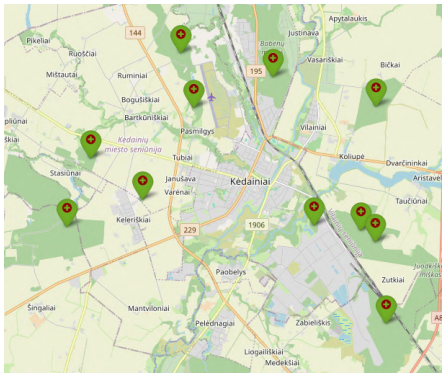


(b) ViT-B (c) ViT-L (d) ViT-H

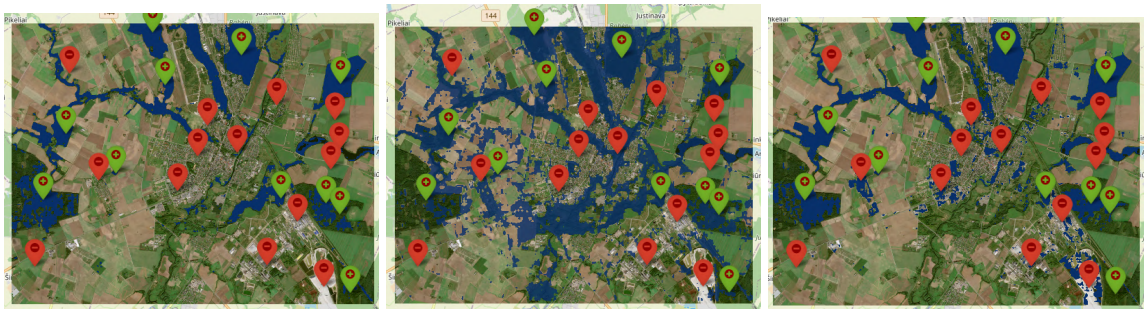
Figure 16. Varėna input prompt results

3.4.1.2. SAM-HQ

Previously mentioned input coordinates were identically used for the successor model SAM-HQ and its checkpoints to test out the SamGeo tool capabilities utilizing those model checkpoints. And in identical manner - Kėdainiai municipality was the first to be analyzed and compared between the predecessor models in this resulting section 3.4.2.2. These model checkpoints were different as it can be seen in figure 10 due to different training, whilst being the same model size. The probable results, just by definition of the model of being High Quality and being SAM successor, are expected to be better than the previously analyzed SAM model with its ViT checkpoints.



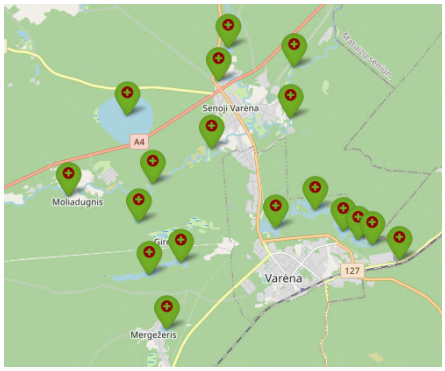
(a) Input Prompts of Forests



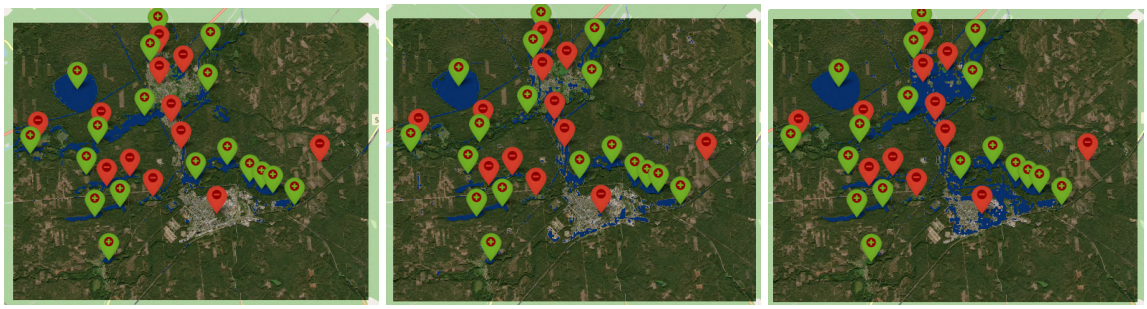
(b) ViT-B (c) ViT-L (d) ViT-H

Figure 17. Kėdainiai SAM-HQ input prompt results

And consequently the identical segmentation and annotation was performed on Varėna region.



(a) Input Prompts of Water Entities



(b) ViT-B (c) ViT-L (d) ViT-H

Figure 18. Varėna SAM-HQ input prompt results

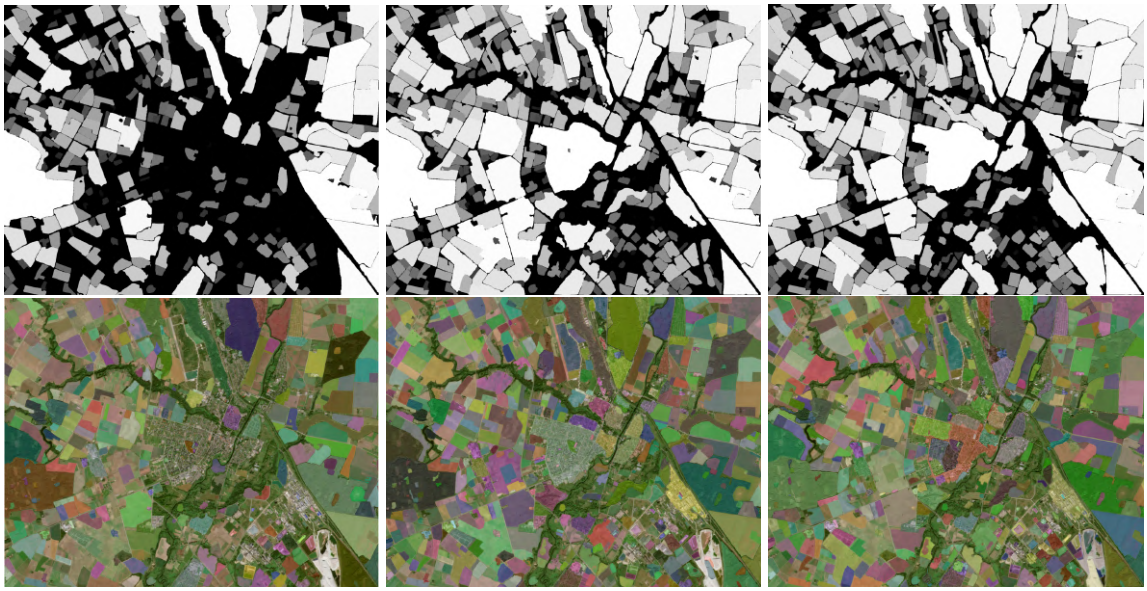
3.4.2. Automatic mask generation and annotation

After input prompt experimentations, automatic mask generation and annotation was used after using the input prompt segmentation model. The main objective is to semantically segment processed regions data so it would automatically annotate as many unique objects as accurately as possible, then check how it visually compares to the input prompt segmentation and calculate resulting scores for results.

