NTNU

Kunnskap for en bedre verden

Department of Geomatics

Deep learning-based building detection from oblique aerial images

Author:

Sigmund Sekkesæter Mestad

August, 2023

Abstract

In recent years the advancements in deep learning and the availability and quality of high-resolution aerial images have pushed the limits of what is possible when it comes to the accuracy and reliability of automatic building detection. This thesis proposes a novel approach to building detection using oblique aerial images, which offer a more comprehensive view of the environments compared to nadir (top-down) images.

The proposed method involves the automatic generation of a training dataset and the training of a deep learning model. This model is then used to detect buildings on images from multiple perspectives before the predictions are combined to offer more accurate and reliable building footprints.

An experimental study evaluates the model's ability to detect buildings and examines its performance on oblique aerial images compared to nadir aerial images. Moreover, an area analysis is conducted to assess the accuracy of the proposed method to combine the predictions from multiple images.

The results demonstrate that oblique aerial images provide more features, enabling better recognition of building structures and improving building detection. The process of combining the segmentation results from several images also proved to be accurate. However, occlusion remains a significant challenge in oblique aerial imagery and improvements are needed to better deal with scenarios where the building structures are not visible in some of the images. Furthermore, it emphasizes the importance of high-quality training data and the need for more diverse datasets to enhance the model's robustness and generalizability.

Sammendrag

I de senere årene har fremskritt innen dyp læring og tilgjengeligheten og kvaliteten på høyoppløselige flyfoto presset grensene for hva som er mulig når det gjelder nøyaktighet og pålitelighet i automatisk bygningsdeteksjon. Denne avhandlingen foreslår en ny tilnærming for å detektere bygninger ved hjelp av skrå luftbilder, som gir en mer omfattende framstilling av omgivelsene sammenlignet med nadir (vertikale) bilder.

Den foreslåtte metoden innebærer automatisk generering av et treningsdatasett og trening av en dyp læringsmodell. Denne modellen brukes deretter til å oppdage bygninger på bilder fra flere perspektiver før prediksjonene kombineres for å oppnå mer nøyaktige og pålitelige bygningsomriss.

En eksperimentell studie evaluerer modellens evne til å oppdage bygninger og vurderer hvordan den presterer på skrå flyfoto sammenlignet med vertikale bilder. Videre utføres en områdeanalyse for å vurderehvor nøyaktig den foreslåtte metoden klarer å kombinere prediksjoner fra flere bilder.

Resultatene viser at skrå flyfoto inneholder flere kjennetegn og trekk som gjør det lettere å gjenkjenne bygningsstrukturer og forbedre deteksjonen av bygninger. Prosessen med å kombinere segmenteringsresultatene fra flere bilder viste seg også å være nøyaktig. Imidlertid er okklusjon en betydelig utfordring på skrå flyfoto, og det er behov for forbedringer for å bedre håndtere tilfeller der bygningsstrukturene ikke er synlige i enkelte bilder. Det legges også vekt på betydningen av høykvalitets treningsdata og behovet for mer varierte datasett for å styrke modellens robusthet og generaliserbarhet.

Preface

I am pleased to present this master's thesis from the Department of Civil and Environmental Engineering at the Norwegian University of Science and Technology in Trondheim, Norway. This thesis is the culmination of my study program in Engineering and ICT with a specialization in Geomatics.

This thesis is the result of a collaboration with Norkart and Kartverket and serves as a contribution to the KartAI research project, which aims to enhance the quality of the Norwegian real estate register using Artificial Intelligence (AI). I am grateful to Norkart for providing 3D building models, to Kartverket for providing oblique aerial images and all those who have supported and shown interest in my work.

I want to express my gratitude to my supervisors at Norkart, Mathilde Ørstadvik and Rune Aasgaard, for their invaluable ideas, insights, and enthusiasm for my project. I would also like to thank my supervisor at NTNU, Hongchao Fan, for the knowledge and guidance provided during my studies and for the assistance and encouragement with this project.

To my amazing wife, you have been my rock and my constant source of support throughout this journey. Your love and encouragement have kept me going, and I cannot thank you enough for believing in me and my abilities. You bring so much love and joy into my life, and I am truly grateful to have you by my side.

Sigmund Sekkesæter Mestad
Trondheim, Wednesday 9th August, 2023

Sigmund S. Mestad

Contents

List of Figures

List of Tables

1 Introduction

This content presented in this thesis is written as a contribution to the research project KartAI. Their main goal is to make the the process of sending and processing building permit applications more effective. This will be achieved by using artificial intelligence to improve the quality of the norwegian cadastre and map databases. A central part of this project is to use machine learning to detect and delineate buildings in aerial images and compare the findings with existing building data. While the KartAI project has achieved promising results so far, there is still room for improvement. Until now only nadir photos has been used and this thesis will explore the possibility of utilizing oblique aerial imagery to improve the building detection.

Building detection and building extraction are two closely related terms in the field of computer vision and remote sensing. Building detection refers to the process of identifying and locating buildings within an image or a set of images. Building extraction goes a step further and involves outlining or delineating the exact boundaries of the detected buildings, resulting in precise building footprints. In this thesis these terms will be used interchangeably as the process involves extracting building footprints, but the ultimate goal is to detect buildings or rather buildings that deviate from the existing building data.

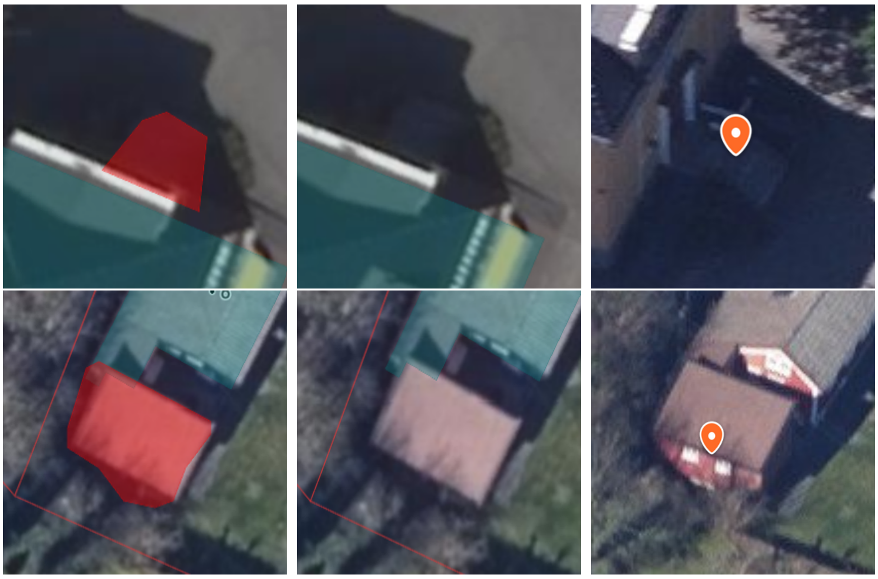
1.1 Motivation

Achieving the best possible accuracy and reliability of the building detection is of utmost importance for this project. While the AI models for detecting buildings keeps getting better, there is still need for manual work to verify the detections and update the databases. Reducing the number of incorrect detections (false positives), would be of great value because it would result in less manual labour. On the other hand, if there are too many undetected buildings (false negatives), the errors in the databases would not be detected, and the project would not fulfill its purpose.

Looking at some examples of wrong suggestions, it becomes evident that in some cases it can be difficult even for the human eye to recognize buildings on a nadir photo. In these cases the real problem is that the nadir images lacks the semantic features required to recognize a building. This thesis will try to deal with this problem by using oblique aerial images. There are two ideas for why this potentially could improve the building detection:

- The oblique images can reveal more features, like vertical extent and facade details, which may give the machine learning model a better basis to detect buildings and to avoid false predictions on object that is similar to buildings from a birds-eye perspective.
- Combining the predictions from several perspectives gives a redundancy, which means that even if the machine learning model fails to recognize a building from one perspective, the final result can still be correct if it is easier to detect from some other perspectives.

Figure 1 gives an example of two detected buildings from the KartAI project, clearly demonstrating the advantages of the oblique perspective. From the orthophotos it difficult to confirm or deny these detections, but looking at the oblique images it is obvious that the first detection is incorrect.



(a) Detections **(b)** Orthophoto **(c)** Oblique images

Figure 1: Examples where oblique photos are helpful to confirm or deny building suggestions

1.2 Challenges and research questions

Using oblique images does however bring some challenges, compared to using orthophotos. Orthophotos are photos that has been geometrically corrected and transformed into a planar representation so that the scale is uniform and it follows a given map projection. This is easy to work with because the real world coordinate for each pixel is implicit and it allows for precise distance, area and angle calculations. With oblique images it gets more complicated because the scale is not uniform and the real world coordinates also depends on the height and the vertical extent of the objects. The oblique perspective also presents a drawback in terms of increased occlusion caused by terrain, vegetation, or tall buildings. However, this limitation can be mitigated by the advantage of having multiple perspectives, which allows us to overcome some of the occlusion challenges.

The goal of this thesis is to do an experiment to put these ideas to a test and see if we can get more accurate and reliable building detections with the help of oblique images compared to only using nadir images. This will be done by using semantic segmentation to detect buildings on oblique aerial images from several perspectives and find a method to project and combine these predictions to extract the building footprints. The objective of the research is to answer to following questions:

- How does the segmentation model perform on oblique images compared to nadir images?
- To what extent can the transformation from image pixels to real-world coordinates provide accurate georeferenced building footprint of the detected buildings on oblique images?
- What are key challenges when it comes to building detection on oblique aerial images and how can these challenges be addressed to enhance performance?

1.3 Structure of thesis

This thesis is structured into several chapters to provide a comprehensive exploration of the proposed approach for building detection using oblique aerial images:

1. **Introduction:** Introduces the context and motivation behind the research, highlighting the importance of building detection in various applications and the limitations of traditional methods relying solely on nadir images. Moreover, it outlines the objective of the research and presents the research questions that will be addressed. Finally, the chapter presents the overall structure of the thesis.

2. **Background knowledge and related work:** Presents the background and related work on building detection from oblique aerial images. It covers the history and applications of oblique imagery, the collinearity equations for georeferencing, and reviews existing research on building extraction methods. The chapter lays the foundation for the subsequent experimental analysis.

3. **Method:** Outlines the proposed approach for building detection from nadir and oblique aerial images. The methodology consists of two main parts. In the first part, a new dataset is created which is then used to train semantic segmentation models for building detection on both nadir and oblique images. The second part presents the purposed method for analyzing an area by using the trained segmentation models on images from multiple perspectives and combining the results to extract the building footprints.

4. **Data:** Describes the data used in the experimental study, including oblique aerial images, 3D building models, and the digital surface model (DSM). Additionally, it presents the training dataset, created from these sources to train the segmentation models.

5. **Experimental study:** Describes the training of the segmentation model, including hyperparameter selection, data augmentation, and training time. It examines the model's ability to detect buildings on oblique compared to nadir aerial images. Finally, it presents the results of an area analysis, showcasing the effectiveness of combining predictions from multiple images for improved and reliable building detections.

6. **Conclusion and futher research:** Concludes the thesis by summarizing the key findings, discussing the contributions of the research, and offering suggestions for future improvements and avenues for further research in the field of building detection using oblique aerial images.

2 Background knowledge and related work

The following chapter provides an overview of historical background, the theoretical foundation and existing research relevant to the topic of building detection on oblique aerial images. It starts with a brief presentation of the history of oblique aerial images, highlights its distinctions from vertical imagery and explores its various applications. Next, the chapter delves into the process of georeferencing and determining the spatial extent of objects captured in the images. The collinearity equations gives mathematical framework for linking the image space to real-world geographic coordinates by considering the geometry of the camera system. Finally the chapter focuses on the task of building detection and segmentation from remote sensing images, outlining the significance and challenges associated with this task. By thoroughly reviewing existing research on the topic, the aim is to provide an understanding of the current state-of-the-art, laying the foundation for the subsequent experimental analysis presented in the following chapters.

2.1 Oblique aerial images

An oblique aerial image is a type of aerial photograph taken from an angle rather than directly from above (nadir view). In contrast to vertical or nadir imagery, which captures the earth's surface directly beneath the aircraft, oblique images are captured with the camera tilted, providing views of the landscape from an oblique perspective.

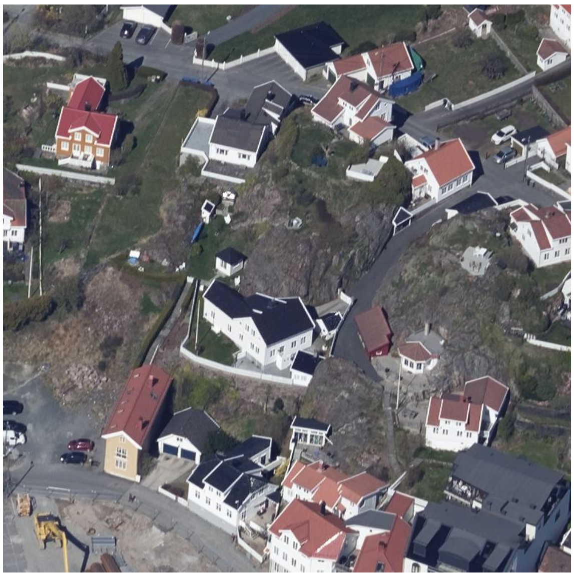


Figure 2: Example of an oblique aerial image.

