Automatic LoD-2 Building Reconstruction with Building Boundary Constraints from Satellite-Derived Data

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Timucin Bulmus

Graduate Program in Civil Engineering

The Ohio State University

2023

Thesis Committee

Professor Rongjun Qin, Advisor

Professor Charles Toth

Copyrighted by

Timucin Bulmus

2023

Abstract

Urban areas present diverse architectural designs, posing challenges for semantic segmentation and reconstruction tasks. The complexity arises from the presence of multiple peaks, slopes, and variations in roof structures, leading to potential misclassification and incomplete representations. The densely packed nature of city centers further exacerbates the problem, causing occlusions and interference between adjacent structures, making accurate isolation of individual buildings and their sections more difficult. This study endeavors to address these challenges by employing state-of-the-art techniques for segmenting complex and nearby building borders. The segmentation task utilizes the High Resolution Network (HRNet) architecture on the combination of building ground truth mask and satellite derived orthophoto, and manually generated borders. These borders are instrumental in separating the building prediction mask to achieve heightened accuracy in building extraction. The study culminates in 3D building model reconstruction through a model-driven approach, enhancing the representation and understanding of complex urban structures.

Acknowledgments

Words cannot express my gratitude to Professor Rongjun Qin for his invaluable support and feedback. Without his support, I would not have been able to embark on this journey. Additionally, I would like to express my deep appreciation to Professor Charles Toth, a member of my defense committee, whose thoughts and feedback have been immensely significant for my research.

I am grateful to my lab colleagues, Shengxi Gui, Guixiang Zhang, and Ningli Xu, for their constant support and companionship throughout my research.

I want to send a heartfelt thanks to the Ministry of National Education of the Republic of Turkey and the General Directorate of Land Registry and Cadastre for granting me the incredible opportunity to pursue my education abroad. Words can't express how honored and proud I feel to have received such a prestigious scholarship. It's truly a dream come true, and I am overflowing with gratitude.

Lastly, I want to extend a special thank you to my family for their continuous love, encouragement, and unwavering support. Their presence and belief in me have been instrumental in my life, and I am forever grateful for their sacrifices and dedication.

Vita

B.S. Geomatics Engineering, Erciyes University	January 2018
M.S. Civil Engineering, The Ohio State University	July 2023

Fields of Study

Major Field: Civil Engineering

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Statement of Purpose

In remote sensing and computer vision applications, building segmentation is essential, especially for identifying and analyzing complex building structures. Accurate segmentation of buildings into their component sections is now crucial for activities like urban planning, architectural design, and historic preservation due to the growing population and urban congestion. Deep learning techniques have become effective building segmentation tools in recent years, enabling exact recognition of intricate structural components.

Building segmentation from aerial or satellite photos has been shown to be an invaluable tool for deep learning methods like convolutional neural networks (CNNs). In order to identify patterns and features that separate buildings from their surroundings, these algorithms are trained using labeled data. Deep learning algorithms can be used to identify complex building components and separate them into their component parts with a high degree of accuracy.

The problem statement revolves with identification of complex building components. To identify and distinguish different sections inside the complex building, a deep learning method has been applied on the generated building borders which is a useful information to detect those pixels on the border. Building components can be distinguished with the help of borders by using component analysis tools such as connected component analysis or watershed algorithm.

Building model reconstruction is another significant problem in the field of remote sensing and computer vision. Separating the different parts of a structure may be done using satellite pictures and trained boundary models. Building model production is made possible by the incorporation of Digital Surface Models (DSMs), which improves the reconstruction procedure further. In that case, buildings can be modeled more precisely.

The combination of deep learning-based identification and separation of complex building parts, along with the integration of DSMs, offers a comprehensive approach to building model reconstruction. This method makes use of DSMs for exact alignment and reconstruction, component analysis methods for part separation, and deep learning algorithms for constructing segmentation. The rebuilt building models help with urban planning, architectural design, and numerous geospatial applications by offering useful insights into the precise composition, spatial arrangement, and architectural aspects of complex structures.

1.2. Related Work

An early segmentation method of graph-based segmentation on the edge weights and spatial coherence entails creating a graph representation of the picture. A technique optimizes the graph and achieves precise picture segmentation using algorithms like Graph Cut. The research methodology entails specifying a border measurement predicate, creating a productive segmentation algorithm, and exhibiting global attributes. A variety of local neighborhoods are employed to identify boundaries effectively. In order to segment images meaningfully, the method combines graph representation, effective algorithms, and consideration of nearby neighborhoods (Felzenszwalb & Huttenlocher, 2004). Watershed segmentation is another early work that simulates floods to divide areas based on local minima and considers pixel intensities as a topographic surface. It may be used to a variety of picture formats, including grayscale or color, and is frequently used to define objects or borders in an image based on intensity gradients. Improvements have been made in order to address the shortcomings of the watershed transform in medical image analysis. One enhancement is adding past data through a previous probability calculation, allowing the system to use more information for better segmentation outcomes. Another enhancement is to combine the watershed transform with other approaches, such as atlas registration using markers, to increase segmentation accuracy by combining the advantages of the watershed algorithm (Grau et al., 2004).

CNN-based architectures have demonstrated to be quite good at identifying spatial relationships and picking up useful representations for pixel-level classification. Fully Convolutional Networks (FCN) (Shelhamer et al., 2017) which introduced the idea of completely convolutional layers for end-to-end pixel-level classification. The FCN model is utilized in a study to evaluate satellite data in conjunction with a Digital Surface Model (DSM) to conduct class segmentation. Label is divided into 2 classes: buildings and building borders. The addition of height information from the DSM improves segmentation outcomes (Schuegraf et al., 2022). Another study introduces a novel methodology that employs the SkipFuse-U-Net-3+ architecture for the partitioning of architectural structures into segments characterized by geometric and spectral homogeneity. The approach entails the prediction of individual building pixels and separation lines, followed by the conversion of semantic outcomes into distinct instances through the application of the watershed transform. The model is trained using pixel-level and topology-conscious loss functions on satellite imagery and Digital

Surface Models (DSMs). Empirical findings highlight the effective adaptability of the proposed method to diverse geographical regions, surpassing established techniques in the production of well-defined building segments characterized by crisp boundaries (Schuegraf et al., 2023). U-Net (Ronneberger et al., 2015) uses skip connections to collect both local and global contextual information while DeepLab (Chen et al., 2018) captures multi-scale characteristics using dilated convolutions. Another study introduced InternImage which utilizes deformable convolutions, allowing it to have a large effective receptive field for tasks such as detection and segmentation, while also adapting spatial aggregation based on input and task information. This reduces the strict inductive bias of traditional CNNs, enabling the model to learn stronger and more robust patterns with large-scale parameters from massive datasets, similar to the gains seen in ViTs. The results demonstrate the effectiveness and potential of large-scale CNN-based models in computer vision tasks (Wang et al., 2022).

In computer vision research, using transformers for semantic segmentation has gained traction. The Vision Transformer (ViT) (Dosovitskiy et al., 2020) is a remarkable transformer-based model that adapts the transformer architecture for image tasks and achieves competitive performance in capturing global context. Another technique is TransUNet (Chen et al., 2021) which blends convolutional neural networks with transformers to obtain accurate pixel-level predictions. Furthermore, Swin Transformer (Liu et al., 2021) has demonstrated outstanding performance in a variety of computer vision applications, including semantic segmentation, by quickly capturing long-range dependencies and processing high-resolution inputs. These transformer-based models show the power of using self-attention processes for complex semantic segmentation tasks. Additionally, UNetFormer uses a Transformer-based decoder. UNetFormer leverages the advantages of both UNet architecture and Transformer in a distinctive manner to achieve highly efficient segmentation. To ensure computational effectiveness, the encoder is implemented using the lightweight ResNet18 model. Moreover, the researchers develop a novel global-local attention mechanism for the decoder, which enables effective modeling of both global and local information (Wang et al., 2022).