In this section, all four regions data will be used for analysis and similarly to input prompting sequence - will be starting from SAM and finishing with SAM-HQ.

3.4.2.1. SAM

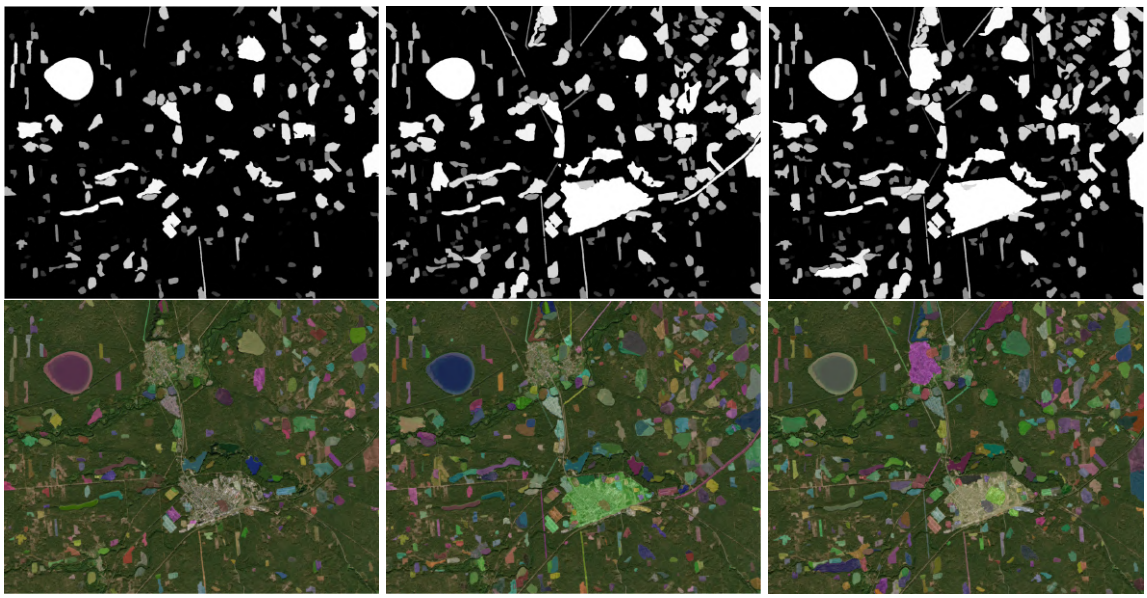
Similarly to input prompt analysis, Kėdainiai municipality region was the first one to be segmented and annotated through the generator, and as per the generated results in figure 19, it can be visually compared that the largest model checkpoint Vit-H, was the most accurate identifying unique segments in the map, which is expected due to it being the largest model checkpoint.



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 19. Kėdainiai SAM generation results

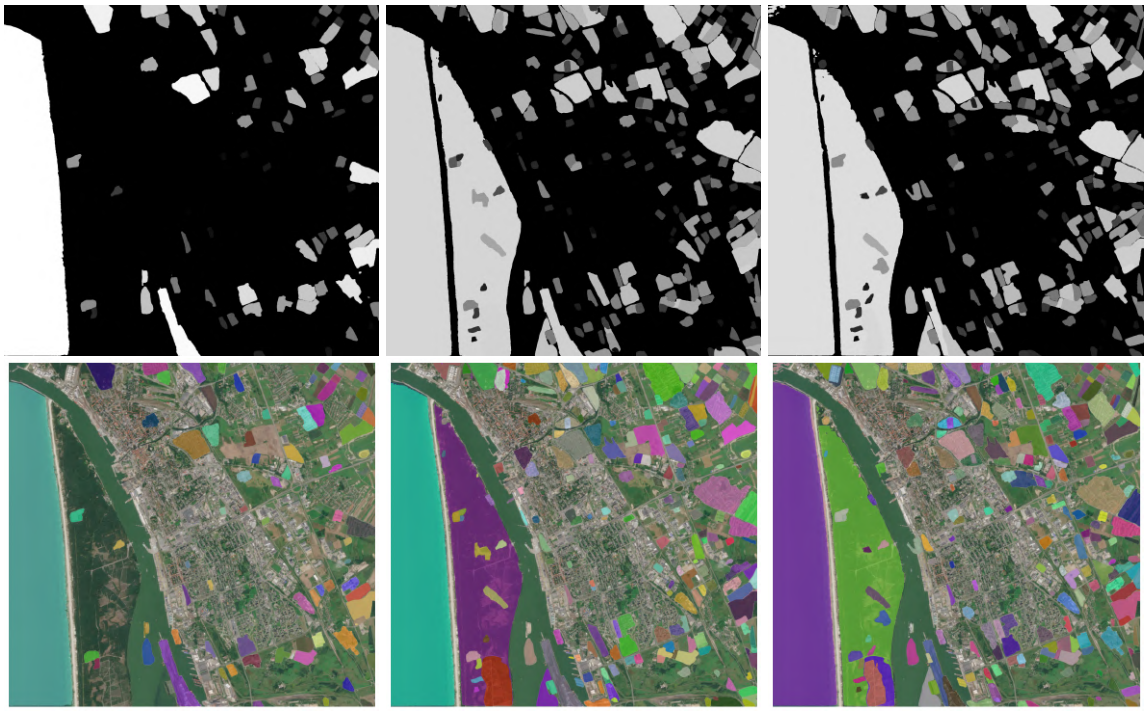
The second region was Varėna municipality, and it did worse in comparison to Kėdainiai region, as seen in the generated results in figure 20. Visually, the largest model checkpoint Vit-H seemed to have performed the best, but only slightly in comparison to ViT-L.