Oblique images has been a known and used concept since the beginning of aerial imagery. The first recorded aerial photo in the US was an oblique shot from a balloon in 1860. These early oblique images were often used for reconnaissance and military purposes. In the last part of the 20th century, the oblique aerial imagery faced competition from other technologies such as vertical aerial photography and satellite imagery. They were often preferred for mapping and surveying due to its ability to produce accurate, planimetric maps.

In the last 20 years the use of oblique aerial images has significantly increased. This is a result of the progress made in photogrammetry and computer vision algorithms, the increasing market availability and the advantages they provide compared to vertical imagery. These advantages includes visibility of vertical structures, multiple views, higher redundancy and reliability, more detailed 3D information and a lot of hidden potential (Verykokou and Ioannidis, 2018a).

Oblique aerial imagery has been a topic of research for various applications, including but not limited to texture mapping, 3D building reconstruction, dense image matching, 3D scene classification and facade detection, partitioning and reconstruction (Verykokou and Ioannidis, 2018b).

There has also been some research on extraction of building footprints from oblique aerial images, which coincide with the objective of this thesis (Nex et al., 2013; Frommholz et al., 2016). The methods found in previous research all have in common that the buildings are not directly detected in the images. Instead a dense image matching algorithm is used to generate dense point clouds. The buildings are then detected based on geometric features like edges and the verticality of building facades. In contrast, the method proposed in this thesis uses deep learning to detect buildings based on the semantic features in the images.

2.2 Collinearity equation

The only way to really make use of a detected building in an aerial image, is if it can be expressed in real world coordinates. As mentioned in 1.2, this gets more complicated with oblique images than with orthophotos. However, assuming we have accurate information about the cameras interior and exterior parameters, this can be solved by using the collinearity equations.

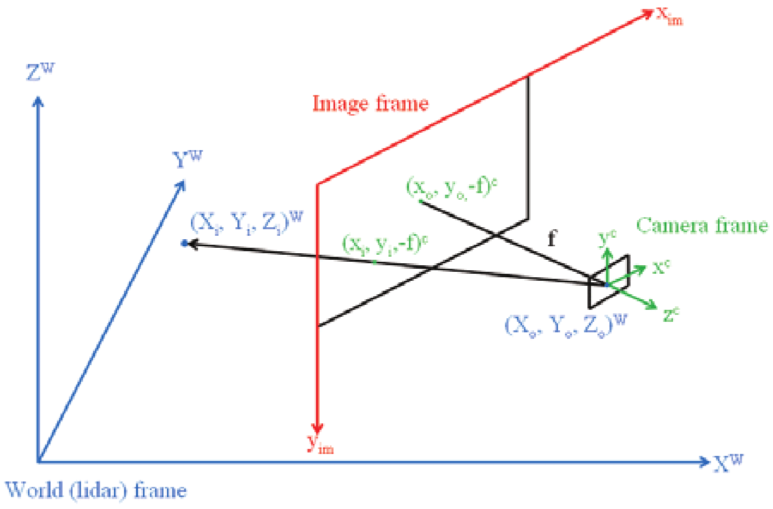


Figure 3: Camera geometry and parameters

Source: Li-Chee-Ming and Armenakis, 2012

The interior orientation refers to the focal length (f), the distance between the camera lens and the camera sensor, and the principle point (x_0, y_0) which is the location on the image sensor where the optical axis intersects. These values are constant and specific for each camera. The exterior parameters refers to the geographic location (X_0, Y_0, Z_0) and orientation ($yaw(\kappa)$, $pitch(\omega)$, $roll(\phi)$) of the camera perspective centre when each image is taken.

The collinearity equations are based on the fact that the perspective center, photo image point and the object point must lie on a straight line (**schenkIntroductionPhotogrammetry**). This concept results in a set of two equations that can express the image coordinates (x, y) as a function of 12 parameters: The object coordinates (X, Y, Z) and the nine camera parameters mentioned above.

$$x(X, Y, Z) = x_0 - f\frac{(X - X_0)r_{11} + (Y - Y_0)r_{12} + (Z - Z_0)r_{13}}{(X - X_0)r_{31} + (Y - Y_0)r_{32} + (Z - Z_0)r_{33}} \tag{1a}$$

$$y(X, Y, Z) = y_0 - f\frac{(X - X_0)r_{21} + (Y - Y_0)r_{22} + (Z - Z_0)r_{23}}{(X - X_0)r_{31} + (Y - Y_0)r_{32} + (Z - Z_0)r_{33}} \tag{1b}$$

where the $r_{i,j}$ represent the corresponding entry in the rotation matrix:

$$R = \begin{bmatrix} \cos\phi\cos\kappa & \cos\omega\sin\kappa + \sin\omega\sin\phi\cos\kappa & \sin\omega\sin\kappa - \cos\omega\sin\phi\cos\kappa \\ -\cos\phi\sin\kappa & \cos\omega\cos\kappa - \sin\omega\sin\phi\sin\kappa & \sin\omega\cos\kappa + \cos\omega\sin\phi\sin\kappa \\ \sin\phi & -\sin\omega\cos\phi & \cos\omega\cos\phi \end{bmatrix} \tag{2}$$

Equation 1a and 1b describes the transformation from world coordinates to image coordinates. In order to do the opposite, transform image coordinates to world coordinates, the equations can be solved for X and Y:

$$X(x, y, Z) = x_0 + (Z - Z_0)\frac{r_1 x + a_4 y - a_7 f}{a_3 x + a_6 y - a_9 f} \tag{3a}$$

$$Y(x, y, Z) = y_0 + (Z - Z_0)\frac{r_2 x + a_5 y - a_8 f}{a_3 x + a_6 y - a_9 f} \tag{3b}$$

Note that with only the image coordinates (x, y), the equations would only describe a straight line from the camera perspective centre on which the object must lie. Thus, to determine the coordinate the height value (Z) is also needed.

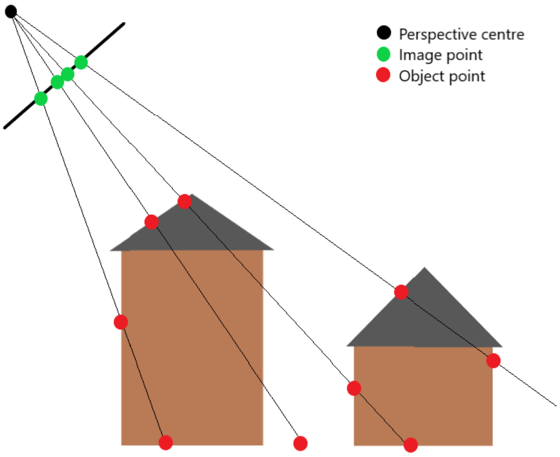


Figure 4: Illustration of the collinearity model.

2.3 Building extraction from remote sensing images

Building extraction from remote sensing images refers to the process of identifying building and non-building pixels in remote sensing images. This task plays a crucial role in various fields, such as urban planning, disaster management, and environmental monitoring. The ability to accurately identify buildings from aerial or satellite imagery provides valuable insights for decision-making processes and contributes to efficient resource allocation.

There are many reasons for why building extraction is a challenging task. Buildings can be complex, with a diversity of sizes, shapes and architectural details. Similar features can look very different in terms of colors or spectral properties when observed at different times and in different locations(H. Chen et al., 2022). Surrounding objects introduces potential sources of interference and occlusion and the contrast between the roof and the surrounding ground surfaces may be low, especially in urban ares with a lot of asphalt (Wei et al., 2004). Other factors that makes it more complicated is variation in scale, shadows and man-made features that can easily be confused with buildings, such as trucks, terraces, tents and containers.

Due to these challenges there has been much research and many methods proposed to solve the problem of building detection. Traditional methods for building detection, such as watershed (Serra, 1982), active contours (Kass et al., 1988) and Markov random field (Kato, 2012), all have the drawback that they heavily depends on handcrafted feature selection, which are difficult to optimize (Yang et al., 2018).

In recent years there have been an increased use of AI, specifically deep learning, in the field of remote sensing. The rapid development of better technologies and algorithms, combined with better sensors and more powerful computers has opened a new world of possibilities. Unlike in traditional machine learning methods, where the process of feature extraction is performed manually by the data scientist, deep learning methods can extract features automatically. Convolutional Neural Networks (CNN) uses convolution to extract features. The images are broken down into smaller, simpler features, represented by filters. These filters are optimized through automated learning and applied to different regions of the input to extract the relevant information. Long et al. (2015) adapted the CNN model to a Fully Convolutional Network (FCN) to enable the task of semantic segmentation, that is making a dense pixel-wise segmentation map of an image, where each pixel is categorized into a class or object. Newer methods has gradually achieved better segmentation results by modifying the structure of the FCN. For example, U-Net (Ronneberger et al., 2015) has played a crucial role by finding a way to capture both local and global contextual information. However, the increase in receptive field while retaining local information requires a lot of computational and memory overhead, making it unsuitable for processing remote sensing data, given its large volume and resolution (K. Chen et al., 2021).

To overcome these limitations, there has recently been a shift towards attention-based models with transformers. Transformers can learn long-range dependencies, making them suitable for pixel-wise segmentation in remote sensing images. H. Chen et al. (2022) demonstrates this by implementing an efficient transformer method for remote sensing image change detection and achieve state-of-the-art results, compared to other methods.

Despite these advancements, standard transformation models are still computationally expansive and memory-intensive. To increase efficiency, Carion et al. (2020) introduced DETR, a hybrid CNN-transformer, but the result was slow convergence and bad performance on small objects. Deformable DETR (Zhu et al., 2021) addresses this by focusing on key sampling points using deformable convolution, outperforming DETR in detecting small objects.

K. Chen et al. (2021) builds on this concept and presents a method that is optimized for building extraction by utilizing the fact that buildings usually only occupy a small part of aerial images. They call the method Sparse Token Transformers (STT), and the idea is to represent buildings as a set of "sparse" feature vectors and only consider the key vectors for self-attention. This allows the model to acquire a large perceptive field with contextual information while also greatly reducing the computational complexity.

3 Methodology

The project aims to detect and delineate buildings using a combination of nadir and oblique aerial images. The proposed approach involves two parts and relies on deep learning techniques. The first part is to create a new dataset which is then used to train semantic segmentation models to recognize buildings on nadir and oblique aerial images. The second part is to analyze an area by using the semantic segmentation model on images from several perspectives and combine the results to extract the footprint of the buildings. An overview of the complete process is illustrated in figure 5.

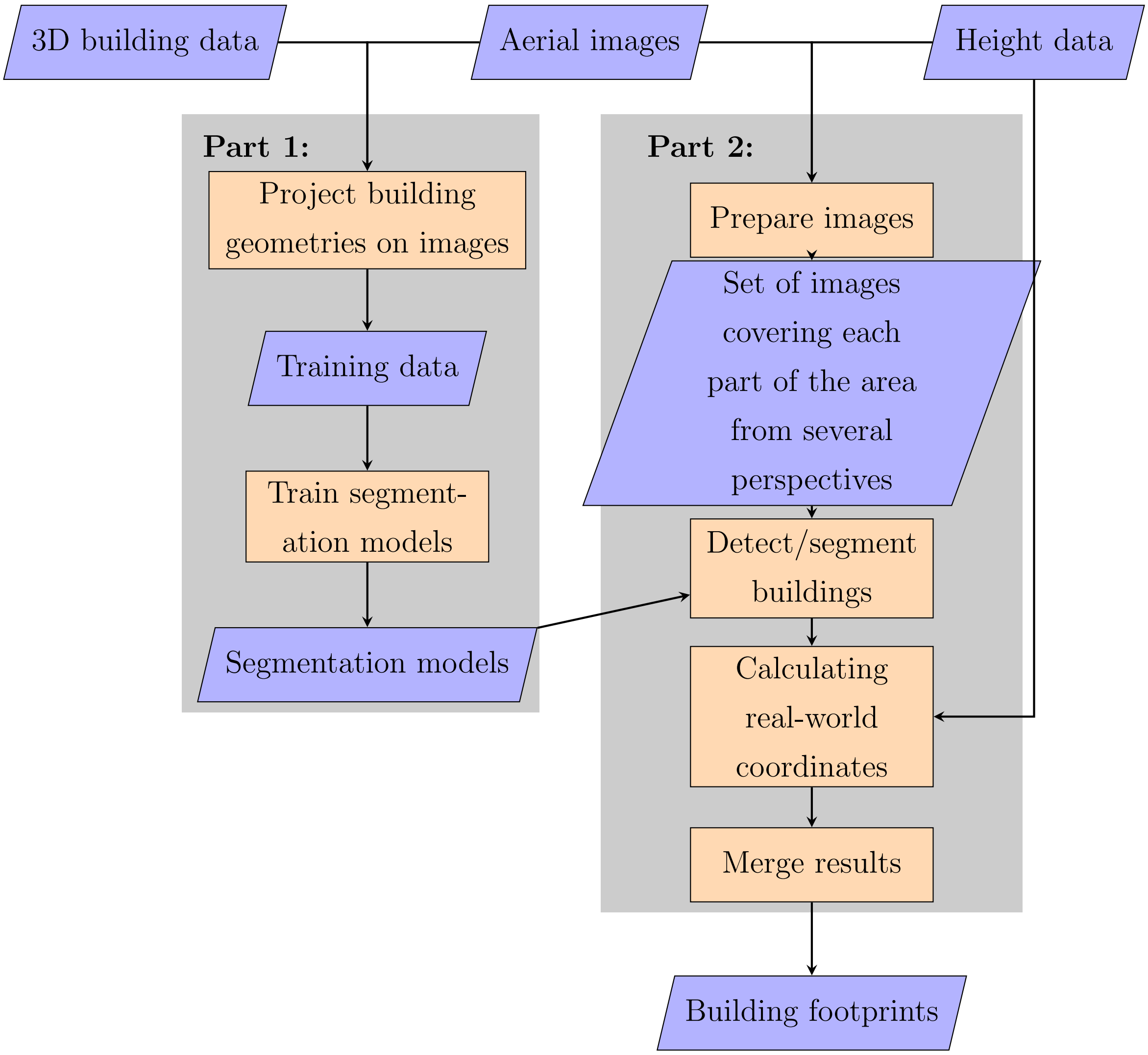

3D building data
Aerial images
Height data
Part 1:
Project building geometries on images
Training data
Train segmentation models
Segmentation models
Part 2:
Prepare images
Set of images covering each part of the area from several perspectives
Detect/segment buildings
Calculating real-world coordinates
Merge results
Building footprints

Figure 5: Overview of method

The method depends on three types of input data:

- **Aerial Images** including the exterior and interior parameters of the camera.
- **3D building data** which are used as ground truth to label aerial images to create training data for the segmentation models. To be able make accurate annotations on oblique images, this should preferably be accurate 3D models including detailed roof structures.
- **Height data** is required to transform the image coordinates to real world coordinates. To make accurate transformations, this should preferably be a Digital Surface Model (DSM) with a resolution of at least 1 m.