Outline extraction, in addition to image segmentation, plays a critical role in model reconstruction. The retrieved building outlines serve as a fundamental framework for future model reconstruction. They provide the basis for creating accurate 3D models and comprehensive architectural blueprints. These outlines are invaluable in understanding the spatial arrangement of buildings, identifying distinct building components, and documenting essential architectural aspects. A study is performed using normal vector estimation to comprehend the local orientation and geometry of the surface on a DSM derived from a satellite picture and filters the vegetation area. Then, based on the height difference, it determines the orientation of the roof faces and fits a line along the edges (Nex & Remondino, 2012). Another research suggests using a U-Net architecture-based semantic segmentation algorithm to extract building footprints from high-resolution multispectral satellite pictures and GIS databases like OpenStreetMap (Li et al., 2019). Another work focuses on transforming building masks into boundary lines and then changing their orientation using orthophoto-derived line segments. The postprocessing stage entails recognizing shared characteristics between polygon surrounds and OpenStreetMap data in order to improve the polygons, with the assumption that building polygons align with the direction of the buildings, supported by OpenStreetMap road vectors (Gui & Qin, 2021).

The diverse and unique nature of building outlines poses significant challenges for training a model to accurately identify ground truth polygons. The variability in building shapes makes it impractical to achieve a one-size-fits-all approach, necessitating alternative methodologies to address this complexity. Therefore, a study proposes a novel method to simplify building footprints in topographic map generalization from large to medium scales. This method formulates the simplification problem as a joint task, combining node removal classification and node movement regression. To accomplish this, the study introduces a multi-task graph convolutional neural network model (MT_GCNN) that can effectively learn and address both node removal and movement tasks simultaneously. The ultimate goal is to improve the efficiency and accuracy of building footprint simplification, thereby enhancing the quality of topographic map generalization at medium scales (Zhou et al., 2023).

In the process of building model reconstruction, researchers frequently depend on point cloud data or Digital Surface Models (DSMs) to produce exact representations of structures. A study uses robust estimation approach and Support Vector Machine to generate the best roof models, assuring accuracy and a model-driven approach (Henn et al., 2013). Another study focuses on semiautomatic building model generation from vector base maps and format aerial imagery in a large scale (Buyukdemircioglu et al., 2018). Another study performs a hybrid technique for reconstructing 3D building models using WorldView-2 satellite data. Mask refining, building outline extraction, decomposition, and roof type categorization are among the techniques used in the method, which blends data-driven and model-driven techniques. A roof type library with parameter initialization is used. The study adopts a discrete search space and alters the optimization approach to identify the most dependable 3D model. In addition, it explores the reconstruction of linking roofs and the interaction between nearby roof models (Partovi et al., 2019). Furthermore, a study introduces a novel RANSAC-MPR framework for reconstructing buildings from point clouds. The framework uses the RANSAC paradigm to iteratively estimate parameters of building primitives from planar patches. It enhances primitive selection by utilizing nearest planar sampling, increasing the likelihood of successful primitive identification. The approach simultaneously determines all parameters of a segmented building primitive and employs a non-learning score function for primitive selection. Overall, the RANSAC-MPR framework offers an efficient and accurate method for building reconstruction from point clouds (Li & Shan, 2022).

1.3. Thesis Structure

Chapter 2 of this study outlines the processes involved, beginning with the generation of building footprints and borders. It further discusses the methodology for separating building sections through line extraction.

In Chapter 3, the experimental findings are provided, along with descriptions of the datasets that were employed. The chapter also discusses border creation, training, and evaluating state-of-the-art models for both classes of building masks and borders. In addition, the technique of creating the building mask is presented, followed by line vectorization on the expected borders. Finally, the chapter concludes with an experiment on building model reconstruction and evaluation of experiment outcomes.

Chapter 4 concludes the experimentation with a comprehensive study overview and a concise summary of the findings. Furthermore, it provides a thorough examination of the limitations encountered during the research and outlines potential avenues for future investigations to enhance the study's outcomes.

Chapter 2. Methodology

The building reconstruction workflow in this study encompasses several interconnected stages, aiming to generate accurate and realistic 3D models of buildings from satellite images of urban areas. The process commences with data preparation, where suitable datasets containing high-resolution images of urban regions are collected and preprocessed for analysis.

The workflow for building reconstruction in this study involves data preparation, where high-resolution urban images are collected and preprocessed. Building mask and border segmentation models are trained using HRNet to predict building masks and borders for the provided testing dataset. The predicted border results are then vectorized to refine borders and extract individual building polygons. Further geometric accuracy is achieved through Graph-Cut Labelling for building rectangle orientation refinement. The process moves to 3D modeling, where decomposed building rectangles are transformed into 3D representations. Consistency in 3D building types is enforced, and a novel approach for recovering complex 3D buildings through model-level merging is explored. This comprehensive workflow aims to generate accurate and detailed 3D building reconstructions from the input images.

This chapter is divided into three sections that go through the various steps of the suggested technique. The first section employs deep learning approaches to generate a building footprint and border segmentation model. This entails training a model to recognize and define building footprints and borders based on input data such as aerial or satellite images.

The second section deals with the problem of dividing building components. Following the prediction of border pixels from the previous stage, the study applies line fitting algorithms to refine and divide the different parts inside the buildings. The study seeks to properly discern and designate specific components or portions of the buildings by fitting lines to the predicted border pixels.

The third section in this chapter focuses on fitting a Level of Detail 2 (LoD-2) model to the extracted building components. LoD-2 models are more realistic reconstructions of the structures, accurately reflecting their geometrical and architectural aspects. The project intends to recreate those buildings in a more exact and detailed manner by fitting a LoD-2 model to the segmented building components.