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 20. Varėna SAM generation results

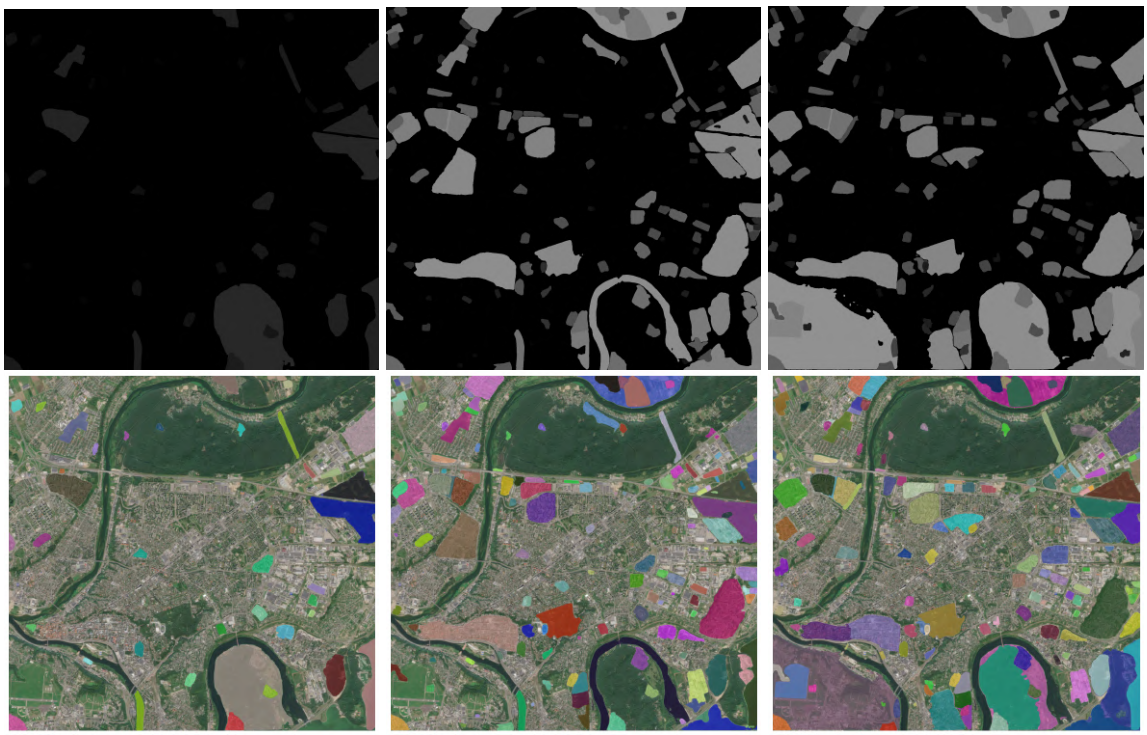
The third region was Klaipėda city region and as seen in the generated results of figure 21, we can see that ViT-H managed to annotate slightly more unique objects than ViT-L; ViT-B did not annotate Smiltynė region.



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 21. Klaipėda SAM generation results

The last region was Kaunas city, and from the generated results in figure 22, it can be seen, that SAM was not as confident as previous regions and performed the worst. It can also be noted, that neither of the three managed to annotate Kleboniškis forest park.

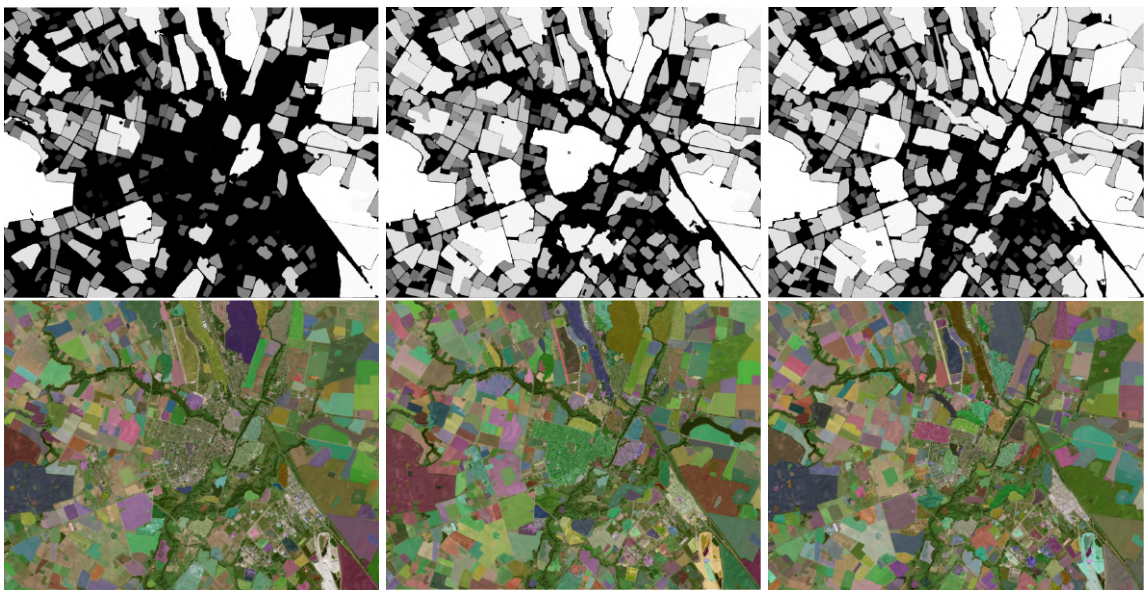


(a) ViT-B (b) ViT-L (c) ViT-H

Figure 22. Kaunas SAM generation results

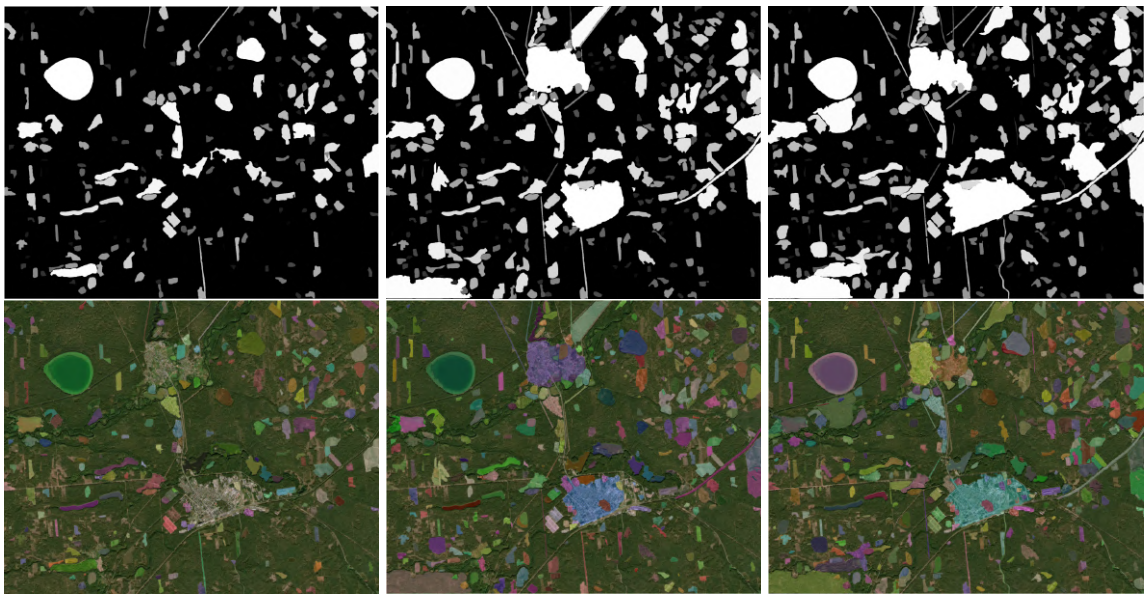
3.4.2.2. SAM-HQ

After SAM, identical analysis was performed on SAM-HQ and its model checkpoints. Identical parameters andarguments (in section 3.4) were used for SAM-HQ as they were for the predecessor model so that the results could be compared in an equal manner for the four Lithuanian regions.