In part 1, the first step is to create training data by using the collinearity equations to project 3D building geometries onto the aerial images. This data is then used to train a segmentation model. Because the oblique and nadir images are inherently different, they are treated separately. This implies that two separate dataset is created and used to train two separate segmentation models.

In part 2, the area is first divided in smaller parts, and each part is extracted from aerial images from different perspectives. The previously trained segmentation models are then used to detect buildings in these images before the collinearity equations once again are used, this time to transform image coordinates of the detected buildings to real world coordinates. Finally, the results from each image can be merged into one mask. In the following chapter, each step of the method will be explored in detail, providing comprehensive explanations of the process involved in creating training data, training the segmentation models, and analyzing the area from multiple perspectives.

3.1 Building detection

The first part of the process is to create training data and train a semantic segmentation model to detect buildings on both nadir and oblique aerial images. Because the oblique and nadir images are inherently different, they are treated separately. This implies that two separate dataset is created and used to train two separate segmentation models.

3.1.1 Create training data

The method and tools for creating training data for building detection on oblique imagery was developed as part of a preliminary study. The aerial images are first divided in tiles of 512x512 pixels before the collinearity equations are used to project 3D building models on the images as explained in 2.2. After projecting the geometries to the image plane, the polygons are rasterized on a canvas, and the masks are saved as PNG files. For this project only one class is used, which means each pixel is either classified as a building or background.

When creating the training data there are some considerations that are important. Firstly, it is important that the samples are diverse. The images should contain buildings of various types, sizes and shapes. They should also include a variety of environments and non-building objects. These variations is important to make the segmentation model more robust and detect buildings in general and not only specific types of buildings in specific surroundings. A good way to achieve this is to get images from more than one region and make sure you cover both urban and rural environments.

The dataset is split in three partitions: training data, validation data and test data. Each partition serves its own purpose. The training data should be the majority of the images (80%) and is what the model uses during training. The model learns by adjusting its internal parameters based on patterns in the training data. The validation data is a smaller partition (10%), which is used during training to assess the model performance on unseen data. This helps in the process of fine-tuning the model parameters and to prevent overfitting. The test data (10%) is used to evaluate the model after the development process is complete, to provide an unbiased estimate of how well the model is able to generalize and perform on unseen data.

When dividing the dataset in partitions, a simple and common solution would be to create one large dataset and randomly pick samples for each partition. However, it is important that the three partitions are independent. If the images in the training data covers the same objects as in the validation or test data, one wouldn't be able to assess how well the the model have learnt general features or merely memorized specific features for those particular objects. When it comes to aerial images there is often a overlap between images and for oblique images you cover the same area from different angles. In this case it would be better to use images from different locations for the different partitions.

3.1.2 Training the segmentation model

When training a segmentation model, the first step is to choose a model architecture. The STT, presented in 2.3, is a suitable model for building detection, but it is also possible to experiment with other models. The original implementation of the STT model is found on "Hugging Face"[1], an AI community where people can deploy and share pre-trained machine learning models and datasets.

The STT repository includes two models, pretrained on the WHU (Ji et al., 2019) and the INRIA (Maggiori et al., 2017) dataset respectively. It may be possible to use these models directly, but because they are only trained on nadir images, the result are not very good on oblique images. In addition the WHU and INRIA dataset have a spatial resolution of 0.7 and 0.3m respectively and will probably not perform optimally with other resolutions. It is however useful to use one of these models as a starting point instead of starting the training from scratch. The technique of using the weights and parameters from a model trained for a similar task is referred to as transfer learning and is very useful to reduce training time and achieving better results with limited training data.

Data augmentation is another useful technique which involves applying various transformations or modifications to existing data samples. These transformations can include image rotations, translations, scaling, flips, or adding noise, among others. The purpose of data augmentation is to increase the diversity and variability of the training data, which helps the model generalize better and improve its performance.

[1] https://huggingface.co/KyanChen/BuildingExtraction

3.2 Area analysis

When analyzing an area, the idea is to first prepare sets of images that cover each part of the area from different angles. Then the segmentation model from part 1 will be used to detect buildings on each image before the predictions will be merged into one orthographic mask by determining the real-world coordinates of each pixel of the 5 masks and aggregating the confidence scores.

3.2.1 Prepare images

The first step is to divide the area to be analyzed by defining tiles of 50x50 meters, as illustrated in figure 6. Each tile will then be analyzed one by one. To identify which photos that are covering each tile, the coordinates of the tile corners is transformed to image coordinates using the collinearity equations as explained in 2.2. To do this a height value is also needed for each of the four corners. If all the image coordinates are within the bounds of the photo the tile is present in the photo.

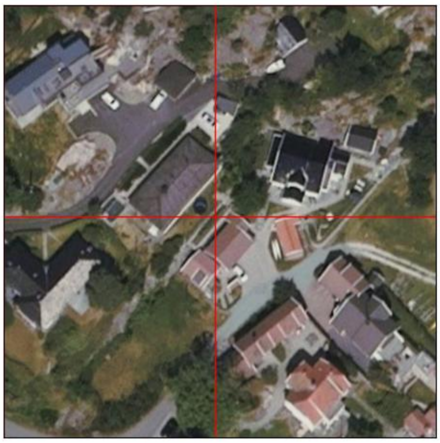


Figure 6: Example of a test area divided in tiles of 50x50 meter.

Because aerial imagery generally have a considerable overlap between the photos, the tile will be covered by multiple photos from each direction. While retaining all images would maximize data, it could lead to an imbalance because of different number of observations from each perspective. To avoid this, a weighting system or averaging method could be used, but the benefit would probably not be worth the extra complexity and computational time. Hence, a single image from each direction is selected by determining the photo where the tile area is closest to the center, based on the minimal sum of the image coordinates of the tile corners.

Finally the tiles are obtained from the selected photos. Because the spatial resolution is around 10 cm, 50 meters will correspond to approximately 500 pixels. However, due to the oblique perspective, the tiles will have irregular and inconsistent sizes and shapes. To meet the requirements of the segmentation model, images with a resolution of 512x512 are necessary. Resizing the images by interpolation is one way to address this requirement, but it can distort the images and potentially affect the performance of the segmentation model. To avoid distortion, a squared area is selected that includes the complete tile along with some surrounding area. If the selected area is too large, the image is resampled to achieve the required resolution of 512x512 pixels.

3.2.2 Segmentation

When the test area is split into smaller tiles and each tile is represented by 5 images from north, south, west, east and above, the data is ready for analysis. Each image is analyzed by the segmentation models and the result is a 512x512 mask where each pixel has a score between 0 and 1, indicating how confident the model is that the given pixel represents a building. Pixels with a value closer to 1 means the model has high confidence that the pixel depicts a building. Figure 7 shows an example of images and corresponding segmentation masks from different perspectives.

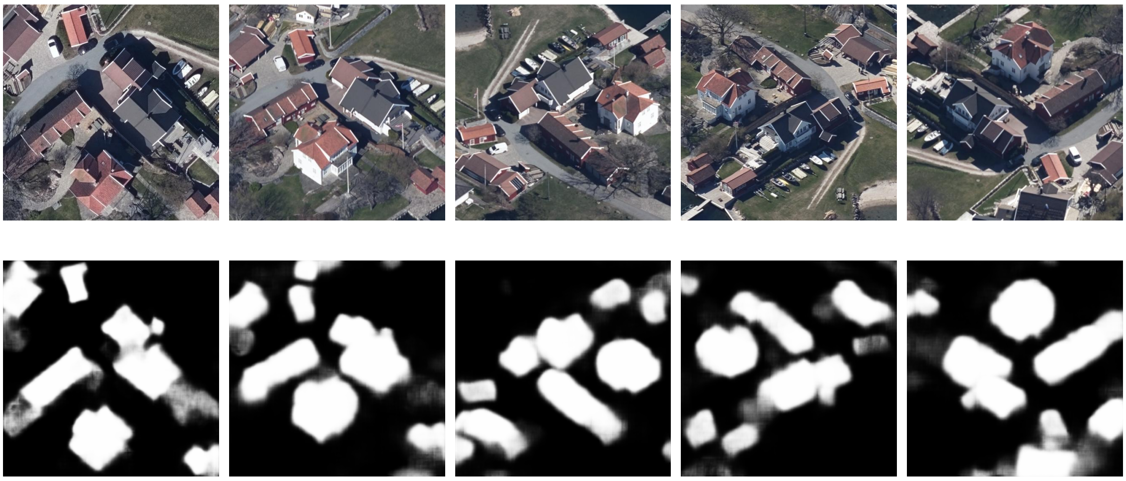


Figure 7: Images from different perspectives and corresponding predictions from the segmentation model.

3.2.3 Calculating real-world coordinates for the AI predicted buildings

In part 1, the collinearity equation was used to project 3D geometry of buildings on oblique images. The following section will describe the opposite process, that is finding the real-world coordinates of the buildings detected on the oblique images. As mentioned in 2.2, the collinearity equations only describes a straight line from the image sensor and through the camera perspective centre. To know where this line intersects with the object on the image, we also need to know the height value (Z) for each pixel. The height values are obtained from a Digital Surface Model (DSM) with a spatial resolution of 1m, as illustrated in Figure 8.

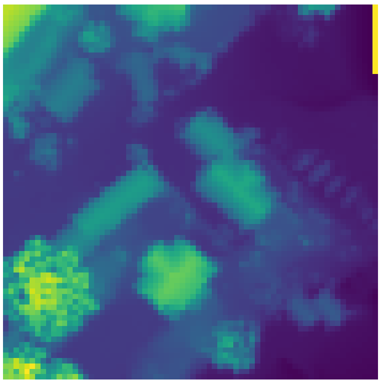
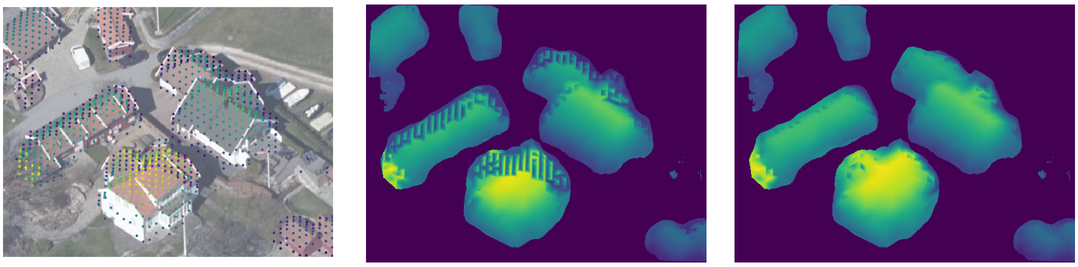


Figure 8: Digital Surface Model for the area.

To determine the height value for each pixel in the oblique image, the DSM must first be transformed to image coordinates. Taking the real world xy-coordinate of each pixel along with the given height value, the collinearity equations (1) is used to assign each height value to a coordinate in the oblique image. Figure 9a shows the set of heigh values after discarding all points outside an area with detected buildings, as they are not needed. Following this step, linear interpolation is used to estimate the height values for all the remaining building pixels (figure 9b). Upon reviewing the results, it becomes apparent that the upper parts of the buildings have some conflicting height values. This is because both the height value from the roof and the ground behind it, is both projected to the same area in the oblique images as illustrated in figure 4. To solve this problem, every point that has 5 or more neighbors within $10px * \frac{image\ size}{500} \approx 1m$ and has a lower height value than the average in this neighborhood are discarded. Figure 9c displays the final result after this filtering process.



(a) Height values in image coordinates.

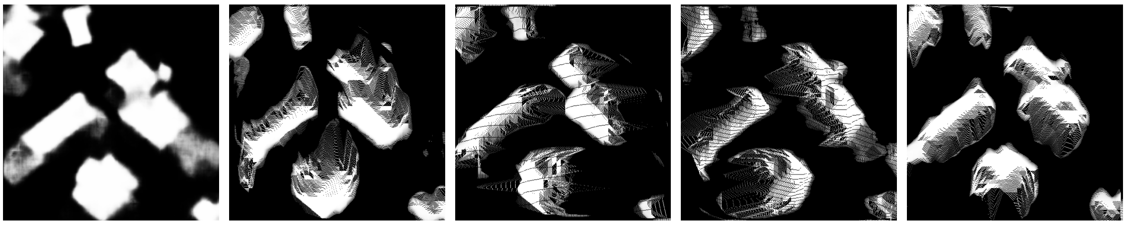
(b) Interpolated height values.