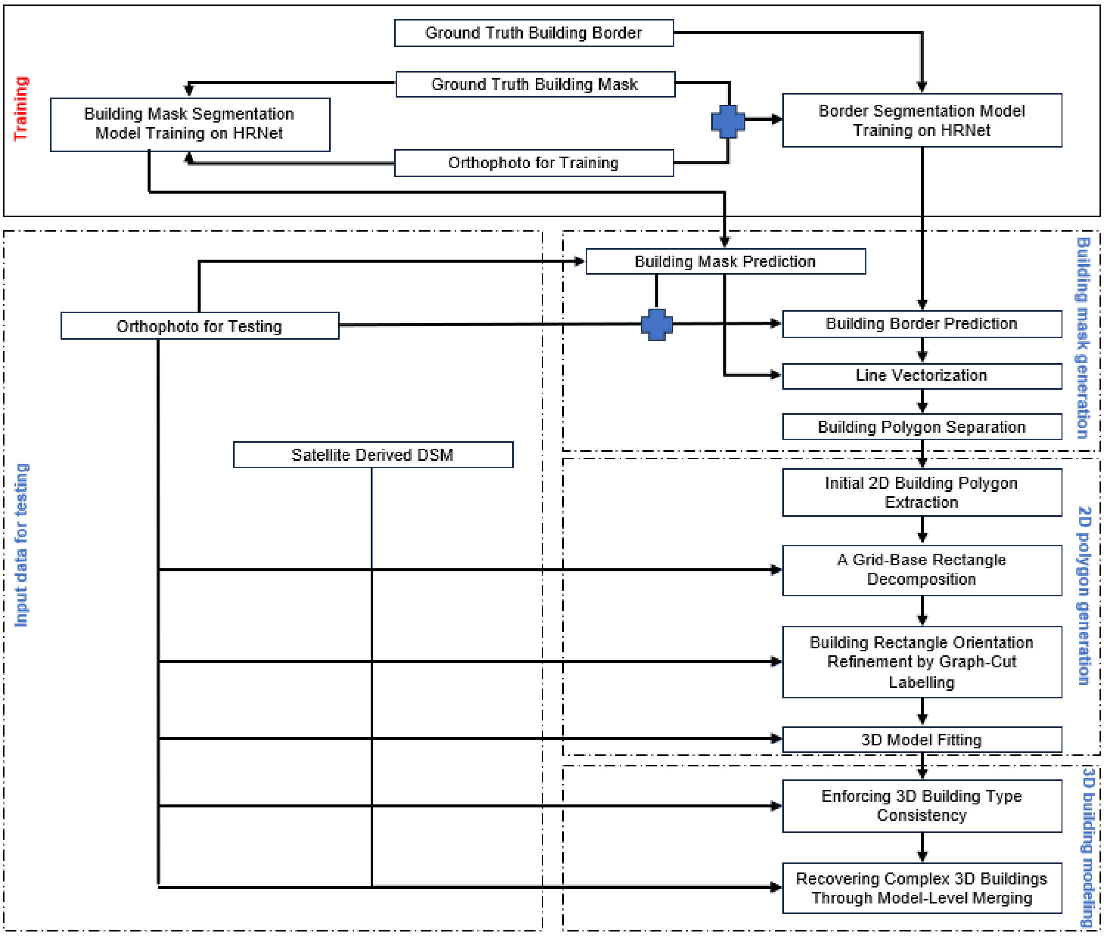

Training
Ground Truth Building Border
Ground Truth Building Mask
Building Mask Segmentation Model Training on HRNet
Border Segmentation Model Training on HRNet
Orthophoto for Training
Building Mask Prediction
Orthophoto for Testing
Building Border Prediction
Line Vectorization
Building Polygon Separation
Building mask generation
Satellite Derived DSM
Initial 2D Building Polygon Extraction
A Grid-Base Rectangle Decomposition
Building Rectangle Orientation Refinement by Graph-Cut Labelling
3D Model Fitting
2D polygon generation
Input data for testing
Enforcing 3D Building Type Consistency
Recovering Complex 3D Buildings Through Model-Level Merging
3D building modeling

Figure 1: Workflow starting from model training for building mask and border segmentation and resulting in 3D building model reconstruction.

Beginning with the building model training of building footprints and borders using deep learning, this chapter offers a thorough description of the suggested technique. Following that, it moves on to separating building polygons using line vectorization on the predicted border pixels. The fitting of LoD-2 models to the specified building components brings the chapter to a close, allowing for more accurate and thorough building reconstructions.

2.1. Building Footprint and Border Segmentation

Image segmentation enables the identification and separation of various objects and structures within a scene. In the context of building segmentation, it helps to differentiate buildings from other elements like roads, vegetation, and water bodies, which may coexist in the urban landscape. Building segmentation aids in automatic feature extraction, enabling rapid and efficient mapping of urban areas from satellite or aerial imagery. These maps serve as critical inputs for building model reconstruction, urban planning, and city management.

Classification tasks may result efficiently with lower resolution in some circumstances since they emphasize overall output without taking the exact location of information into account. However, for more complex applications like image segmentation, when pixels need positioning, a higher resolution technique is required to extract detailed segments successfully. Since HRNet architecture maintains high resolution layer from start to end, it enables extract properties. Therefore, this section presents model training on HRNet architecture for building mask and building border.

A cutting-edge network architecture known as High-Resolution Network (HRNet) (Sun et al., 2019) is used in this study to examine the extraction of building polygons and borders for semantic segmentation task. HRNet introduces a multi-resolution method that maintains high-resolution representations throughout the network to solve the weaknesses of conventional deep convolutional neural networks (CNNs). Because of its distinctive architecture, HRNet is particularly useful for jobs that call for exact localization and segmentation, including detecting complex architectural structures. It can capture both fine-grained features and global context.

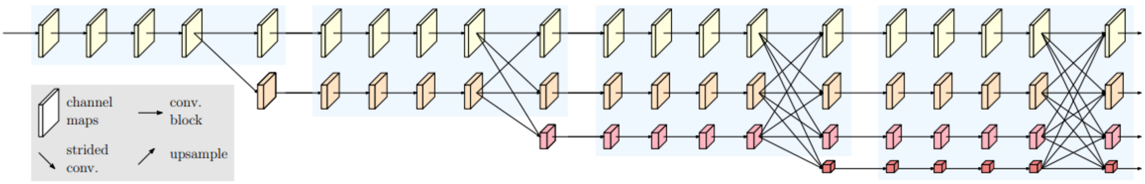


Figure 2: A visual representation for HRNet architecture (Sun, Zhao, et al., 2019)

HRNet design consists of four phases, the last three of which are composed of up of modularized multi-resolution blocks. Both a multi-resolution group convolution and a multi-resolution convolution are included in these blocks. In order to accommodate for various spatial resolutions, the multi-resolution group convolution separates input channels into subgroups and executes individual convolutions on each subset. This makes it possible to extract features at various resolutions. The multi-resolution convolution component facilitates the integration of information across resolutions by combining the output features from several branches. HRNet is useful for complicated building structures due to its capacity to retain high-resolution representations across the network, which maintains spatial information and improves localization and segmentation accuracy.

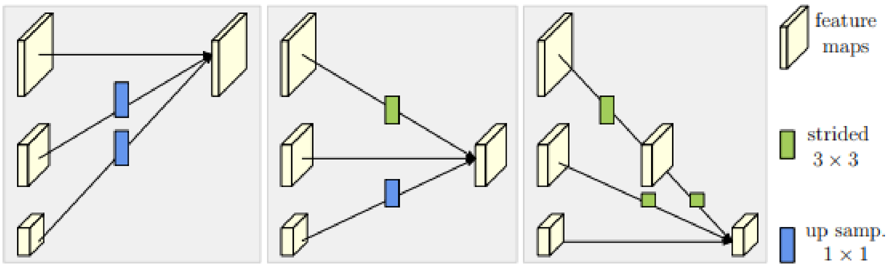


Figure 3: Fusion representation of multi-dimensional layers (Sun, Xiao, et al.,2019).

Each pixel in the lower-resolution picture is mapped to a block of pixels in the higher-resolution image during nearest neighbor upsampling. Upsampling maintains these crucial qualities while downsampling aids in identifying important elements within the global environment. HRNet achieves a balance between obtaining crucial global information and keeping high-resolution features for accurate segmentation. This is done by utilizing downsampling and upsampling.

HRNet design uses 2-strided 3x3 convolutions to reduce the dimensionality of input layer. This setup downscales and reduces the dimension of the model by moving the convolutional kernel across the input feature map every two pixels. To reduce dimensionality of high resolution by 4 times, 3x3 convolution with 2 stride is applied twice. The higher-level features from the input feature map can be captured by this downsampling procedure. The model uses a basic closest neighbor sampling strategy followed by a 1x1 convolution operation to accomplish upsampling. This procedure assists in matching the number of channels in the upsampled feature map to the desired output.

In classification problems, the cross-entropy function is a popular loss function to assess the difference between expected probability and actual class labels. The model is encouraged to reduce the discrepancy between the anticipated probability and the real probabilities contained by the class labels by quantifying the average information or uncertainty in the predictions.

In HRNet architecture, Stochastic Gradient Descent (SGD) is used as an optimization technique that trains incrementally changing its parameters. With regard to the network's parameters, it calculates the gradients of the loss function and updates them in a way that minimizes loss. SGD estimates the gradients and modifies the parameters in accordance using a subset of training data (mini-batch) in each iteration.

In conclusion, HRNet is applied in this research to extract building masks and borders and segment images. HRNet improves localization and segmentation accuracy while capturing detailed information by keeping high-resolution layers throughout the network. Extraction of fine-grained features and global context is possible in HRNet thanks to the multi-resolution blocks, downsampling, and upsampling procedures. This all-encompassing strategy helps with precise segmentation of complex building components and advances picture segmentation tasks.