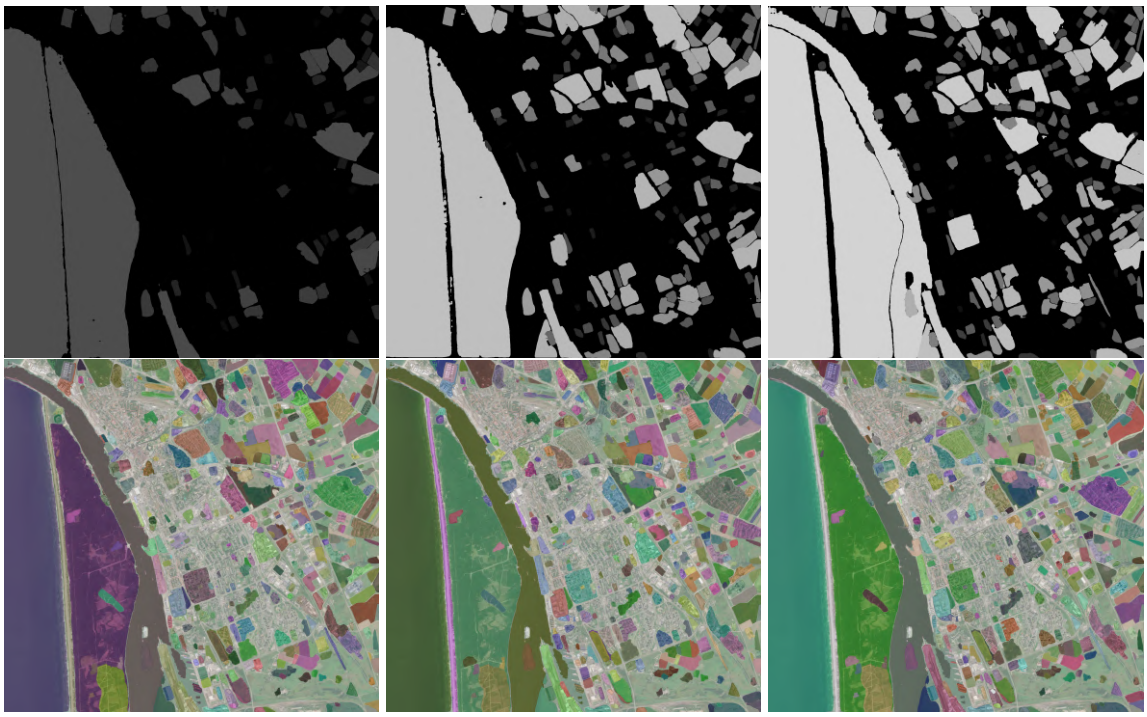
(a) ViT-B (b) ViT-L (c) ViT-H

Figure 23. Kėdainiai SAM-HQ generation results



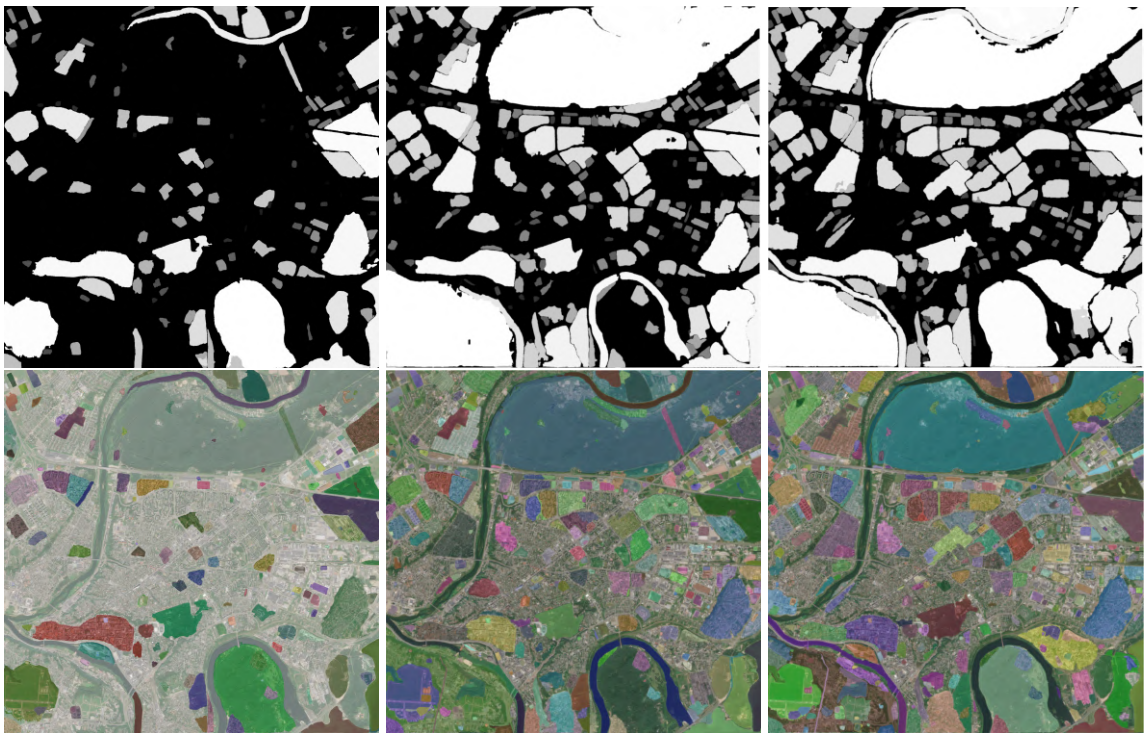
(a) ViT-B (b) ViT-L (c) ViT-H

Figure 24. Varėna SAM-HQ generation results



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 25. Klaipėda SAM-HQ generation results



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 26. Kaunas SAM-HQ generation results