(c) Final height values

As all pixels identified as a building now have a height value, it is now possible to unambiguously determine the real world coordinate by using the collinearity equations solved for X and Y (equation 3).

3.2.4 Merge results

When each pixels is transformed to real-world coordinates, the prediction score is added to an orthographic raster for each image, as shown in figure 10. The raster has a resolution of 500x500 pixels, where each pixel represents a square of 10x10 cm. When transforming the pixel coordinates to real world coordinates there will be some pixels that will have the same real world coordinate. For example will an area representing a wall, be collapsed to a line in the orthographic projection. To aggregate these values to a single score between 0 and 1, the maximum value is chosen.



(a) Nadir

(b) South

(c) West

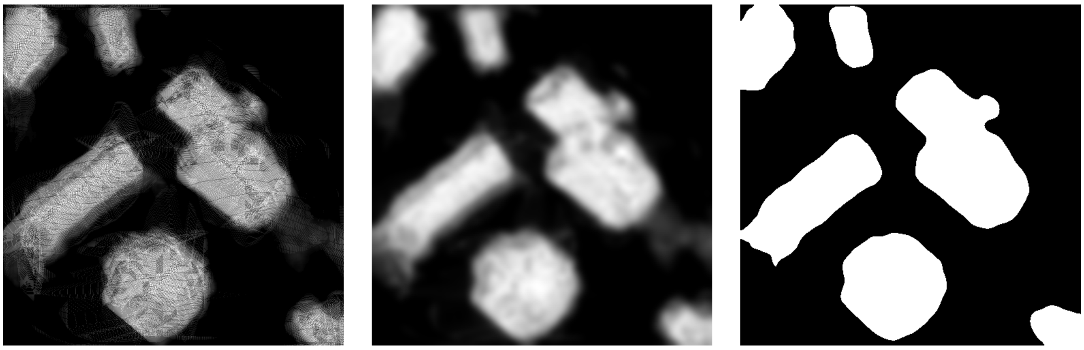
(d) East

(e) North

Figure 10: Ortographic mask based on the nadir image and 4 oblique images.

Finally the score for all five images are summarized. The resulting mask, displayed in figure 11a, is somewhat noisy, mostly because of some wrong transformations caused by inaccurate height values. There are also points that hasn't got any "observations". To account for this, the result is smoothed by using a gaussian filter (see figure 11b). Finally a binary mask is made by selecting a threshold value. Figure

11c shows a binary mask with a threshold of 1.5. Intuitively this value could mean that for example two images are 75% confident that there is a building or one image is 100% confident and another is 50%.

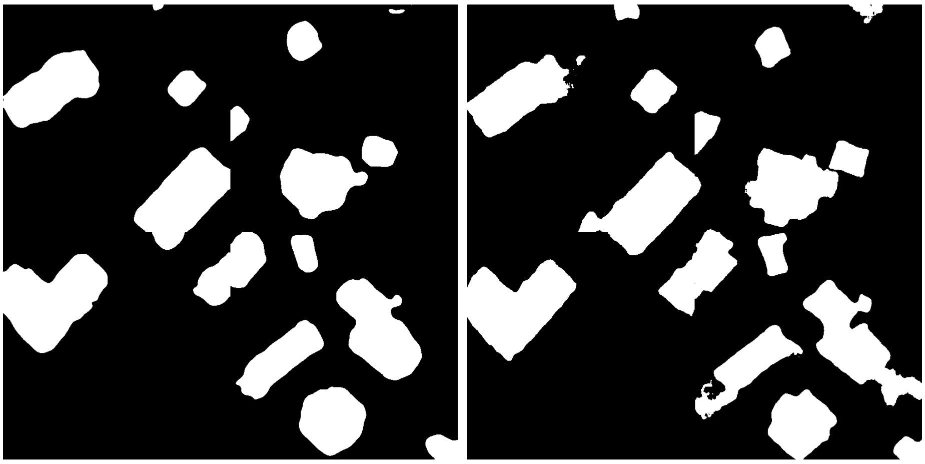


(a) Summarized ortographic mask.

(b) Smoothed with gaussian filter.

(c) Binary mask

When the result has been merged to a single orthographic raster, it can easily be compiled into one large raster covering the whole test area as shown in figure 12.



(a)

(b)

Figure 12: Compiled result based nadir and oblique images (a) and only nadir images (b)

4 Data

To test the proposed method, an experimental study is conducted with data from the southern part of Norway. This chapter will describe the raw input data, that is the aerial images, 3d building data and height data. It will also describe the training dataset that is created and used to train the segmentation models.

4.1 Oblique aerial images

There are two sets of aerial images used in this project. The first set was captured in 2021, covering Grimstad city, and the second set was captured in 2022, covering Lindesnes. Both set of images were commissioned by Kartverket Agder and delivered by Terratec AS.

The camera system used was UltraCam Osprey Prime II for Grimstad and UltraCam Osprey 4.1 for Lindesnes. Both systems consists of one nadir camera and four oblique cameras with an eccentric angle of 45 degrees. The resolution of the different cameras is displayed in the table below (1).

Table 1: Resolution of cameras.

	Nadir	**Oblique**
UltraCam Osprey Prime II	13470 x 8670	10300 x 7700
UltraCam Osprey 4.1	20544 x 14016	12840 x 8760

Aerotriangulation through bundle block adjustment was already applied by Terratec and the estimated exterior parameters were delivered as metadata along with the images. The interior parameters are found in the calibration reports for each camera system. The metadata was delivered on a different format for the two image sets. The first set was delivered with the SOSI-format (not to be confused with the SOSI Standard), which is a plaintext format for representing geodata. To parse these files to a python dict, a small python library named ”sosi” (Dillon, 2017) was utilized. The second image set was delivered with the more well-known shapefile structure. This includes a .shp-file containing geometries describing the covered area for each image, a .shx file that indexes the geometry and a .dbf file that stores attributes including the cameras exterior parameters. To read these files a python library named ”PyShp” (Gillies et al., 2022) was used.

4.2 3D building models

The images are annotated using 3D building models provided by Norkart AS, which are based on the norwegian common map database (FKB). FKB is a collection of primary geodata collected and managed by multiple parties within each municipality in Norway. The FKB data includes building geometries such as outlines, height, roof attachments, and ridge lines. This data is used to create detailed roof structures, and the wall surfaces are generated from the roof edge until they meet the terrain or a lower roof surface.

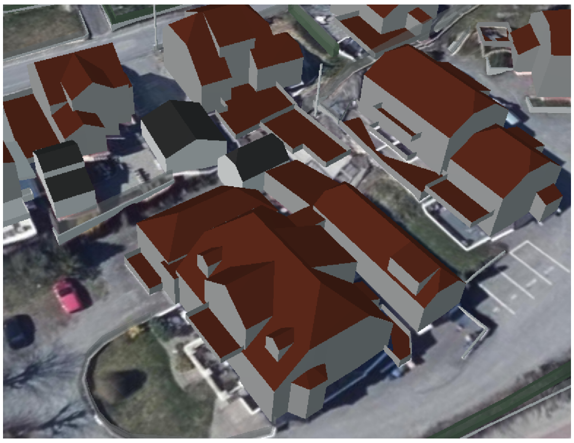


Figure 13: Example of the 3D building models used as ground truth (screenshot from www.norgei3d.no)

An specific issue with these 3D models is that they include terraces which are labelled as roof surfaces. The terraces are also surrounded by autogenerated railings labelled as wall surfaces. As terraces are not considered to be a part of buildings, this would introduce significant errors in the segmentation. To overcome this problem, some custom logic has been developed to recognize and remove these terraces and railings. This is done by examining the building polygons, looking for certain patterns and characteristics observed to match the terraces and the autogenerated railings. Below is a list of such characteristics.

- All autogenerated railings have an height of 75 cm.
- All terrace polygons are flat (all vertices have the same height).
- All terraces intersects (shares a vertex) with at least one autogenerated railing.
- All terraces have edges without a proper wall beneath.

4.3 Digital Surface Model (DSM)

Height data is necessary in the second phase to do the transformation from image coordinates to world coordinates. A digital surface model (DSM) is a type of digital elevation model that represents the earth's surface, including both natural features like terrain, vegetation, and built structures like buildings and infrastructure. This model is acquired from geonorge[2], an open service for a national digital height model, serving a surface model with a resolution of up to 1 meter.

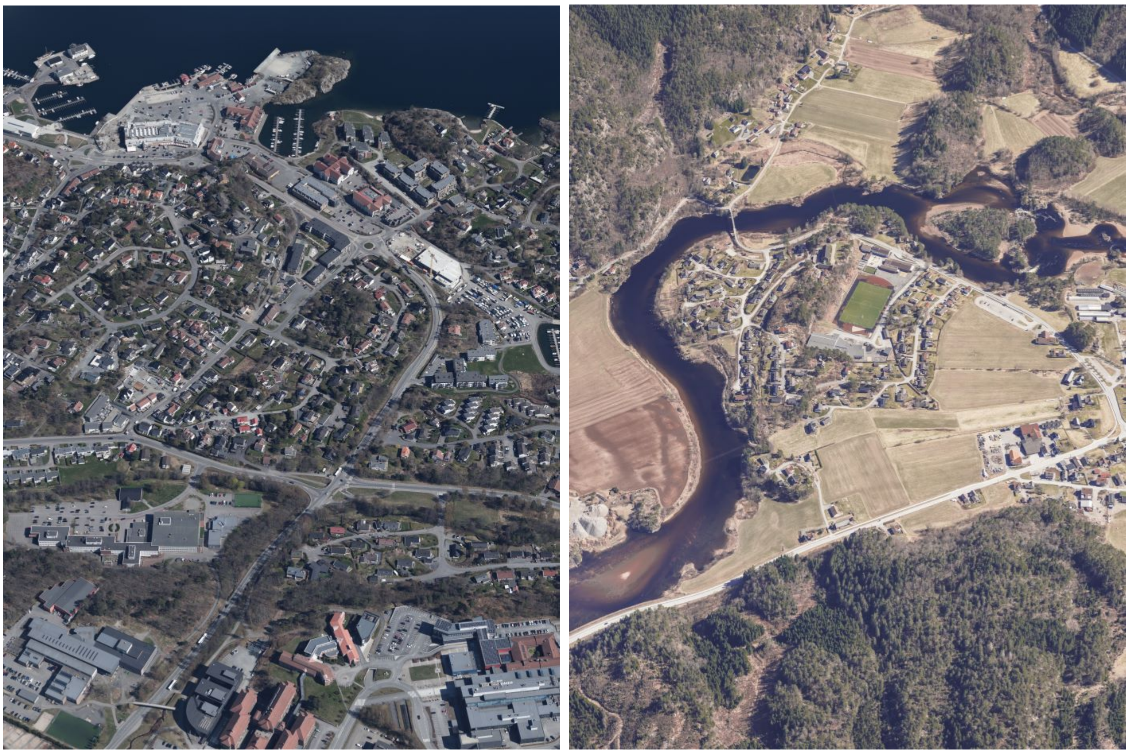
4.4 Training dataset

To create the training data a total of 125 images was used from Lindesnes, of which 25 are nadir and 100 are oblique images. From Grimstad there are only 1 nadir and 14 oblique images. After the images are divided in tiles of 512x512 and labeled as described 3.1.1, the resulting dataset contains 181 nadir and 1830 oblique samples from Grimstad and 5634 nadir and 12216 oblique sample from Lindesnes.

It is important to acknowledge that the dataset is automatically generated based on known building geometry, leading to potential errors. Two sources of error are outdated or incorrect data and labeling buildings even when they are occluded. The latter is even more prominent on oblique images because of more occlusions.

Figure 14 shows that the characteristics of the images are quite different for the two areas. The images from Grimstad has a denser population, a lot of asphalt, some sea and not much vegetation. The images from Lindesnes has a lot of vegetation and fewer buildings per image. There is also a noticeable difference when it comes to the colors due to different light conditions.