2.2. Border Vectorization as Line Segment

Model reconstruction solely based on building mask prediction often yields suboptimal performance in model fitting due to the possibility of considering multiple buildings as a single entity. In the process of building mask prediction, the model may result in low performance to distinguish individual buildings, resulting in a merged representation of multiple structures. Consequently, when fitting a model on such merged buildings with diverse features, the resulting model does not accurately reflect reality, as buildings possess distinct attributes like height, shape, and roof type. This limitation underscores the necessity for building separation before reconstruction.

The incorporation of building borders helps address this challenge, as they facilitate the division of distinct buildings into multiple polygons. By leveraging the information from borders, the individual buildings can be correctly delineated, allowing for more accurate and realistic model reconstructions. The separation of buildings into their respective polygons ensures that each building's unique features are properly represented in the reconstruction process, leading to more faithful and reliable model fittings. This integration of border information plays a vital role in improving the overall performance of model reconstruction and enhances the quality of the final 3D models.

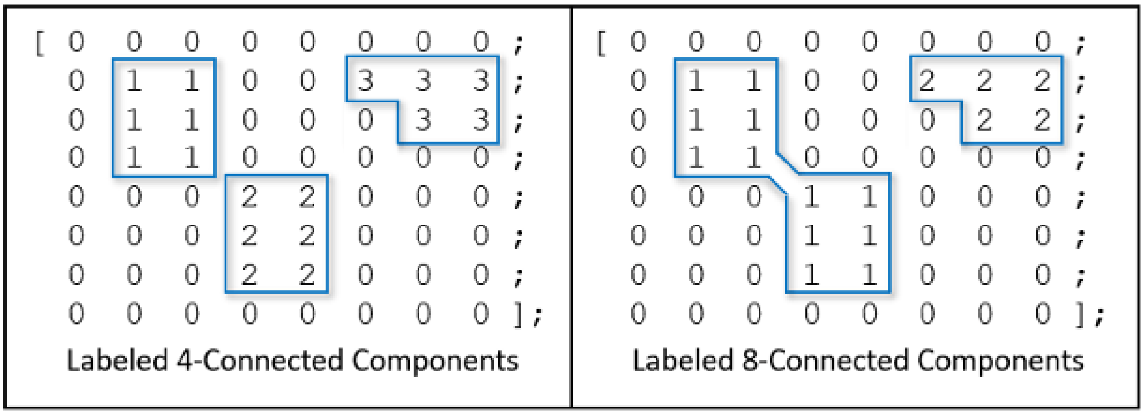
This section demonstrates how building and border prediction results are used to separate distinct buildings and their components. By combining both predictions, the method achieves accurate segmentation of individual buildings, enhancing the realism of the final 3D models. This approach ensures that each building's unique attributes are correctly represented, contributing to advancements in urban modeling and computer vision research.

To identify building polygons and border for each separately, connected component analysis is applied and building polygons from building mask prediction result are separated with the use of border prediction result.

Connected component analysis is used to separate the boundaries and constructing polygons. Identifying and labeling unique regions or components within an image is accomplished via connected component analysis, a fundamental technique used in image processing and computer vision. It is a crucial step in several processes, including object recognition, segmentation, and pattern recognition. Tracing the boundaries of linked pixels or areas with common characteristics like color or intensity is known as connected component analysis.

A starting point is chosen inside the picture to start the procedure. A tracing approach is then used to establish the outlines of the related components from this point. Contour tracing is a popular technique that repeatedly tracks a region's border by locating the nearby pixels or points that make up the contour. This is accomplished by employing algorithms like the Moore-Neighbor Tracing or the Freeman Chain Code (Chang et al., 2004).

When contour tracing, the algorithm looks at the nearby pixels in a certain sequence in an effort to find the following contour point. It is common to use the terms "4 connectivity" (horizontal and vertical) or "8 connectivity" (including diagonal pixels) to describe the connectedness of adjacent pixels. The algorithm keeps following the contours until it goes back to the beginning, signifying the end of a linked component.



```
[ 0   0   0   0   0   0   0   0 ;
  0   1   1   0   0   3   3   3 ;
  0   1   1   0   0   0   3   3 ;
  0   1   1   0   0   0   0   0 ;
  0   0   0   2   2   0   0   0 ;
  0   0   0   2   2   0   0   0 ;
  0   0   0   2   2   0   0   0 ;
  0   0   0   0   0   0   0   0 ];
```

Labeled 4-Connected Components

```
[ 0   0   0   0   0   0   0   0 ;
  0   1   1   0   0   2   2   2 ;
  0   1   1   0   0   0   2   2 ;
  0   1   1   0   0   0   0   0 ;
  0   0   0   1   1   0   0   0 ;
  0   0   0   1   1   0   0   0 ;
  0   0   0   1   1   0   0   0 ;
  0   0   0   0   0   0   0   0 ];
```

Labeled 8-Connected Components

Figure 4: Visualization for 4-pixel connectivity on the left and 8-pixel connectivity on the right (*Label and Measure Connected Components in a Binary Image - MATLAB & Simulink*, n.d.).

Each pixel or point has a label indicating that it belongs to a certain connected component. This labeling makes it possible to analyze and process the recognized components.

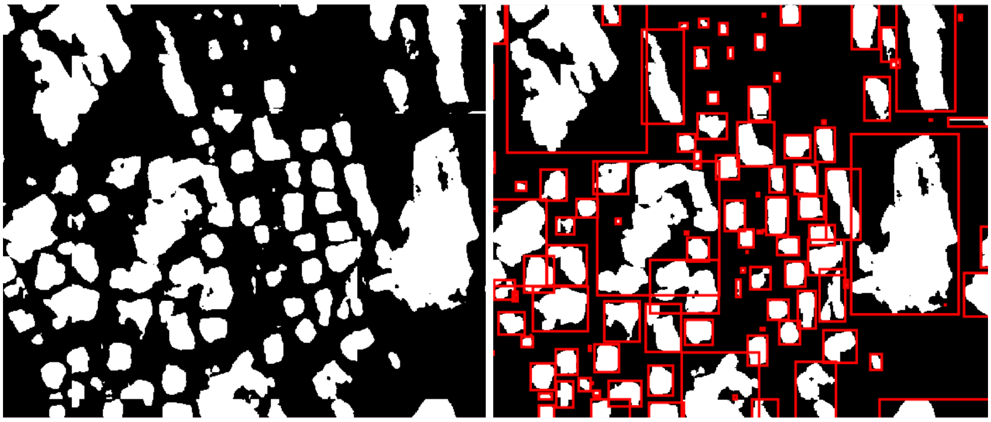


Figure 5: Building mask prediction on the left and building components in bounding boxes on the right.

In each recognized component, predicted border pixels detected and line vectorized with RANSAC line fitting algorithm. RANSAC algorithm is a reliable model fitting technique to successfully handle a certain issue or task compared to other methods such as Hough Transform. Even in the presence of outliers or noise that might negatively impact the accuracy of the findings, the algorithm's main objective is to estimate the ideal parameters of a mathematical model that match the presented data. It was first proposed in the publication Fischler and Bolles (Fischler & Bolles, 1981). Researchers can use RANSAC to get accurate parameter estimates that fully reflect the underlying structures or patterns in the data while reducing the impact of inconsistent or false observations. RANSAC is useful in many different domains, including computer vision, image analysis, and pattern recognition, where the ability to effectively model the data in the face of possible outliers or noise is essential for producing accurate and trustworthy results.