Results

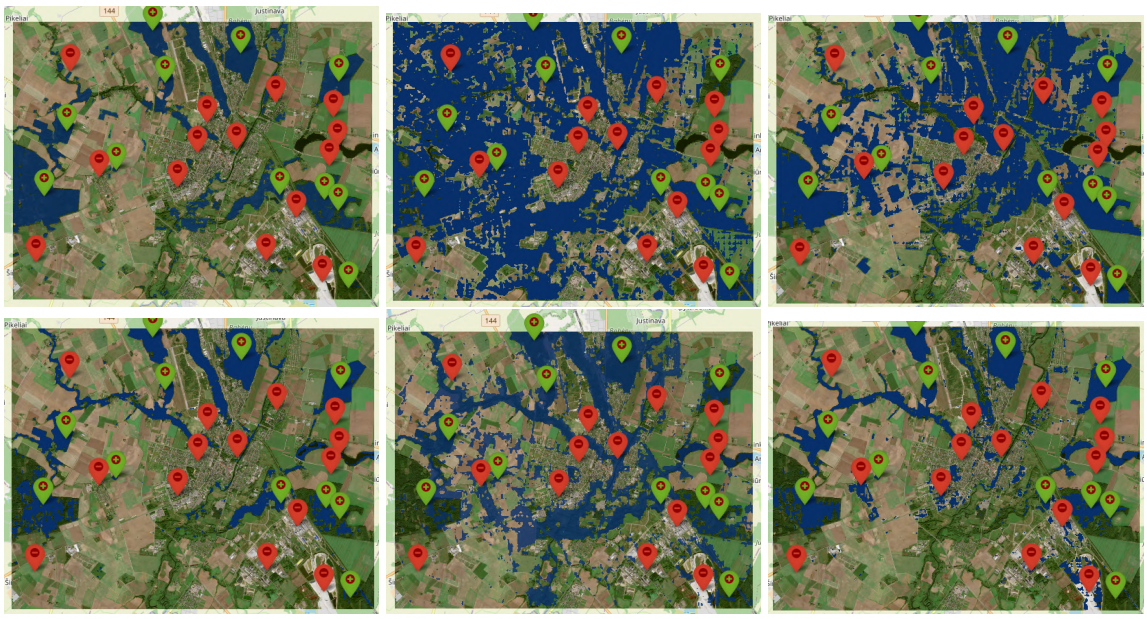
In this section, results from SAM and SAM-HQ models are compared in detail using input prompt segmentation and automatic mask segmentation and annotation. Input prompt segmentation did not provide the masks as it was segmented interactively, so these results can only be compared visually.

Furthermore, Intersection over Union (IoU) scores are calculated as the main metric and compared. As detailed by SAM authors and observed in its implementation [KMR+23] - it includes a mask decoder that, in addition to predicting the mask itself, also outputs a predicted IoU score. This score does not represent a direct calculation of IoU against a ground truth mask (which would require a prior knowledge of the true segmentation), but it serves as an internal estimation of the quality of the generated mask. This predicted IoU score acts as a valuable proxy for mask confidence and is a standard output of the SAM architecture, allowing for relative comparison between different segmentations generated by the model itself across various checkpoints and datasets. These resulting scores can be seen in table 1.

Input prompts results

As mentioned, we can only compare input prompt results visually. For this task Kėdainiai and Varėna regions were selected for their diversity. By checking the results of SAM and SAM-HQ models for these two Lithuanian municipalities (figure 27 and figure 28), it can be seen that for the left most and right most foreground input markers, SAM ViT-B outperformed the SAM-HQ ViT-B by only a slight margin, but it was unexpected that both ViT-L and ViT-H models were outperformed by the smallest model SAM ViT-B.

Moreover, it seems that the topographical landscape analysis and segmentation that has lots of variety with fields, cities, roads, rivers and patches of forests does not perform well when utilizing interactive input prompt segmentation, at least for the Kėdainiai municipality's processed data.

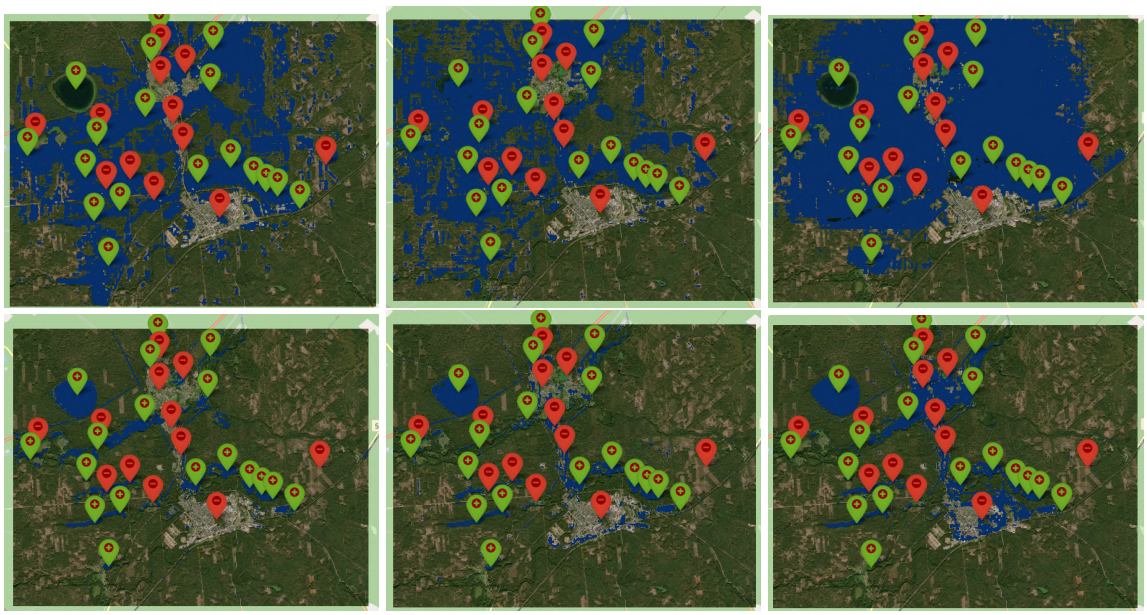


(a) ViT-B (b) ViT-L (c) ViT-H

Figure 27. Kėdainiai SAM and SAM-HQ text prompts result comparison

The results for Varėna region were completely different. It is lush with forests and after targeting water entities in the region, it can be visually seen that SAM model with all included checkpoints were outperformed by SAM-HQ by a long shot, and it can also be seen that only in SAM ViT-L checkpoint Glėbas hidrographical reserve was annotated, but resulted with quite large area of annotation noise throughout the satellite image. When comparing the best results it seems that the smallest SAM-HQ checkpoint ViT-B outperformed its much bigger datasets by a slight margin - it is evident by annotation of Varėna and Senoji Varėna cities, since there are no water entities in the annotated city regions in ViT-L and ViT-H checkpoints results.

Using input prompt segmentation with SAM-HQ is better for regions that have big identical topography objects, in this example - mainly a forest region.