[2] https://kartkatalog.geonorge.no/metadata/nasjonal-hoeydemodell-digital-overflatemodell-25832-wcs/a456f3c2-96f6-42f0-9960-e9888fc0c2de



(a) (b)

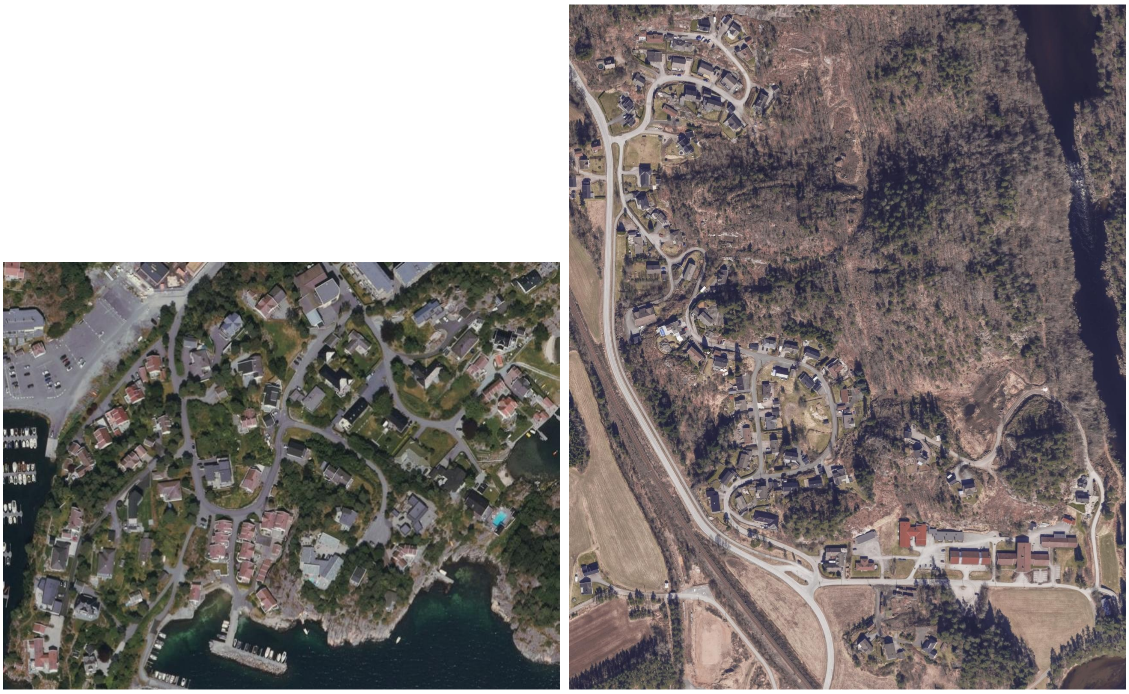
Figure 14: Example of photos from Grimstad (a) and Lindesnes (b)

As explained in 3.1.1, the dataset is divided in three partitions. The size of the dataset and the distribution between the partitions can be found in table 2.

Table 2: Size of the two datasets and their partition distribution.

	Nadir		**Oblique**	
	#	%	#	%
Train	5815	83.9	14046	83.5
Val	506	7.3	1491	8.9
Test	612	8.8	1279	7.6
Total	**6933**		**16816**	

As also mentioned, it is important that the partitions are independent of each other. In this case the images are taken from several perspectives, and there is considerable overlap between each series of images, so there will be several images covering the same area and the same buildings. For this reason two designated areas are chosen for test and validation data. The validation data is based on an area in the southern part of Grimstad city, while the test data are from Marnadal, a small town north in Lindesnes municipality. Both areas are depicted in figure 15.



(a)

(b)

Figure 15: Validation area from Grimstad (a) and test area from Lindesnes (b)

5 Experimental study

In this chapter, the proposed method outlined in Chapter 3 was applied to the data described in Chapter 4. The chapter will explain the process of the experimental study and the following results will be presented and discussed.

The first part of the chapter will initially address the training of the segmentation model, focusing on the selection of hyperparameters, data augmentation, and training time. Furthermore, the model's ability to detect buildings will be evaluated. Examples will be highlighted and discussed to shed light on the model's challenges and to examine its performance on oblique aerial images compared to nadir aerial images. The last part of the chapter will present the results of an area analysis, investigating how accurate and effective the proposed method is in combining predictions from multiple images to achieve improved and more reliable building detections.

The objective of this study is to evaluate the accuracy and reliability of the proposed approach for building detection using oblique aerial images, while also providing valuable insights into its limitations and challenges, and suggesting potential improvements and avenues for future research.

5.1 Training the segmentation model

For this study the STT model is used as described in 3.1.2. The model which is pretrained on the INRIA dataset is used as a basis, and two separate models are further trained separately on the oblique and nadir dataset (4.4). The only hyperparameters changed from the original implementation is the learning rate, which is reduced from 10^{-3} to $5 * 10^{-4}$, and the mean values and standard deviation (STD) used to normalize the data, which are calculated based on the training set.

The data augmentation used is the same as in the original implementation. This includes color distortion, random cropping, mirroring and vertical flip. The training is performed on 2 GPUs of type Nvidia GeForce GTX 1080. The training time for the two models is found in table 3.

Table 3: Training time

	Epochs	Training speed [min/epoch]	Total training time [min]
Nadir	15	5	75
Oblique	14	12.5	175

5.2 Result of building detection

To evaluate the segmentation models, Intersection over Union (IoU) is used as metric to measure the overlap between the predicted building mask and the ground truth. IoU is defined as follows:

$$\begin{aligned} IoU &= \frac{Area\ of\ intersection}{Areaofunion} \\ &= \frac{True\ positives}{True\ positives + False\ postives + False\ negatives} \end{aligned} \tag{4}$$

To evaluate the performance of the segmentation model, it is tested before and after further training on three different set of test data: The nadir and oblique version of the Lindesnes+Grimstad dataset and a few samples from the INRIA dataset. Table 4 gives an overview of the IoU score for each model on different dataset.

Table 4: IoU scores before and after training on different test data.

Test data			**Pretrained**	**Nadir model**	**Oblique model**
Name	**Perspective**	**GSD [m/pixel]**			
Lindesnes	Nadir	0.1	0.671	**0.840**	-
Lindesnes	Oblique	0.1	0.547	-	**0.831**
Lindesnes	Nadir	0.3	0.73	-	-
INRIA	Nadir	0.3	0.8064	0.5581	0.5324

Initially, the pretrained STT model performs a lot worse on our created dataset, than on the INRIA test samples. This suggests that the segmentation model lacks robustness and generality in making accurate predictions across diverse sets of aerial images. These results aligns with the challenges related to building detection, as discussed in 2.3. It also demonstrates the need to train the segmentation model on more representative data. The most obvious difference between the two dataset is the spatial resolution. While our dataset has a resolution of 0.1 meter, the STT model is trained on the INRIA dataset with a resolution of 0.3 m. This means that the model is used to see images covering an larger area and with less low level details. To see how much this affect the performance the model was also tested with a resampled version of the Lindesnes dataset with the same spatial resolution as the INRIA dataset. As expected this gave much better results, but there is still a IoU drop of around 7. This is probably due to other variations between the datasets like different building types, surroundings and the spectral characteristics of the images.

After training, we see that the tables are flipped and our models performs very good on the Lindesnes test data and terrible on the INRIA dataset. Again this confirms that making a robust and general building detection model is hard. To achieve good results we either need to train on similar images as we want to predict on, or we would need training data of much higher quality with a more diverse selection of samples.

We also see that after training we still get slightly better result on nadir images compared to oblique images. One might expect better result on oblique images, considering that the visibility of vertical structures should make it easier to distinguish buildings from non-buildings. On the other hand, Conversely, oblique images have more occlusions leading to more errors in the dataset, as discussed in 4.4. This might impact the training to some extent and will likely result in a lower IoU score due to errors in the test data.

Example results

The following section will present various examples of the segmentation results to gain a better understanding of how the segmentation models perform on both nadir and oblique images. The goal is to explore the advantages and challenges associated with utilizing oblique aerial images for building detection. Moreover, the focus is on identifying instances where the models struggle and discuss possible explanations for these shortcomings.

Figure 16 depicts some good results. The overlay includes three colors to visualize how the predictions are compared to the ground truth labels. GREEN means the prediction match the label (true positive), RED means labels that hasn't been detected (false negative) and BLUE means predictions that doesn't match with the label (false positive). Note that the test data is not perfect and the red pixels in figure 16a is an example of an inaccuracy in the dataset labels, including the balconies as part of the building.

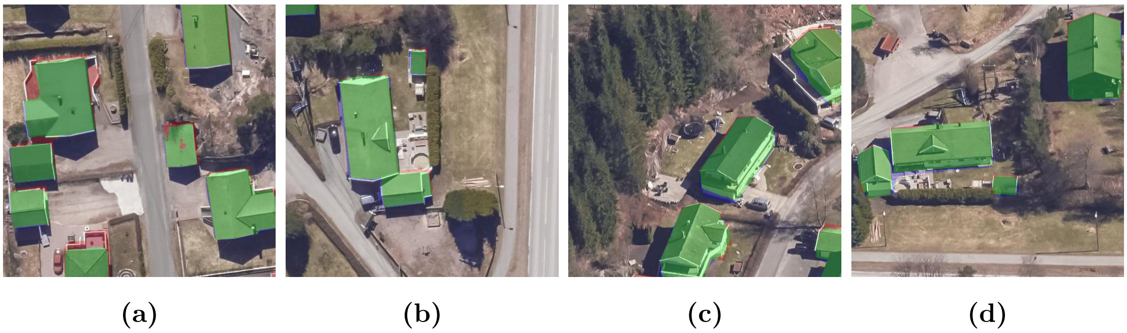

(a) (b) (c) (d)

Figure 16: Good segmentation results on nadir (a-b) and oblique (c-d) images.

Figure 17 displays some instances that illustrate the limitations of using nadir images for building detection. In (a-b), the model struggles to distinguish terraces and misclassifies them as part of the building. Similarly, (c) shows a truck being erroneously identified as a building, while (d) depicts a garage that goes unrecognized.

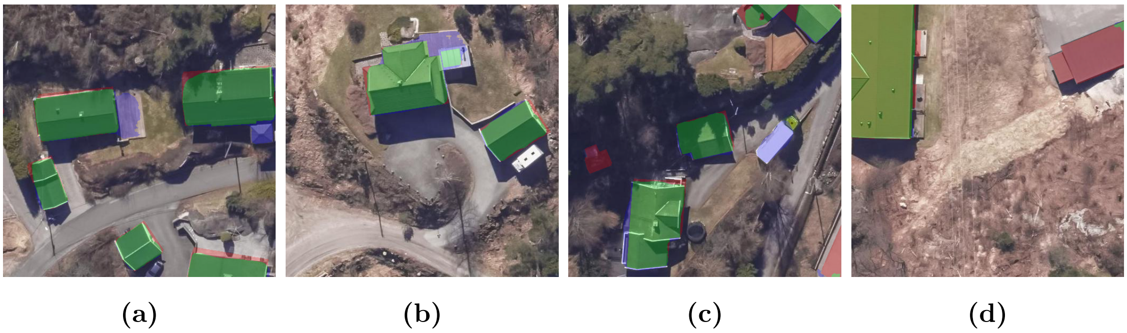

(a) (b) (c) (d)

Figure 17: Examples illustrating the lack of distinguishing features on nadir images.

From a top-down perspective, terraces, trucks, and simple building like garages often appear as regular-shaped polygons without any other distinguishing features. Figure 18 displays the same ares from a oblique perspectives, demonstrating how this reveals more features and makes it much easier to distinguishing these objects.

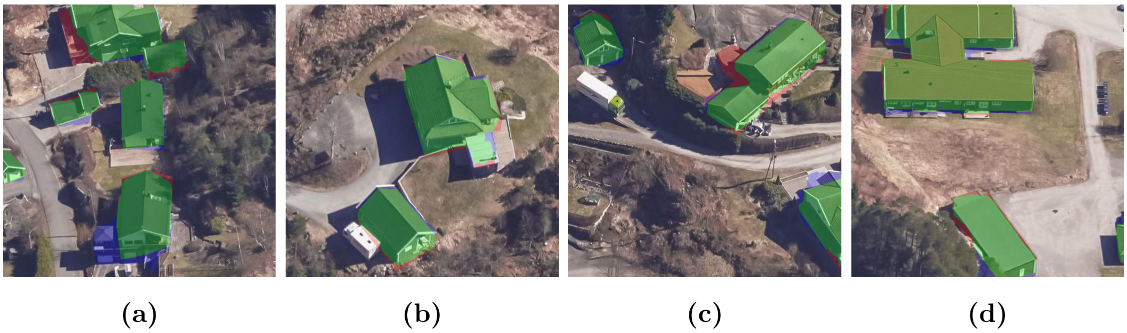


Figure 18: Viewing the same areas from an oblique perspective reveals more features and enables more reliable detection of buildings.

Detecting small buildings below $50m^2$ is a key focus of the KartAI project. These buildings are often exempt from the application requirement and thus remain unregistered. Figure 19 demonstrates the model's capability in detecting small buildings on nadir images. However, it also reveals some inconsistencies. For instance, both (a) and (b) show small buildings that were not detected, while in nearly identical images (c) and (d), the buildings are correctly identified. (e) and (f) shows that the same inconsistency can be found in the oblique images. This disparity can be attributed to the model's indecisiveness, as predictions near the 50% threshold can lead to varying results.