RANSAC is very helpful when working with datasets that could have inaccurate or inconsistent data points. It tackles the problem of outliers by fitting models repeatedly to randomly chosen subsets of the data, or "inliers," which are referred to as the "inliers." RANSAC can determine the most dependable model that most accurately captures the underlying structure of the data by sampling and fitting models iteratively.

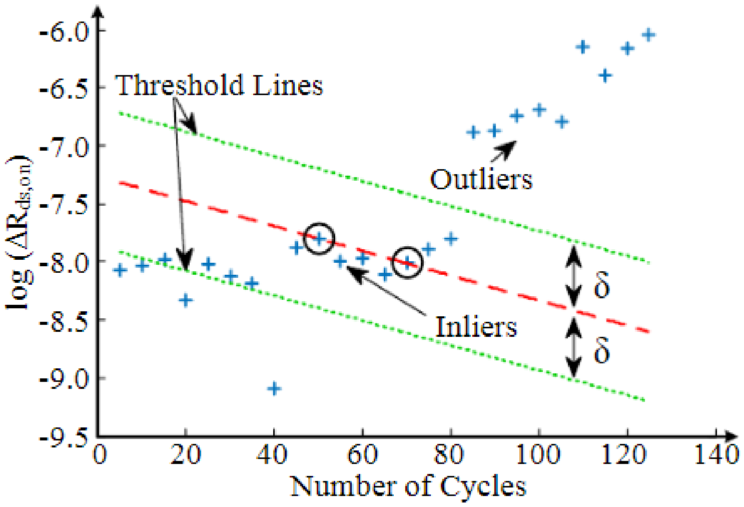


Figure 6: Optimal line of RANSAC algorithm is in red dashed line and threshold lines in green dot lines where points are in blue plus sign (Dusmez et al., 2017).

The capacity of RANSAC to handle datasets with a lot of noise or outliers is one of its strongest points. It successfully removes these outliers from the model estimate process, so they do not unreasonably affect the outcome. RANSAC is a good fit for situations where precise and reliable results are necessary because of its resilience.

In general, the use of connected component analysis for identifying building polygons and borders and line fitting using the RANSAC algorithm allows for the precise separation and characterization of building sections during the reconstruction of a building model.

2.3. 3D Building Model Reconstruction

Building model reconstruction in 3-dimensional is one of the challenging tasks in photogrammetry. Extraction of the building outline is required for 3D model reconstruction, and it is essential for the purpose of constructing comprehensive and realistic 3D representations of urban areas. In order to produce precise and realistic 3D building models, model fitting is crucial because it ensures geometric and semantic coherence, topological consistency, and high geometric accuracy. It makes it possible to produce CityGML LoD2 models and rebuild intricate roofs, offering accurate and thorough representations of buildings (Zhang et al., 2021).

Obtaining building boundaries in the form of polylines while abiding by restrictions like orthogonality and parallelism is the main goal of 2D building polygon extraction. Building polygon extraction consists of 3 procedures: initial line extraction, line correction, and regularization. The study suggests using Douglas-Peucker approach for the first line extraction. This procedure is commonly used in the field of computational geometry to simplify a curve or polyline. The program finds the point on the curve that deviates the greatest from the estimated line segment repeatedly. In the simplified model, this point, known as the "furthest point," becomes an important vertex. At this stage, the curve is split into two parts, and the procedure is performed recursively on each section (Douglas & Peucker, 1973). However, it has the potential to produce erroneous and short line segments (Gui & Qin, 2021).

Using the Line-Segment Detector approach, the regularization phase enhances the line orientations, and the line adjustment phase joins and extends these segments depending on the main orientations. In order to make fitting simplex models easier, a grid-based rectangle decomposition approach is also used to separate large architectural polygons into smaller rectangles. Combination of DSM and orthophoto data help to find suitable lines and conduct maximum inner rectangle extraction.

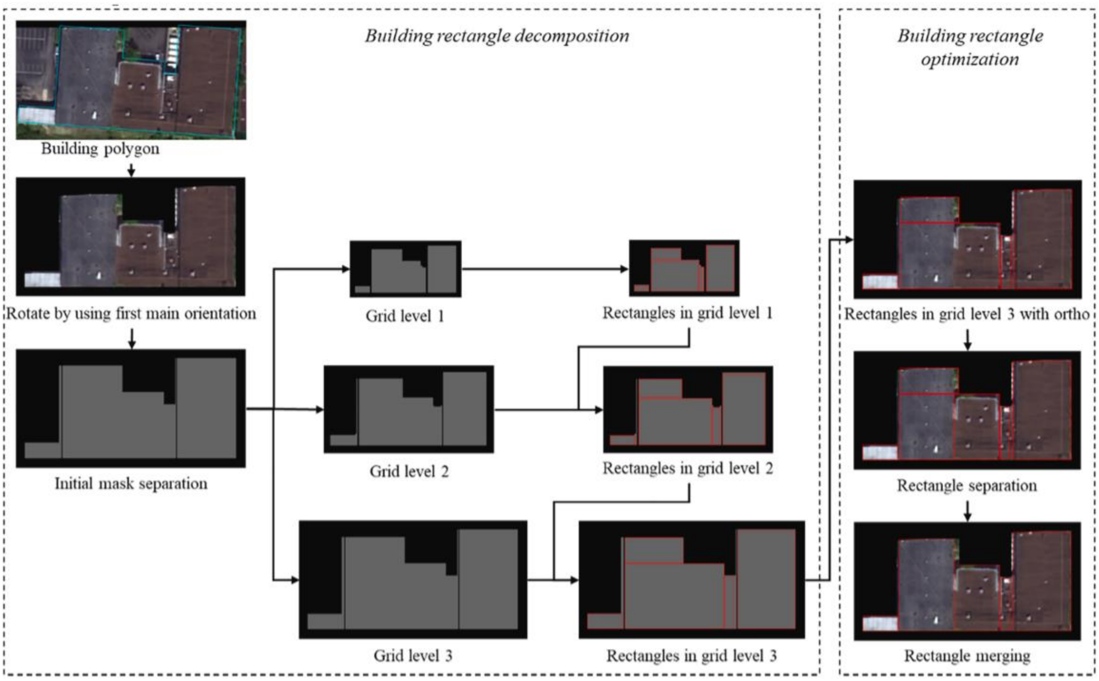

Building rectangle decomposition
Building rectangle optimization
Building polygon
Rotate by using first main orientation
Initial mask separation
Grid level 1
Grid level 2
Grid level 3
Rectangles in grid level 1
Rectangles in grid level 2
Rectangles in grid level 3
Rectangles in grid level 3 with ortho
Rectangle separation
Rectangle merging

Figure 7: Processes of building polygon decomposition starting from building outline extraction (Gui & Qin, 2021).

Together, these methods improve building outline extraction and simplification, facilitating the rapid and accurate reconstruction of 3D models (Gui & Qin, 2021). Basic building models, such as flat, gable, hip, pyramid, and mansard, are used to match the extracted building polygons during the fitting step of the 3D model reconstruction process.

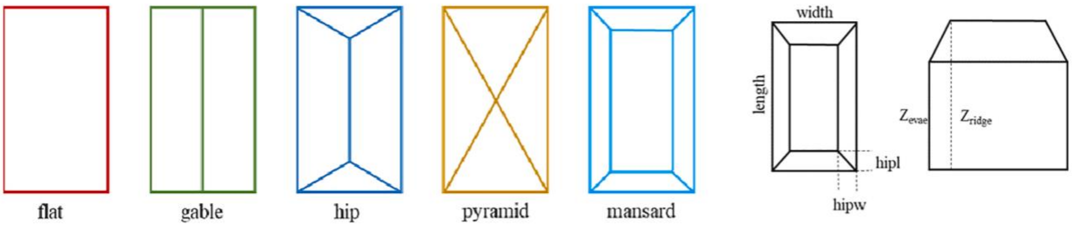


Figure 8: Model shapes for building model reconstruction (Gui & Qin, 2021).