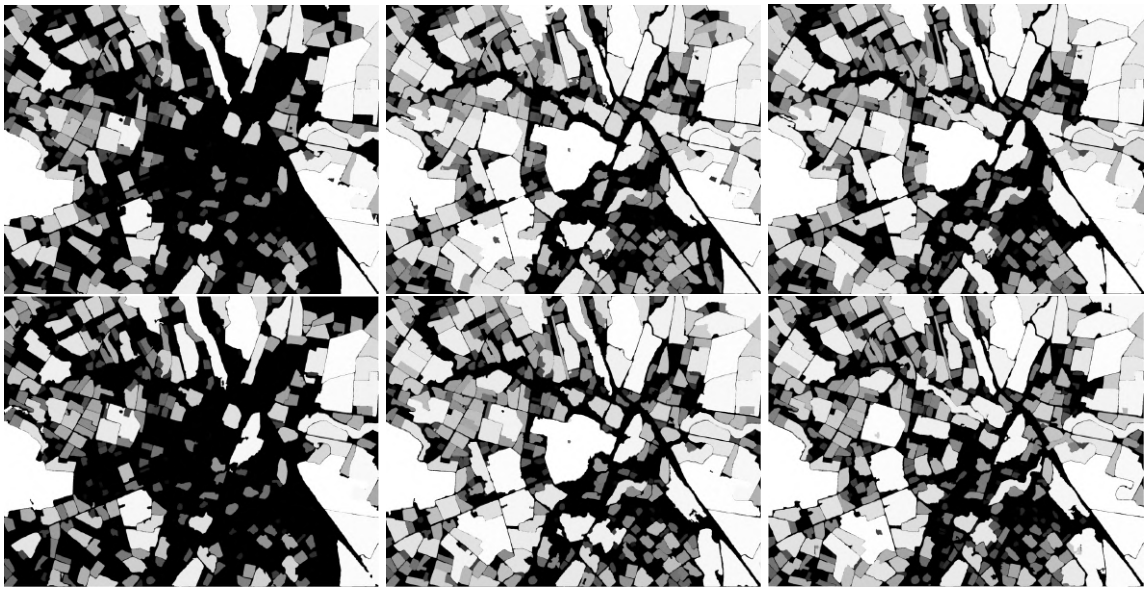
(a) ViT-B (b) ViT-L (c) ViT-H

Figure 28. Varėna SAM and SAM-HQ text prompts result comparison

Automatic mask segmentation results

In the main analytical part of for all four Lithuanian regions using automated mask segmentation, the resulting metrics in table 1 for the models and their corresponding model checkpoints, it is evident that in Kėdainiai region SAM-HQ model checkpoints outperformed their SAM counterparts, with an indication that all three SAM-HQ model checkpoints more confident in predicting unique objects in the topography in comparison to its predecessor.

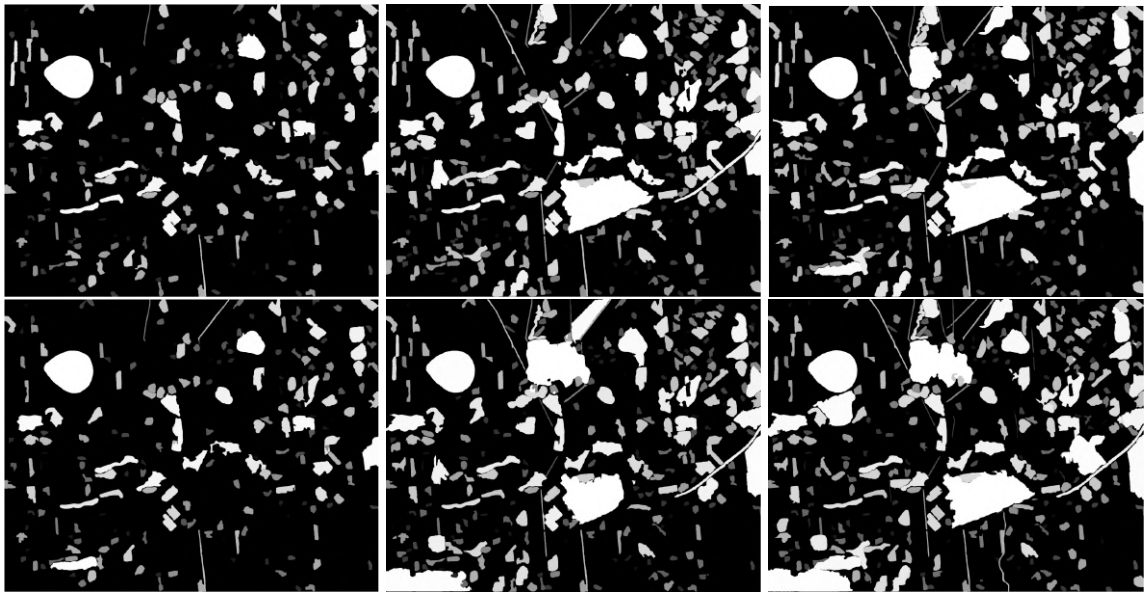
Additionally, by unique object count and IoU metrics Kėdainiai region performed the best in comparison to other regions, suggesting that the greater spectral diversity in Kėdainiai region landscape might facilitate more confident object identification by the models. Also, ViT-B for both models did not manage to segment Kėdainiai city, only small objects around it.



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 29. Kėdainiai SAM and SAM-HQ generated masks comparison

When it came to landscape satellite data, which has the same color scheme, in this case Varėna municipality region, the only underperforming SAM-HQ model checkpoint, when comparing to its predecessor was ViT-L. It can be seen that the region was overly lush in similar green vegetation for both models and their checkpoints to perform well in comparison to Kėdainiai region, which had much more diverse colors in the satellite data. Same as for Kėdainiai segmentation - both Senoji Varėna and Varėna cities were not segmented with in both ViT-B experiments, when comparing their larger checkpoints.