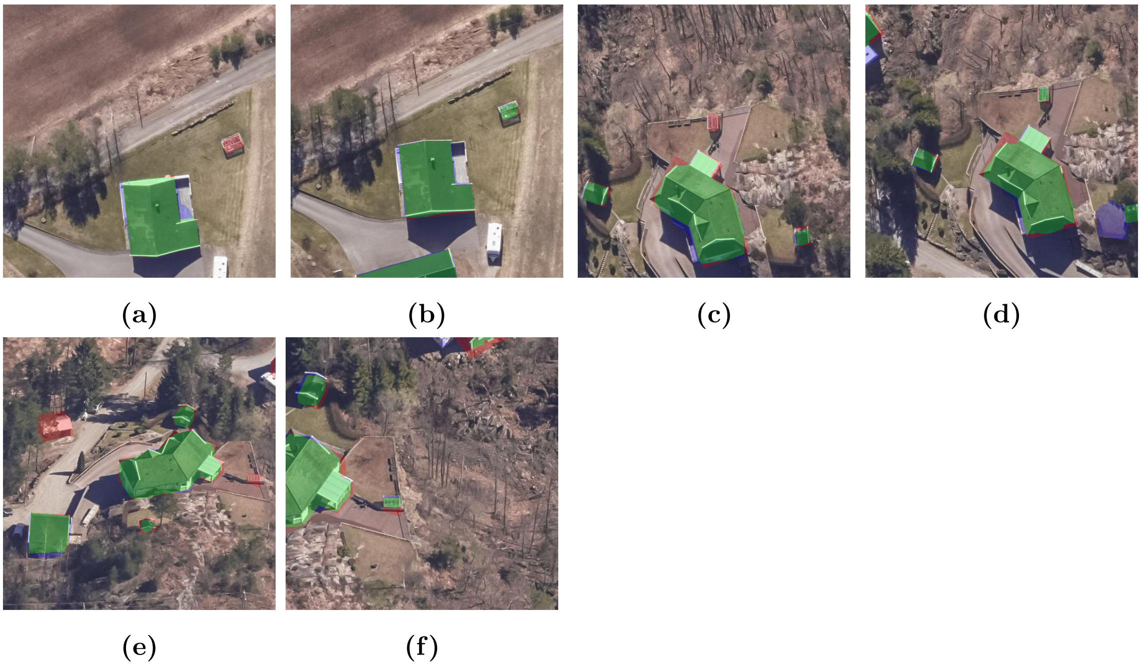


Figure 19: Examples demonstrating the inconsistent performance on small buildings

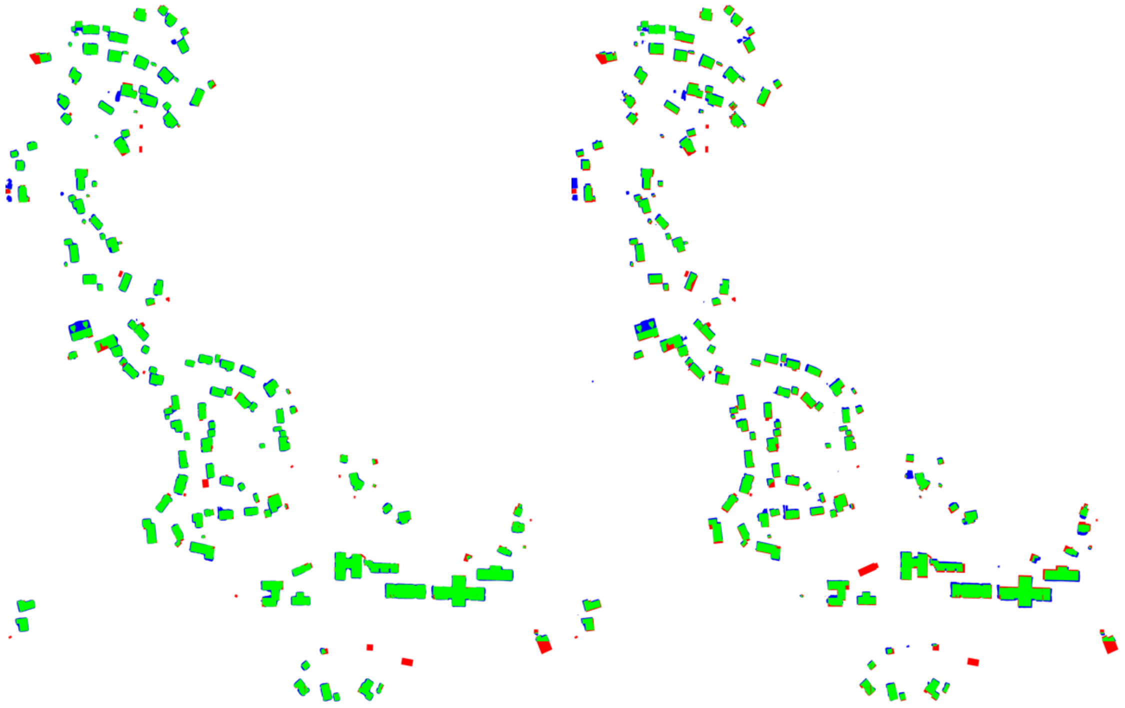
One of the reasons behind the model's struggle with small buildings could be related to shortcomings in the training data. An interesting aspect to consider is that known building data is utilized as ground truth to train the building detection model, which in turn is used to identify errors in the same building data upon which it was originally trained. This recursive approach is not ideal for machine learning, which relies on accurate and reliable training data to learn patterns and make precise predictions. However, it is generally assumed that the building data used as ground truth is of such high quality that the model is expected not to be significantly impacted by the few errors present in the training data. Because, as mentioned earlier, many small buildings are not included in official registries, this can still pose a challenge, especially for smaller structures.

5.3 Result of the area analysis

The following section will present the results of an area analysis, where building footprints are extracted using several images captured from various angles, as explained in 3.2. The previous section explored the performance of the building segmentation models on a single image, and the results demonstrated the advantages of oblique images and their potential to improve the reliability of automatic building detection. While the following section also includes examples to support these results, the primary focus is to evaluate the proposed method for analyzing an area by combining the segmentation results of several images.

To evaluate the accuracy and reliability of this method, the combined results will be compared to the segmentation results from the more traditional approach of only using nadir images. The building data presented in 4.2 will be used as a ground truth. The result will be evaluated by calculating the IoU, accuracy and recall with respect to the ground truth and by a visual inspection of the segmentation results and some relevant examples.

The area that will be analyzed is an area in Marnardal in Lindesnes municipality. It is the same area that is used to create the test dataset (4.4) which is used to analyze the segmentation model in the previous section. The area has a size of 650x800 meter and contains 158 registered buildings. It is divided in 208 tiles of 50x50m and each tile is analyzed one by one before the result is compiled to a large raster and compared to the ground truth. Figure 20 shows the result for both the combined results and the nadir results.



(a) Combined result

(b) Nadir result

Figure 20: Segmentation results. The colors represent true positives (green), false positives (blue) and false negatives (red).

By counting the number of correctly detected buildings (true positives), wrongly predicted buildings (false positives) and undetected buildings (false negatives), we can calculate the recall and precision of each method. The recall is a measurement of sensitivity and tells us how many of the buildings are detected, while precision tells us how many of the predictions that are correct. The metrics are defined as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negative} \tag{5}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{6}$$

The Intersection over Union (IoU) will also be used as a metric to measure the accuracy of the segmentation masks. By summing up the number of pixels classified as true positives, false positives and false negatives, the IoU can be calculated by using equation 4. Table 5 shows the number of counted buildings and all the calculated metrics. Note that while the recall and precision is based on the number of buildings, the IoU is based on the area (number of pixels) of the segmentation masks.

Table 5: Analysis results

	Detected buildings			Recall	Precision	IoU
	TP	**FN**	**FP**			
Combined	150	13	1	0.92	0.99	0.774
Nadir	152	11	6	0.93	0.96	0.684

While the results show that the results based only on nadir had marginally better recall rate, all other metrics indicates that the combined results is both more accurate and reliable. The most surprising result was that the IoU on the nadir images was so low compared to the reported IoU on the test dataset which even covers the same area. The reason for this, though, is assumed to be caused by relief displacement. The dataset used to train and evaluate the segmentation model is based on 3D projections of the building geometries, which implies that the relief displacement is accounted for and walls are correctly classified as building. The ground truth used in this analysis on the other hand, is based on the an orthographic projection of the building which will not be accurate unless the image is taken directly from above.

Figure 21-23 divides the area in three parts and shows a more detailed overview comparing the two analyses. Colors are used to represent all the combinations for how the two analyses have correct/incorrect detections or fail to detect buildings. The meaning of the colors is described in table 6.

Table 6: Explanation of the colors used in the overview in figure 21-23

	Combinations			Description
	Combined	**Nadir**	**Truth**	
⬜	+	+	+	Correct detection of both analyses
🟥	-	-	+	Not detected by neither of the analyses
🟩	+	+	-	Incorrect detection by both analysis
🟨	+	-	+	Detected only by the combined analysis
🟦	-	+	+	Detected only by the nadir analysis
🟪	+	-	-	Incorrect detection by the combined analysis
🟧	-	+	-	Incorrect detection by the nadir analysis

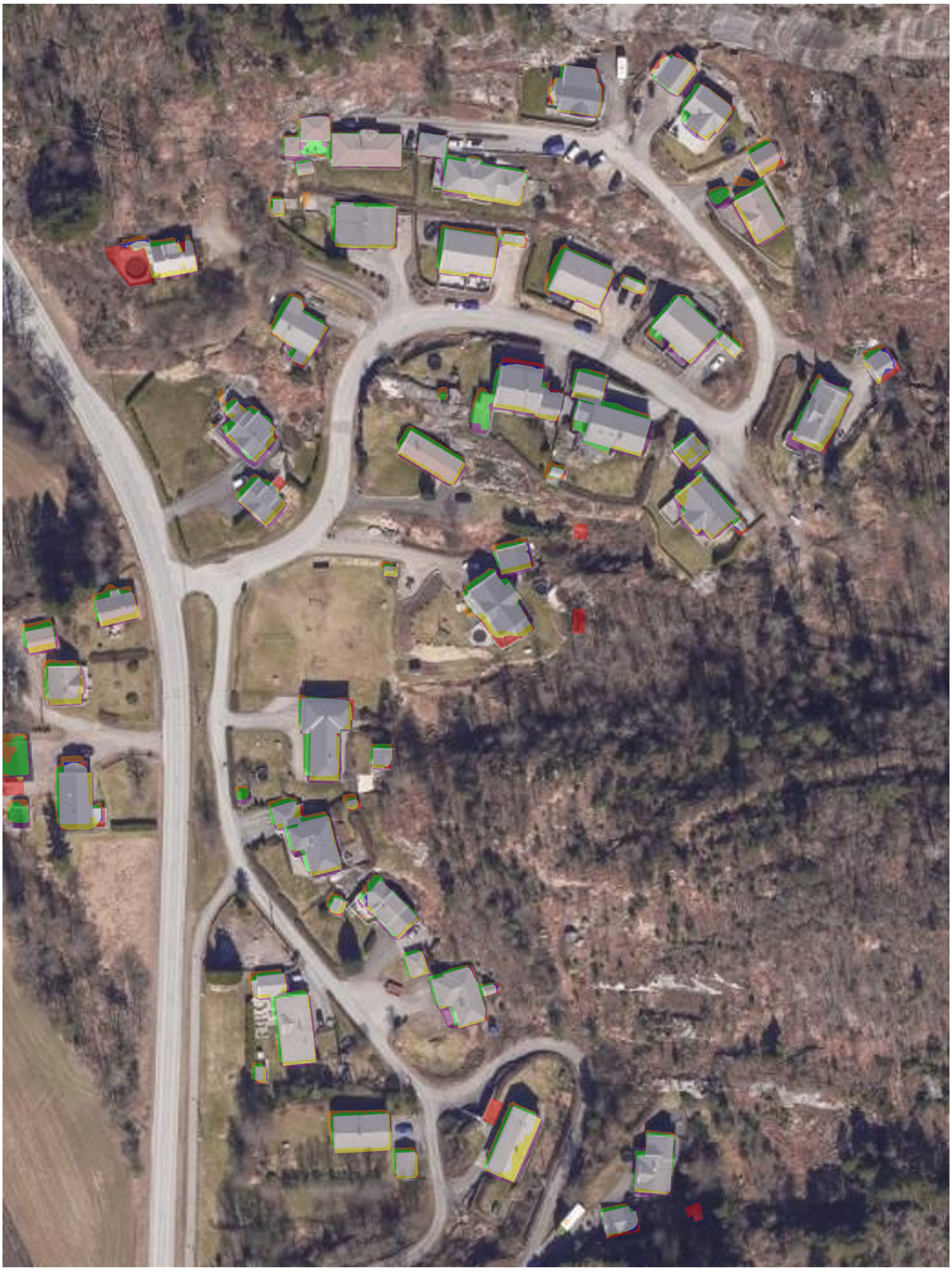


Figure 21: Analysis result (Part 1)