The best-fitting model for each polygon is chosen using an extensive optimization process that considers the DSM data and other geometrical properties associated with each model. Two processes are used in the post-processing step to improve the building models. First, taking into consideration the similarity in color and height of nearby structures, Graph-Cut optimization is used to impose consistency in building types. This guarantees that different building kinds are represented consistently. Second, using model-level merging, which identifies and combines close building rectangles based on criteria like height and color variations, complex 3D structures are reconstructed. The final 3D renderings of the buildings are created by optimizing the merged models. The accuracy and coherence of the building models are enhanced by these post-processing techniques (Gui & Qin, 2021).

Chapter 3. Experiment

This chapter commences by providing a comprehensive description of the dataset employed in the experiment. It then proceeds to explain the manual process of generating borders, followed by the implementation of a state-of-the-art model for generating masks. Additionally, the study proceeds to the training phase, where a specialized model is developed to accurately classify borders, accompanied by a thorough testing procedure. Furthermore, the chapter encompasses the application of line fitting techniques on the predicted borders. Furthermore, it describes the development of a model for reconstructing the data and proceeds with an evaluation of the experiment's outcomes. The chapter concludes with a comprehensive discussion of the findings.

3.1. Dataset

This research includes satellite photos, Digital Surface Models (DSMs), and binary building masks from Trento, Italy (Qin et al., 2022). Satellite imagery captures precise information about buildings and their surroundings by providing high-resolution visual data of the Earth's surface. DSMs, on the other hand, represent the terrain's and constructed buildings' elevation or height values, enabling correct reconstruction of building geometry. The binary building masks identify the presence or absence of buildings in the images, providing useful ground truth data for segmentation and reconstruction tasks. These datasets are critical for effective building segmentation, which involves identifying and delineating individual structures, as well as later reconstruction activities, which include creating detailed 3D models of the buildings.

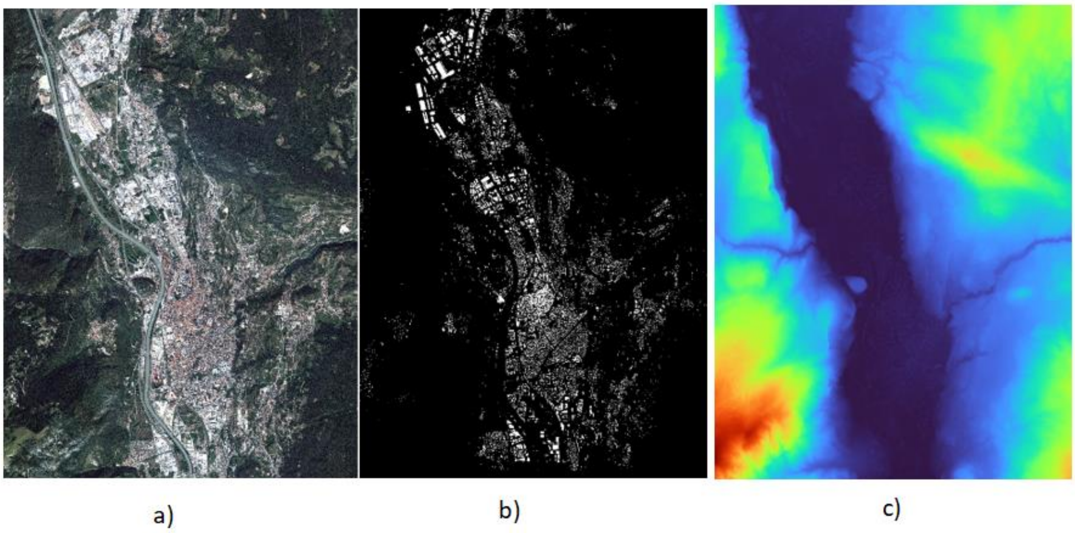


Figure 9: a) Orthophoto, b) building mask and c) DSM for Trento.

Satellite imagery of Trento contains; RGB image, building mask where pixels over building is 1, otherwise 0 and Digital Surface Model (DSM) obtained from satellite image (Shown in Figure 9).

To test model performance, satellite imagery and segmented building mask of London was used. This dataset has the same features as Trento, but image patches extracted from size of 14434x14407 image with an overlapping pixel of 256 in both vertical and horizontal direction and model tested in 1536x1536 size image.

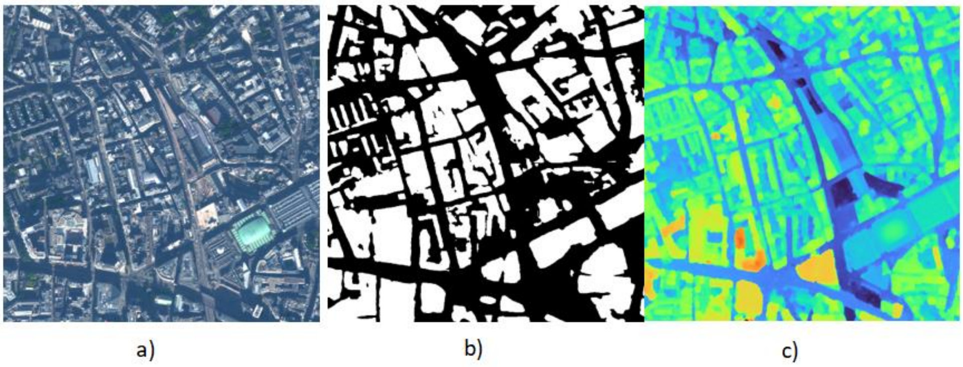


Figure 10: a) Orthophoto, b) building mask and c) DSM for London.

Digital Surface Models (DSMs) developed from LiDAR data are also used in the research to evaluate the accuracy of DSMs created from building reconstruction models. LiDAR technology works by sending out laser pulses and timing how long it takes for the signals to return after reaching the surface of the Earth. The elevation or height values of the topography and built-up areas are captured in very precise DSMs using this data. For assessing the DSMs produced by the building reconstruction models, the DSMs derived from LiDAR are a trustworthy reference.

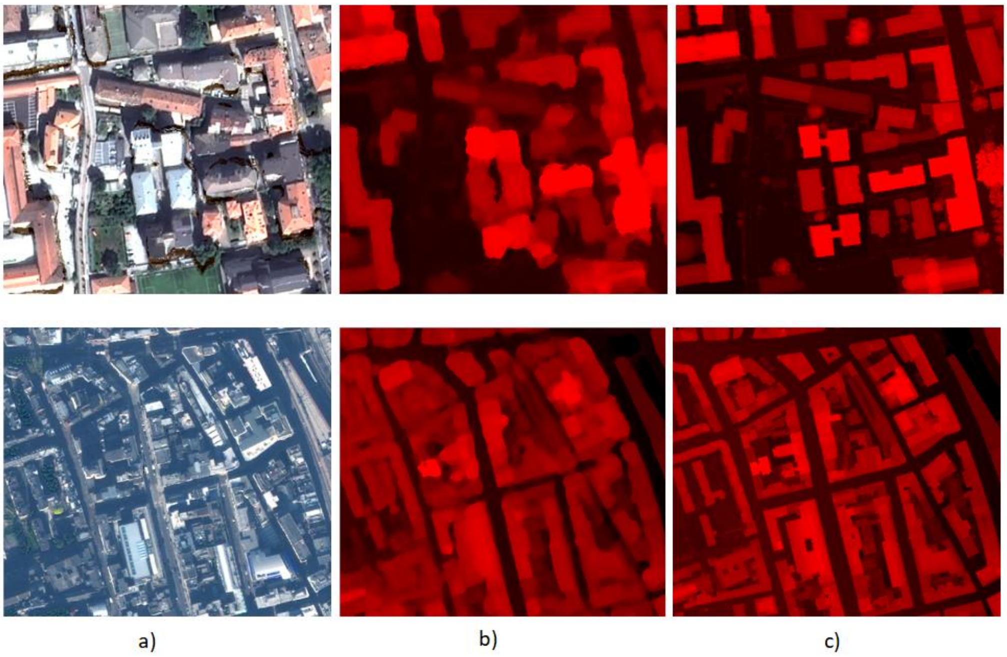


Figure 11: Figures show a) orthophoto, b) satellite derived DSM and c) LiDAR derived DSM for Trento and London.