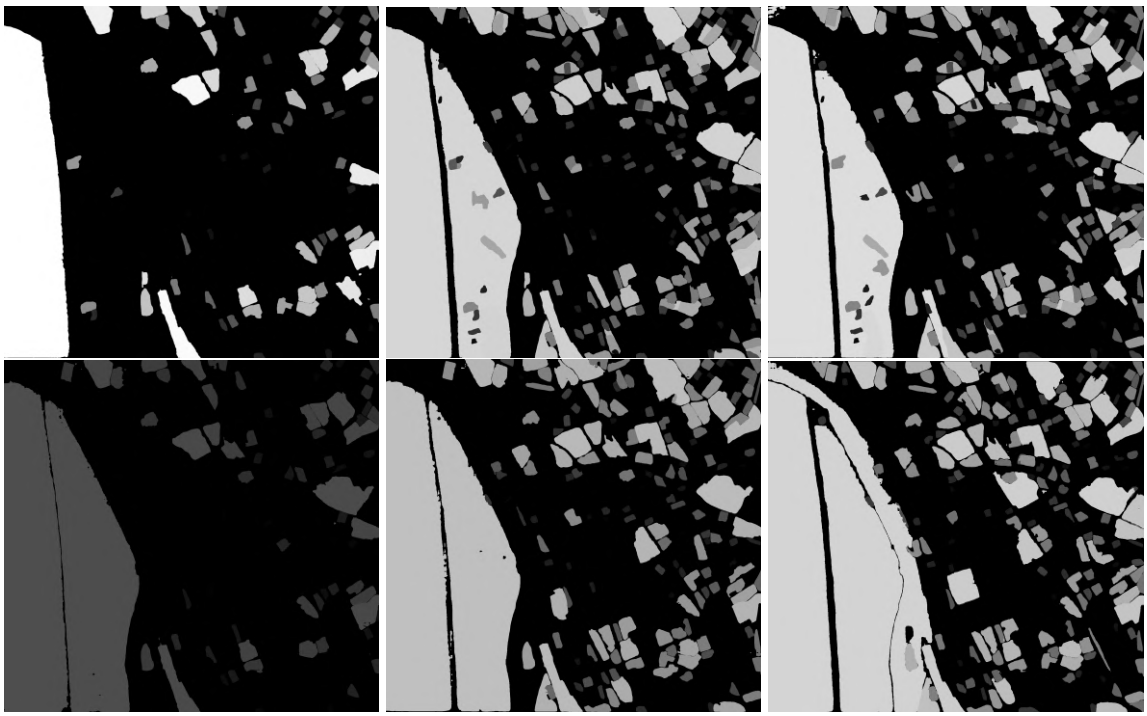


(a) ViT-B (b) ViT-L (c) ViT-H

Figure 30. Varėna SAM and SAM-HQ generated masks comparison

Moving on to more urbanized regions, the third was Klaipėda region. The results were quite different from previously discussed regions. Comparing visually and from the resulting metrics

in table 1, it can be seen that SAM-HQ ViT-H performed the best, but only slightly better than ViT-L. When comparing both models results, SAM seemed to identify more small objects inside the Smiltynė region in comparison with SAM-HQ, which was more accurate on mainland objects. Moreover, the only model and checkpoint which managed to annotate Curonian Lagoon water body was SAM-HQ ViT-H, which was a desired result even for smaller checkpoints, but only this one managed to annotate it. Lastly, as per IoU metrics, it seems that SAM ViT-L and ViT-H accuracy was almost identical to SAM-HQ ViT-L.

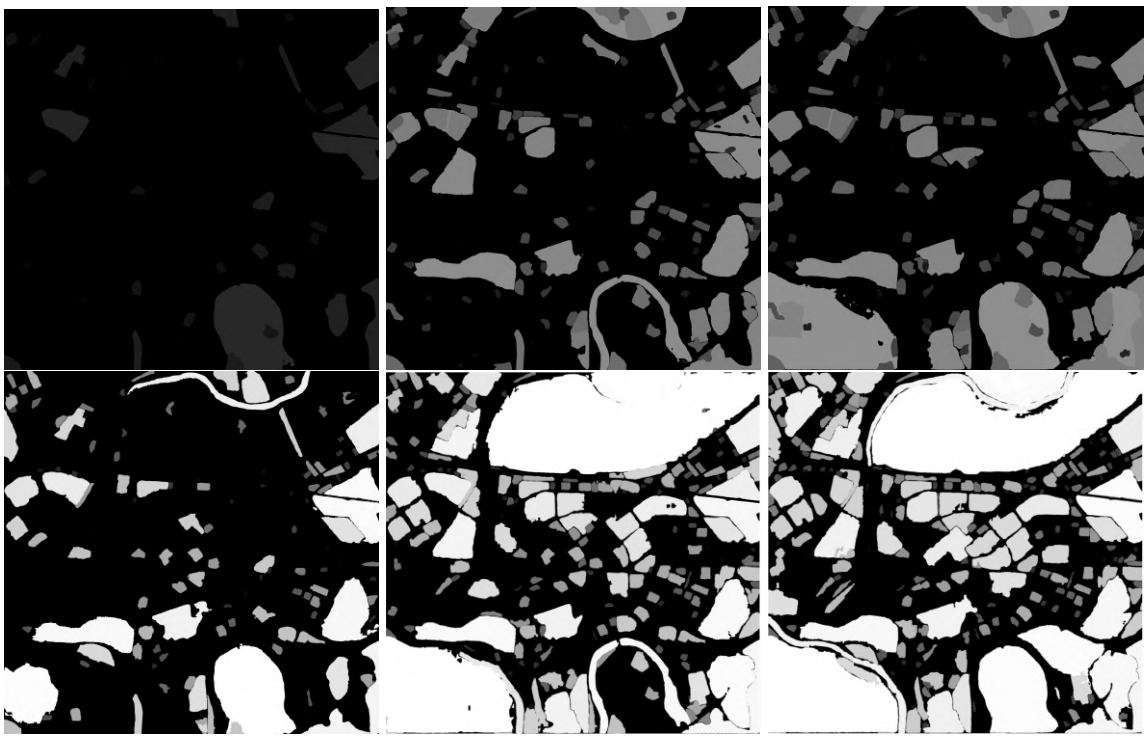


(a) ViT-B (b) ViT-L (c) ViT-H

Figure 31. Klaipėda SAM and SAM-HQ generated masks comparison

Finally, the last and most densely urbanized out the four is Kaunas region. From the resulting metrics in table 1, as expected SAM underperformed when comparing with SAM-HQ. It can be noted that the predecessors smallest model had the worst metrics out of all the analysis, in the generated mask it can be seen that it was not confident segmenting when using SAM ViT-B.

By visual hue and count of unique objects in the masks, SAM was not as confident as its successor and did not identify as many unique objects (especially in the heart of the city) as its successor SAM-HQ. It is also worth noting, that only SAM-HQ ViT-L and ViT-H managed to annotate Kleboniškis forest park, and ViT-L for both SAM and SAM-HQ segmented Nemunas river which surrounds Panemunė forest park, whilst others segmented only the park.



(a) ViT-B (b) ViT-L (c) ViT-H

Figure 32. Kaunas SAM and SAM-HQ generated masks comparison

Table 1. Predicted Intersection over Union Scores between SAM and SAM-HQ models

Predicted IoU	Kėdainiai	Varėna	Klaipėda	Kaunas	**Average**
SAM ViT-B	0.3096	0.1469	0.2284	0.1114	0.1991
SAM ViT-L	0.6441	0.2069	0.4438	0.2360	0.3827
SAM ViT-H	0.6514	0.2148	0.4522	0.3412	0.4149
SAM-HQ ViT-B	0.3593	0.1422	0.3591	0.2494	0.2775
SAM-HQ ViT-L	0.6962	0.1984	0.4565	0.5614	0.4781
SAM-HQ ViT-H	**0.7092**	0.2404	0.5236	0.6295	0.5257

Recommendations

Based on the findings of this thesis, the following recommendations are proposed for future work and practical applications:

- Gathering a higher quality dataset should be the top priority, with a possibility of looking into emerging image upscaling technologies and methods;
- While cloud removal was a pre-processing step, future work could investigate the direct robustness of segmentation models to varying levels of cloud cover or haze, potentially reducing reliance on perfect cloud-free imagery;
- Conduct a more exhaustive exploration of SamGeo tool parameters to potentially enhance segmentation accuracy for specific landscape types within Lithuania;
- Investigate the impact of seasonal changes on segmentation performance by analyzing satellite imagery from different times of the year for the selected regions, as landscape features and spectral signatures can vary significantly;
- Further explore techniques or model adjustments to improve segmentation performance in regions with extensive, homogeneous cover, such as the densely forested areas observed in the Varėna region;
- Explore the practical integration of the generated semantic masks into specific geospatial analysis workflows, such as land cover change detection, urban expansion monitoring, or agricultural field delineation, to assess their real-world utility.