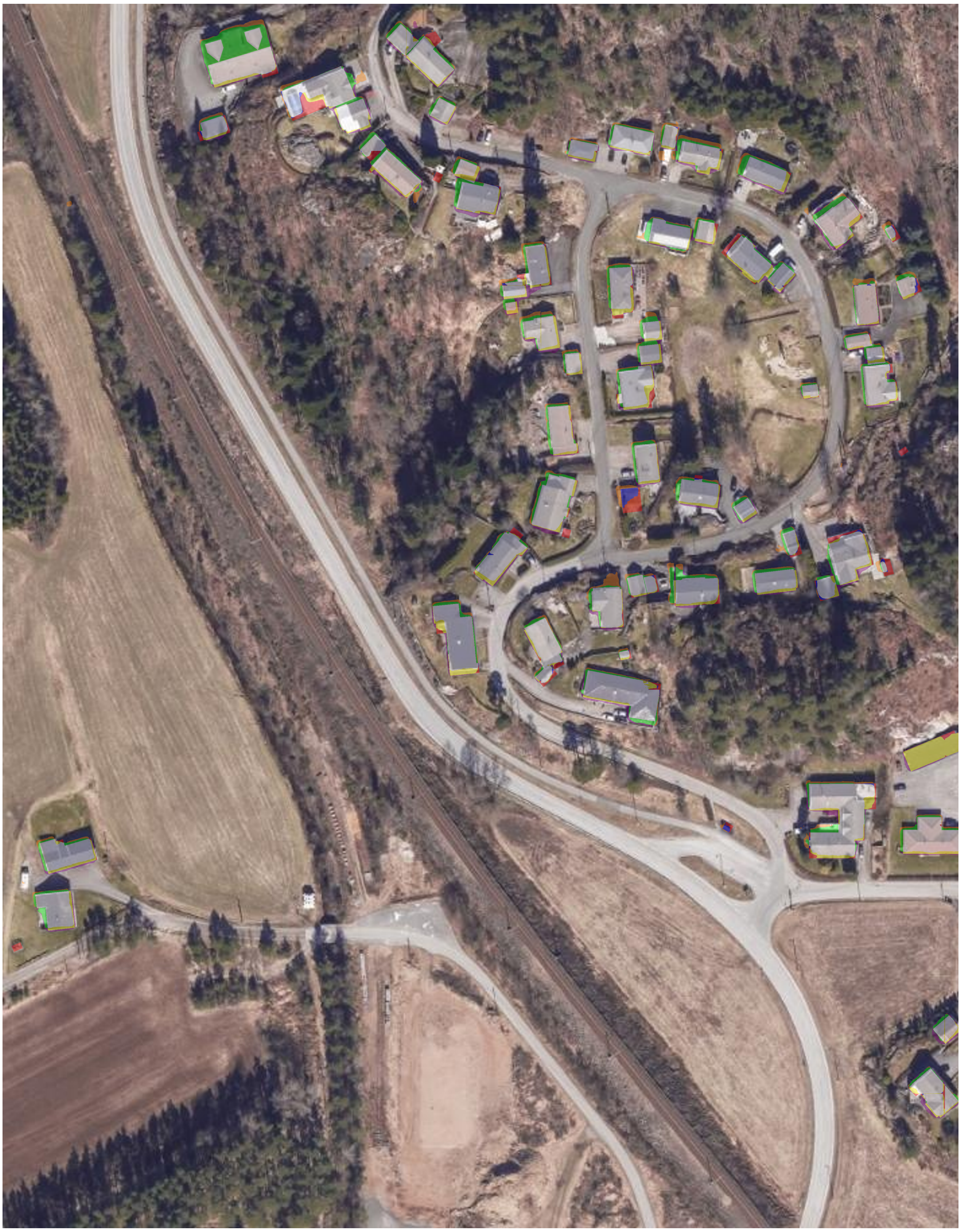


Figure 22: Analysis result (Part 2)

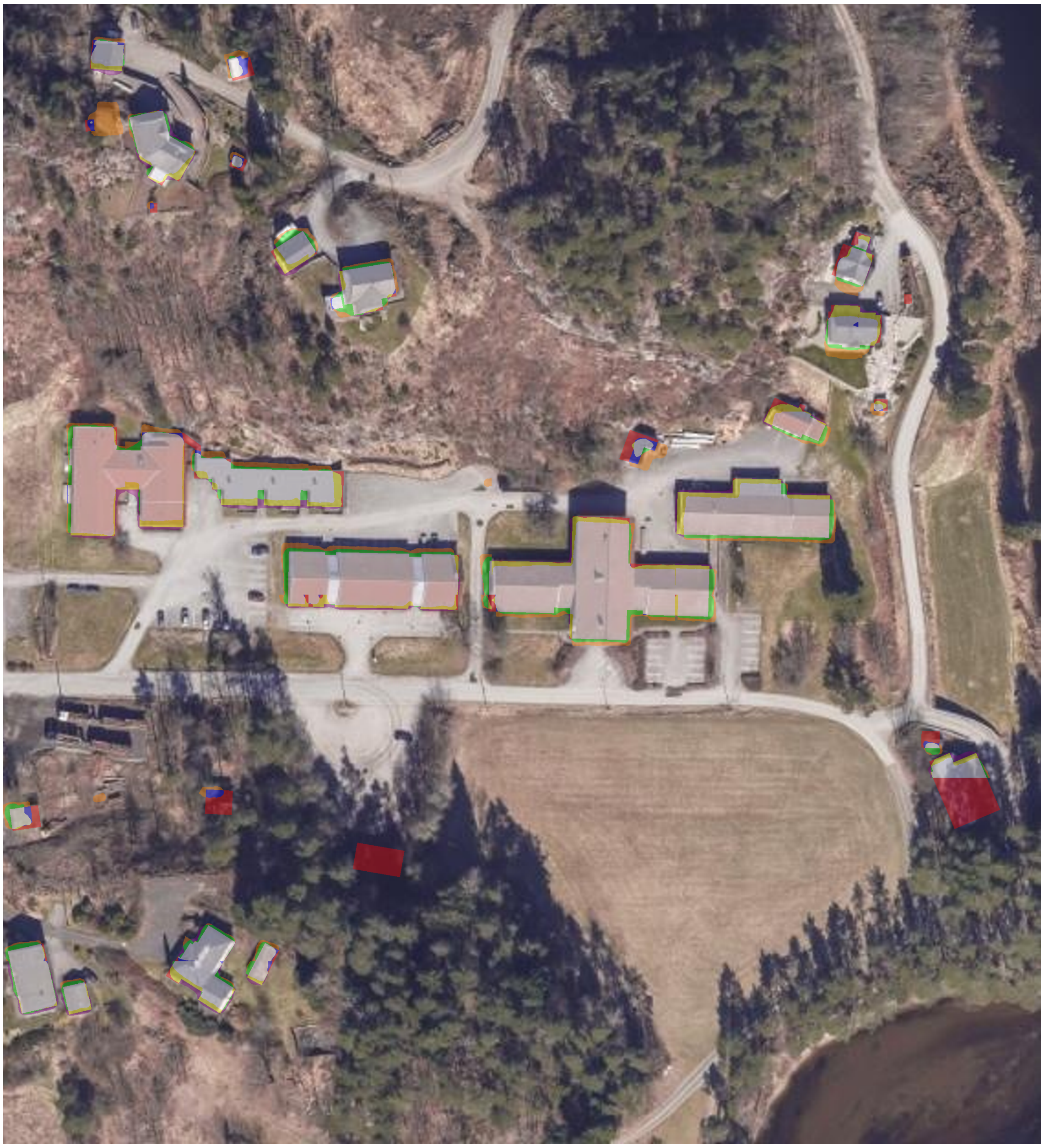


Figure 23: Analysis result (Part 3)

Example results

In this section, a series of examples will be explored to demonstrate and evaluate the proposed method for combining predictions from multiple images. These illustrative examples aim to highlight the strengths and limitations of the method and identify potential areas for improvement.

Figure 24 shows a good example of where the proposed method for combining predictions work very well. The figure shows that the nadir model doesn't even notice the garage, while all the oblique predictions agree that it is a building and gives an accurate outline of the building footprint. It also gives a better result on the building to the left, where the nadir model understandably leaves out a small part that is hard to distinguish from the ground.

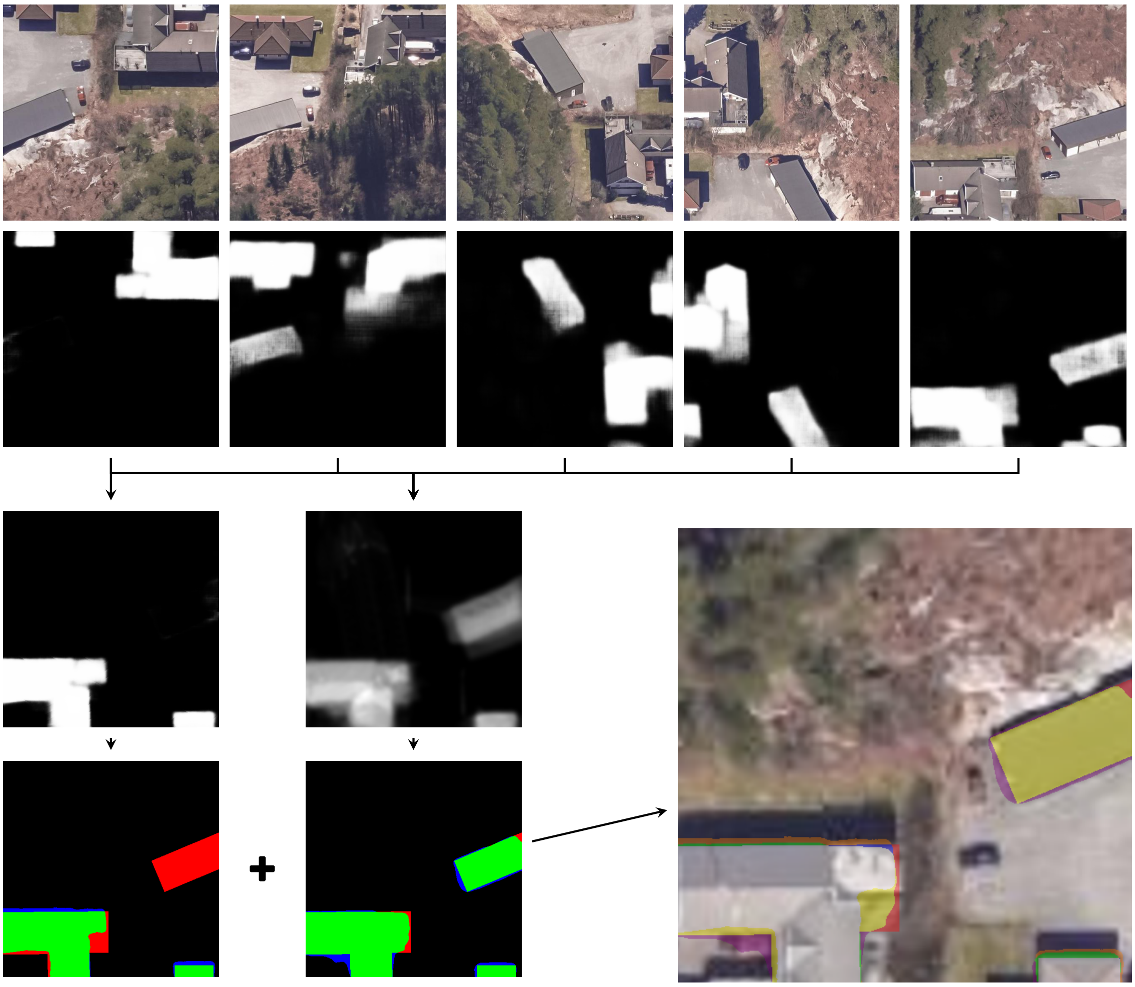


Figure 24: A garage not detected on the nadir image.

Figure 25 gives an example where the nadir models incorrectly classifies a lawn as a building. Once again the prediction make sense from a top-down perspective. All you see is a green rectangle which could be a roof and a steep edge and a shadow on one side, which is very typical for buildings. From an oblique perspective, however, it is easy to see that it is not a building. This examples shows a challenge with the purposed method where there is room for improvement. Right next to the lawn there is a small building, which is noticed in three of four oblique images, but the overall confidence is below the threshold value. This challenge will be further discussed in the next example.

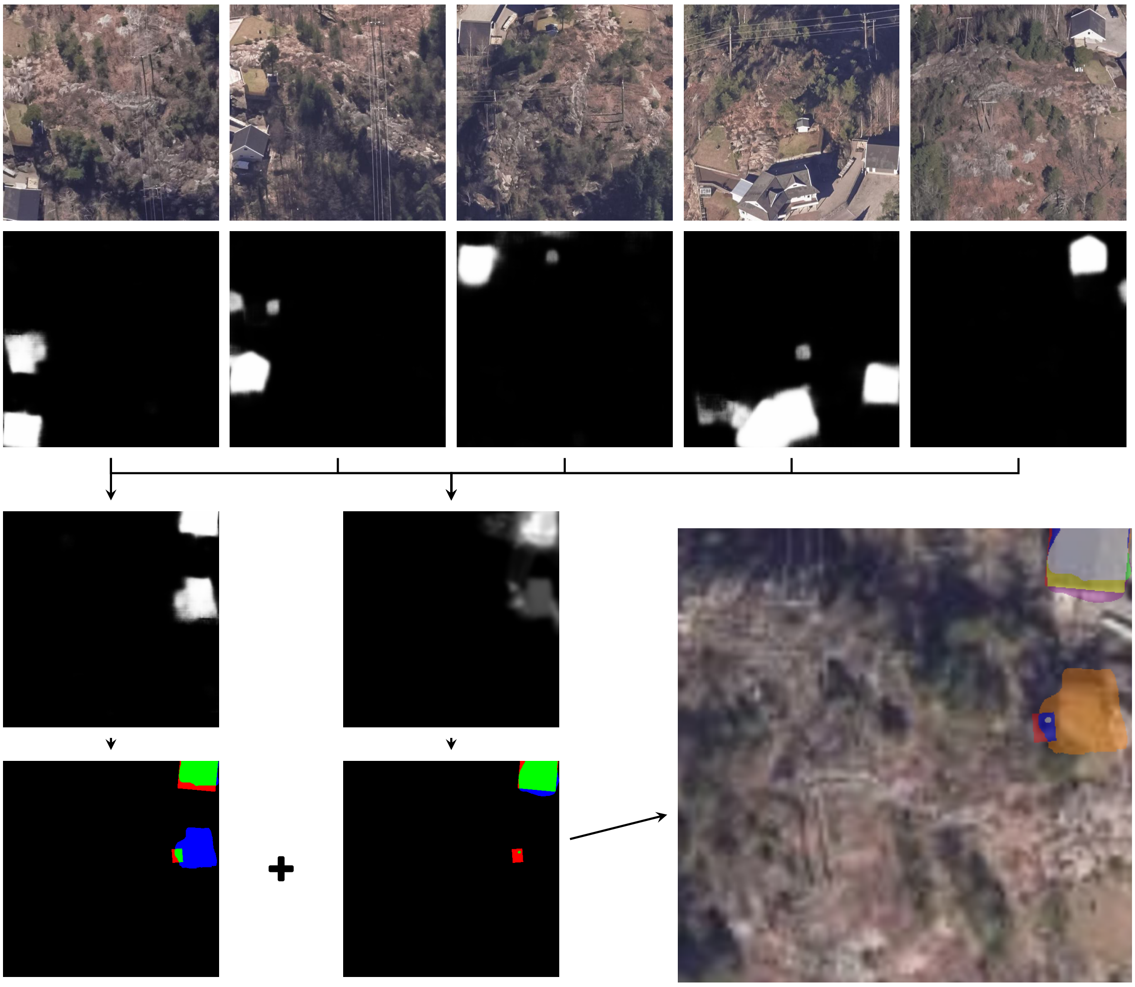


Figure 25: A lawn misclassified as building on nadir image.