In conclusion, the datasets used in this study, including satellite images, DSMs, and binary building masks, offer essential data for building segmentation and reconstruction tasks. They allow for the precise production of 3D models as well as the identification and delineation of specific features.

3.2. Border Generation

Understanding complex and nearby buildings is the main goal for building reconstruction. For that reason, this study starts with building border generation. Generation of building borders are made manually via geographic information system of QGIS. There are two types of borders. One of them is for the nearby buildings and the other one is for complex buildings. The utilization of complex and nearby building borders allows for the distinction of different building components and enhances the comprehension of diverse structures. As a result, complex building components and distinct structures can be effectively separated and modeled based on their unique shapes.

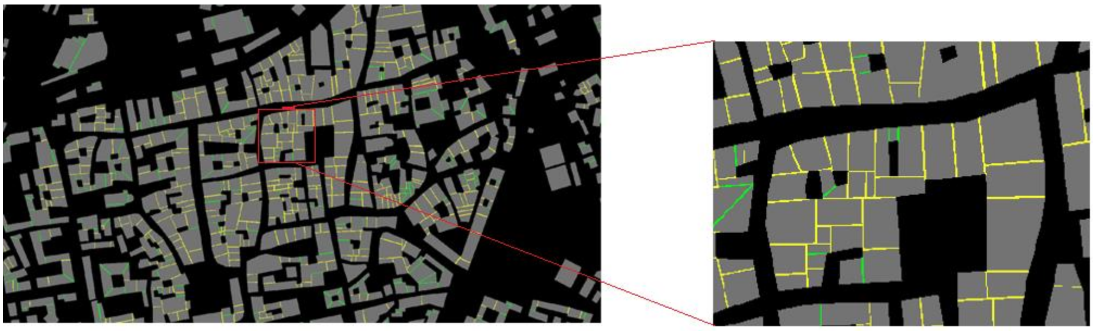


Figure 12: Two classes of borders are shown in green color for complex buildings and yellow for nearby buildings on the ground truth building mask.

3.3. Building and Border Segmentation Model

Following the acquisition of building masks and manual delineation of building borders into two distinct classes, where class 1 represents borders within a single building and class 2 denotes borders between two distinct buildings, a state-of-the-art segmentation model is developed to automatically detect these borders within the provided input dataset. This section elaborates on the application of the input data and provides detailed instructions on the training and testing procedures of the model. Building and building border segmentation model is employed with an open-source code package of mmsegmentation produced by (OpenMMLab, 2023).

3.3.1. Building Segmentation Model Training and Testing

Using a pretrained HRNet model, the building segmentation approach was employed on the input dataset. The dataset used to train the model had 31,612 patches of size 512x512 for training and 3,512 patches for validation. With 160,000 iterations of pretrained data, the accuracy for the building class and the world scale were 73.27% and 83.54%, respectively.

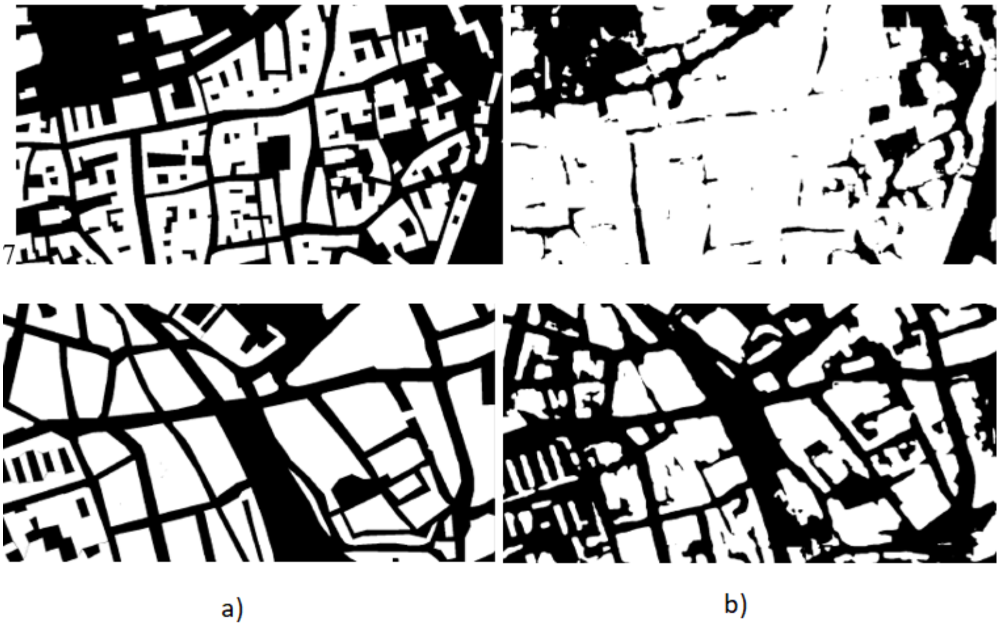


Figure 13: a) Ground truth building footprints and b) the corresponding prediction results for Trento and London.

3.3.2 Border Segmentation Model Training and Testing

The model was trained with combination of RGB satellite imagery and building mask and two classes border as inputs. The dataset consists of 419 image patches, each measuring 512x512 pixels, with a 256-pixel overlap in both the horizontal and vertical directions. Additionally, there are 46 image patches designated for validation purposes. The dataset exhibits an imbalance in the number of samples per class, leading to a higher representation of small areas with dense building structures. To address this issue, class weights were computed and assigned before training the model, as the input labels were found to be imbalanced.

The HRNet architecture was employed for the model, and the optimization process utilized the stochastic gradient descent (SGD) algorithm. The learning rate for the optimization was set to 0.01. After the training process, the model demonstrated

Intersection of Union (IoU) scores of 8% for complex and 9% for nearby building borders.

In addition to evaluating the HRNet model, another model was trained using UNet for comparative analysis. The performances of both models were calculated, and the results are presented in Table 1. The findings demonstrate that UNet exhibits relatively lower performance in contrast to HRNet. Specifically, the UNet model achieved approximately 3% IoU score for complex building borders and 7% for nearby building borders.

Table 1: Performance comparison between UNet and HRNet in terms of F1 and IoU scores.