Conclusion

- Model performance is highly dependent on the specific landscape characteristics. Agriculturally diverse regions (like Kėdainiai) showed higher segmentation confidence compared to densely forested, homogeneous regions (like Varėna), where distinguishing features proved more challenging. This underscores that a one-size-fits-all approach is insufficient for diverse topographies;
- SAM-HQ generally surpasses SAM in segmentation quality, particularly in complex urban environments like Kaunas or a region with spectral diversity like Kėdainiai. This suggests these types of regions offer tangible benefits for Lithuanian landscape geospatial applications;
- Larger ViT backbones, specifically the largest one ViT-H, yields the highest predicted Intersection over Union scores when used with both SAM and SAM-HQ. This confirms the trend that increased model capacity can lead to improved performance, albeit having higher computational costs;
- Input prompt segmentation offers insights for interactive tasks, but automatic mask generation is more reliable and suitable for automated segmentation and annotation;
- The variability in performance across different regions and the challenges with homogeneous landscapes highlight the need for further refinement, potentially through higher resolution data or newer model adaptations.

References

[AA19] Anju Asokan and J Anitha. Machine learning based image processing techniques for satellite image analysis-a survey. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pp. 119–124. IEEE, 2019.

[Abd20] Abdulhakim Mohamed Abdi. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using sentinel-2 data. *GIScience & Remote Sensing*, 57(1):1–20, 2020.

[BLY+21] Qiang Bai, Shaobo Li, Jing Yang, Mingming Shen, Sanlong Jiang, and Xingxing Zhang. Robot three-finger grasping strategy based on deeplabv3+. *Actuators*, 10(12), 2021. ISSN: 2076-0825. DOI: 10.3390/act10120328. URL: https://www.mdpi.com/2076-0825/10/12/328.

[CPS+17] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[CZL+22] Shanjing Chen, Wenjuan Zhang, Zhen Li, Yuxi Wang, and Bing Zhang. Cloud removal with sar-optical data fusion and graph-based feature aggregation network. *Remote Sensing*, 14(14):3374, 2022.

[CZP+18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[EGS+23] Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. Uncrtaints: uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2085–2095, 2023.

[EXS+22] Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. Sen12ms-cr-ts: a remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[GLG+18] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018.

[GS22] Dalia Grendaitė and Edvinas Stonevičius. Machine learning algorithms for biophysical classification of lithuanian lakes based on remote sensing data. *Water*, 14(11):1732, 2022.

[KMC+22] Shreeyam Kacker, Alex Meredith, Kerri Cahoy, and Georges Labreche. Machine learning image processing algorithms onboard ops-sat, 2022.

[KMR+23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. *arXiv:2304.02643*, 2023.

[KOB21] Yao Kai, Alberto Ortiz, and Francisco Bonnin-Pascual. A weakly-supervised semantic segmentation approach based on the centroid loss: application to quality control and inspection. *IEEE Access*, 9:69010–69026, 2021-01. DOI: 10.1109/ACCESS.2021.3077847.

[KYD+23] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015-06.

[MEZ+20] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020.

[MFH+17] Suguru Miyagawa, Hisashi Fukuyama, Masakazu Hirota, Tatsuo Yamaguchi, Kazuo Kitamura, Takao Endo, Hiroyuki Kanda, Takeshi Morimoto, and Takashi Fujikado. Automated measurements of human cone photoreceptor density in healthy and degenerative retina by region-based segmentation. *Clinical Ophthalmology*, 11:781–790, 2017-04. DOI: 10.2147/OPTH.S133070.

[NBB+20] Edoardo Nemni, Joseph Bullock, Samir Belabbes, and Lars Bromley. Fully convolutional neural network for rapid flood segmentation in synthetic aperture radar imagery. *Remote Sensing*, 12(16):2532, 2020.

[Ric07] Mark A Richards. A beginner's guide to interferometric sar concepts and signal processing [aess tutorial iv]. *IEEE Aerospace and Electronic Systems Magazine*, 22(9):5–29, 2007.

[SBA+20] Joel Segarra, Maria Luisa Buchaillot, Jose Luis Araus, and Shawn C Kefauver. Remote sensing for precision agriculture: sentinel-2 improved features and applications. *Agronomy*, 10(5):641, 2020.

[SPE+20] Michael Schmitt, Jonathan Prexl, Patrick Ebel, Lukas Liebel, and Xiao Xiang Zhu. Weakly supervised semantic segmentation of satellite images for land cover mapping–challenges and opportunities. *arXiv preprint arXiv:2002.08254*, 2020.

[WO23] Qiusheng Wu and Lucas Prado Osco. Samgeo: a python package for segmenting geospatial data with the segment anything model (sam). *Journal of Open Source Software*, 8(89):5663, 2023.

[YCS+22] Zhengbo Yu, Zhe Chen, Zhongchang Sun, Huadong Guo, Bo Leng, Ziqiong He, Jinpei Yang, and Shuwen Xing. Segdetector: a deep learning model for detecting small and overlapping damaged buildings in satellite images. *Remote Sensing*, 14(23):6136, 2022.

Abbreviations

- AI – Artificial Intelligence
- ASPP – Atrous Spatial Pyramid Pooling
- CNN – Convolutional Neural Network
- ESA – European Space Agency
- FCN – Fully Convolutional Network
- GPU – Graphics Processing Unit
- IoU – Intersection over Union
- InSAR – Interferometric Synthetic Aperture Radar
- ResNet – Residual Network
- RGB – Red-Green-Blue
- SAM – Segment Anything Model
- SAM-HQ – Segment Anything Model in High Quality
- SAR – Synthetic Aperture Radar
- ViT – Vision Transformer
- ViT-B – Vision Transformer Base
- ViT-L – Vision Transformer Large
- ViT-H – Vision Transformer Huge

Appendix 1
AI tools usage

In the preparation of this thesis, AI powered tools were used to assist with aspects of grammar refinement, rephrasing suggestions and reading material comprehension. However, all core research, experimental design, data analysis, interpretation of results, and the final written content are the original work of the author. The tools that were used:

- OpenAI - ChatGPT. https://chat.openai.com/chat;
- Grammarly Inc. - Grammarly. https://www.grammarly.com.