The next and final example will address one of the major challenges associated with oblique aerial images: occlusion. In figure 26, there is a building surrounded by tall trees, visible only from one out of five perspectives. On the nadir image it is hidden in the shadows, while on the other oblique aerial images it is occluded by trees. The segmentation model is highly confident in detecting the building from the one angle.

However, even if the model is 100% certain that it is a building, the confidence score from this one image will not bring the total score above the threshold. This issue here is that the five observations are equally weighted. As humans, we possess the ability to recognize that the feature visible in this single image is not present in the others. Consequently, we would base our decision solely on this specific image. If this behavior could be replicated it would significantly increase the reliability of the building detection from oblique aerial images. One possible approach could be to train a segmentation model to classify pixels representing trees or very dark shadows and avoid considering these observations when combining the results. Another approach could be to train a model to process multiple images as inputs and ascertain whether they depict the same object observed from distinct perspectives.

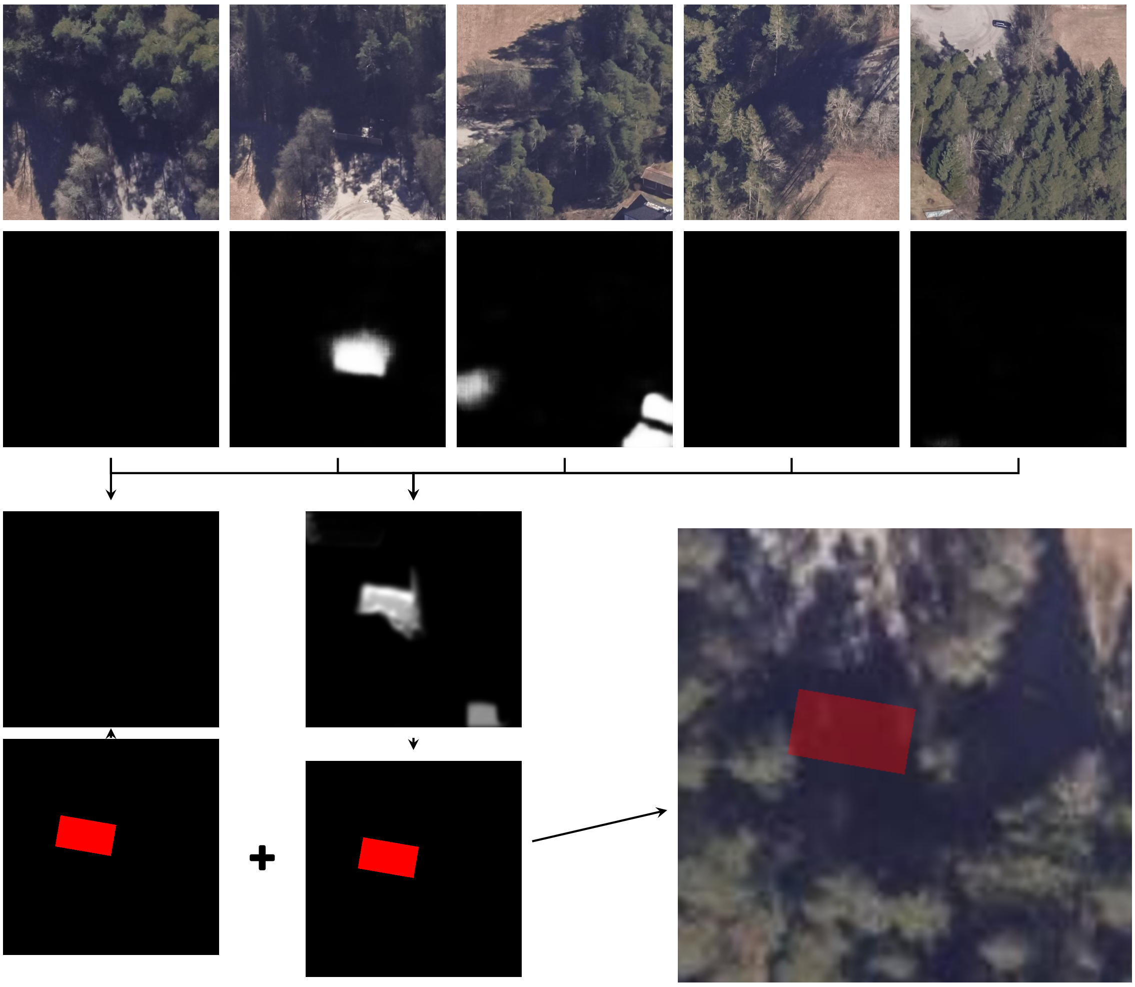


Figure 26: An occluded building visible only in one image.

6 Conclusion and further research

In this research, a proposed method for building detection using oblique aerial images was presented and evaluated. The method combined predictions from multiple images captured from different angles to achieve more accurate and reliable building detections. The approach was compared to traditional building detection using nadir images alone. The experimental study provided valuable insights into the strengths and limitations of the proposed method.

When evaluating the segmentation models performance on single images, the performance on nadir images was slightly better than oblique images in terms of Intersection over Union (IoU). However, the visual inspection of the segmentation results showed that the oblique images provided more distinguishing features, and demonstrated more reliable predictions, especially on objects with few features from a top-down perspective, such as terraces, trucks and simple building structures.

The proposed method of analyzing an area by combining segmentation results from multiple images proved to result in accurate building footprints. In addition, the number of false positives was drastically reduced compared to the nadir result. The result also revealed that occlusion remains a significant challenge and some improvements are needed to handle such situations effectively.

The study also revealed the importance of training data quality and representation. The segmentation model's performance was highly dependent on the similarity between the training data and the target dataset. Training the model on images with a resolution similar to the target data improved its performance significantly. Additionally, the research demonstrated some challenges of detecting small buildings. Some of the explanation for this is assumed to be that many small buildings are not included in official registries and consequently not present in the training data.

The study opens several avenues for further research to improve the suggested approach and general advancements in the field of building detection using oblique aerial images:

- **Weighting of Observations:** Explore methods to detect occluded objects in the images and weight or exclude observations accordingly. This approach aims to improve the reliability of predictions by only considering relevant information.

- **Refine Dataset:** The training data could be further refined by leveraging the trained model to identify potential outliers or mislabeled samples in the dataset. By using the model to predict building detections on the training data, instances with low confidence scores or instances that are consistently misclassified can be flagged as potential outliers. These flagged samples can then be manually inspected and either corrected or removed from the training dataset. This iterative process of model-assisted data cleaning can help improve the overall quality and accuracy of the training data, leading to better performance of the building detection model.

- **Robustness Across Datasets:** Exploring methods to improve the model's robustness and generalization across diverse datasets could lead to more effective building detection in various geographical regions and environments.

- **Incorporating Additional Data Sources:** Leveraging other data sources, such as lidar data or semantic maps, in combination with oblique aerial images could enhance building detection accuracy and provide more comprehensive information.

- **Two-step approach:** A two-step approach, where an initial more simple analysis can locate potential building candidates. Subsequently, additional data sources such as lidar data and semantic maps, combined with higher-resolution imagery, can be utilized for more accurate and detailed analysis to confirm or deny these suggestions and further refine the building footprints. This two-step process helps optimize computational resources and focuses efforts on relevant areas.

In conclusion, this thesis proposed a comprehensive method for building detection using oblique aerial images, demonstrating the benefits of multi-perspective analysis for improved accuracy and reliability. While promising results were achieved, addressing challenges like occlusions and enhancing model robustness across datasets remains essential.

References

Carion, Nicolas et al. (May 2020). *End-to-End Object Detection with Transformers.* DOI: 10.48550/arXiv.2005.12872. arXiv: 2005.12872 [cs]. (Visited on 5th July 2023).

Li-Chee-Ming, Julien and Costas Armenakis (May 2012). 'Fusion of Optical and Terrestrial Laser Scanner Data'. In: 38.

Chen, Hao, Zipeng Qi and Zhenwei Shi (2022). 'Remote Sensing Image Change Detection with Transformers'. In: *IEEE Transactions on Geoscience and Remote Sensing* 60, pp. 1–14. ISSN: 0196-2892, 1558-0644. DOI: 10.1109/TGRS.2021.3095166. arXiv: 2103.00208 [cs]. (Visited on 5th July 2023).

Chen, Keyan, Zhengxia Zou and Zhenwei Shi (2021). 'Building Extraction from Remote Sensing Images with Sparse Token Transformers'. In: *Remote Sensing* 13.21. ISSN: 2072-4292. DOI: 10.3390/rs13214441.

Dillon, Ryan J. (2017). *Sosi.*

Frommholz, D., M. Linkiewicz and A. M. Poznanska (June 2016). 'Inlining 3d Reconstruction, Multi-Source Texture Mapping and Semantic Analysis Using Oblique Aerial Imagery'. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 41B3, pp. 605–612. ISSN: 2194-9034 The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. DOI: 10.5194/isprs-archives-XLI-B3-605-2016. (Visited on 11th July 2023).

Gillies, Sean et al. (Dec. 2022). *Shapely.* DOI: 10.5281/zenodo.5597138.

Ji, Shunping, Shiqing Wei and Meng Lu (Jan. 2019). 'Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set'. In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1, pp. 574–586. ISSN: 1558-0644. DOI: 10.1109/TGRS.2018.2858817.

Kass, Michael, Andrew Witkin and Demetri Terzopoulos (Jan. 1988). 'Snakes: Active Contour Models'. In: *International Journal of Computer Vision* 1.4, pp. 321–331. ISSN: 1573-1405. DOI: 10.1007/BF00133570. (Visited on 4th July 2023).

Kato, Zoltan (2012). *Markov Random Fields in Image Segmentation.*

Long, Jonathan, Evan Shelhamer and Trevor Darrell (Mar. 2015). *Fully Convolutional Networks for Semantic Segmentation.* DOI: 10.48550/arXiv.1411.4038. arXiv: 1411.4038 [cs]. (Visited on 4th July 2023).

Maggiori, Emmanuel et al. (2017). 'Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark'. In: *2017 IEEE Inter-*

national Geoscience and Remote Sensing Symposium (IGARSS), pp. 3226–3229. DOI: 10.1109/IGARSS.2017.8127684.

Nex, F., E. Rupnik and F. Remondino (Oct. 2013). 'Building Footprints Extraction from Oblique Imagery'. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W3, pp. 61–66. ISSN: 2194-9050. DOI: 10.5194/isprsannals-II-3-W3-61-2013. (Visited on 11th June 2023).

Ronneberger, Olaf, Philipp Fischer and Thomas Brox (May 2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation.* DOI: 10.48550/arXiv.1505.04597. arXiv: 1505.04597 [cs]. (Visited on 5th July 2023).

Serra, J. (1982). *Image Analysis and Mathematical Morphology.* Image Analysis and Mathematical Morphology. Academic Press. ISBN: 978-0-12-637241-0.

Verykokou, Styliani and Charalabos Ioannidis (2018a). 'Oblique Aerial Images: A Review Focusing on Georeferencing Procedures'. In: *International Journal of Remote Sensing* 39.11, pp. 3452–3496. DOI: 10.1080/01431161.2018.1444294. eprint: https://doi.org/10.1080/01431161.2018.1444294.

— (June 2018b). 'Oblique Aerial Images: A Review Focusing on Georeferencing Procedures'. In: *International Journal of Remote Sensing* 39.11, pp. 3452–3496. ISSN: 0143-1161. DOI: 10.1080/01431161.2018.1444294. (Visited on 18th June 2023).

Wei, Yanfeng, Zhongming Zhao and Jianghong Song (2004). 'Urban Building Extraction from High-Resolution Satellite Panchromatic Image Using Clustering and Edge Detection'. In: *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium.* Vol. 3, 2008–2010 vol.3. DOI: 10.1109/IGARSS.2004.1370742.

Yang, Hui et al. (Nov. 2018). 'Building Extraction in Very High Resolution Imagery by Dense-Attention Networks'. In: *Remote Sensing* 10.11, p. 1768. ISSN: 2072-4292. DOI: 10.3390/rs10111768. (Visited on 4th July 2023).

Zhu, Xizhou et al. (Mar. 2021). *Deformable DETR: Deformable Transformers for End-to-End Object Detection.* DOI: 10.48550/arXiv.2010.04159. arXiv: 2010.04159 [cs]. (Visited on 5th July 2023).