Model Type	**Class Name**	**F1**	**IoU**	**Accuracy**
UNet	Non-border	0.9888	0.9778	0.9783
	Complex Border	0.0573	0.0295	0.5281
	Nearby Border	0.1309	0.0700	0.6872
HRNet	Non-border	0.9984	0.9968	0.9980
	Complex Border	0.1485	0.0802	0.1319
	Nearby Border	0.1741	0.0954	0.3273

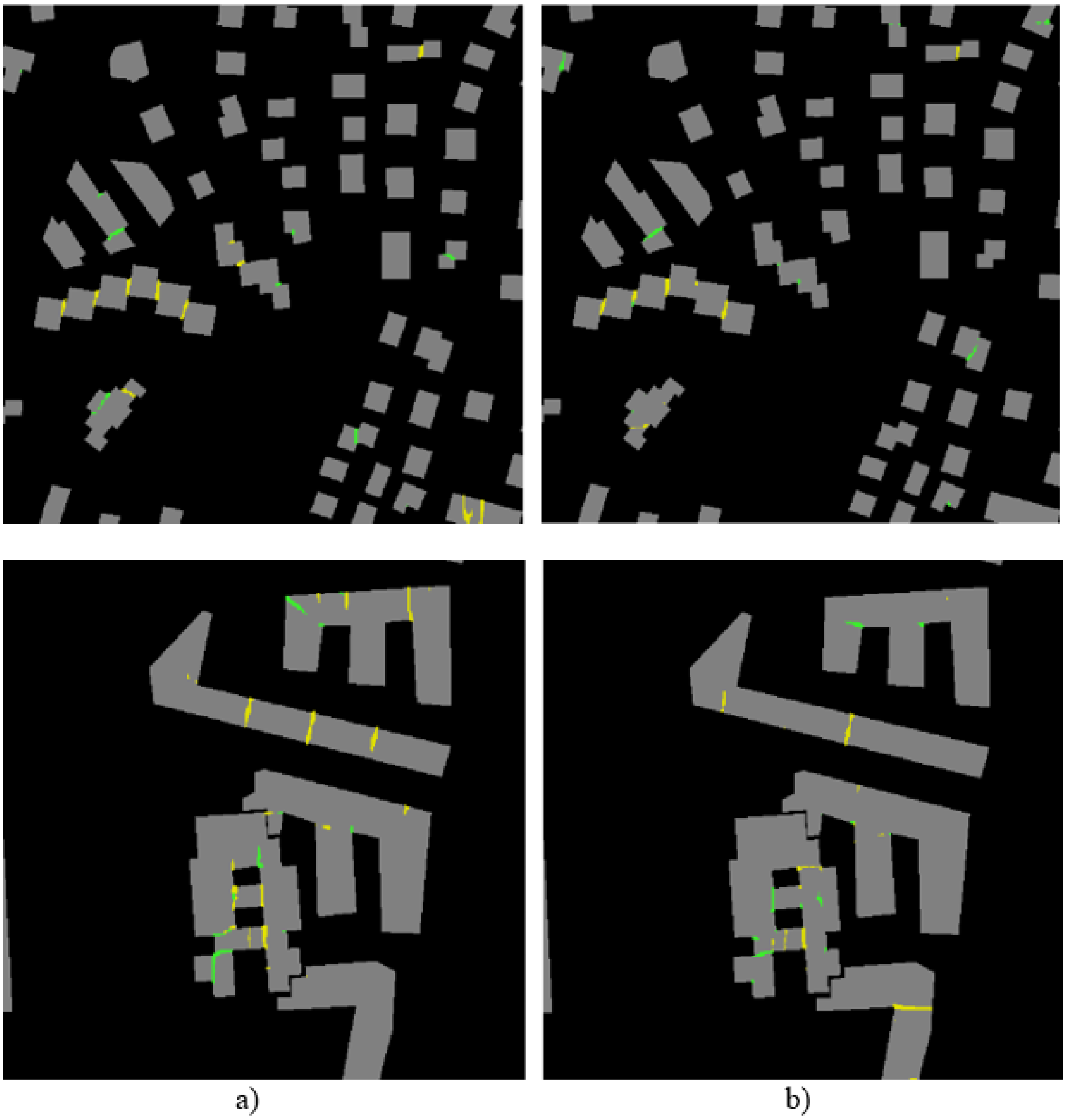
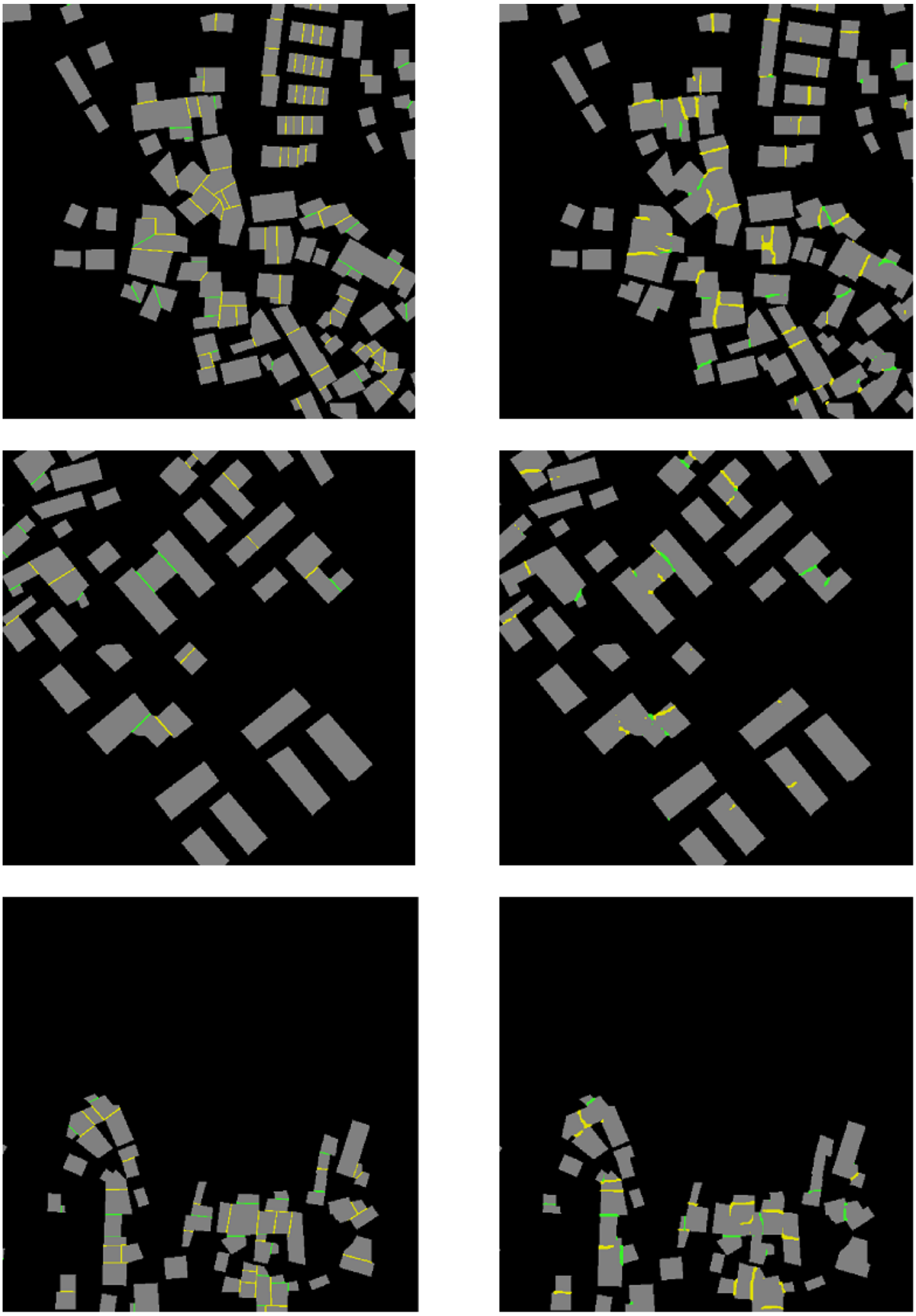


Figure 14: Figures show prediction results a) from HRNet and b) from UNet.

Since there are not significant differences between performance of HRNet and UNet models, their visual representations in Figure 14 show similarities.



a) b)

Figure 15: a) Ground truth and b) predicted borders from HRNet model are shown where green and yellow color indicate complex and nearby building borders respectively for Trento.

The model performed around the same mIoU result of 8% for complex and 9% for nearby building border on the testing images.

The London region has a complicated building structure, and the model, which was mostly trained on the Trento dataset, had difficulty properly recognizing the majority of borders there. In contrast to Trento, the model performed less well in this complex environment, showing decreased performance.

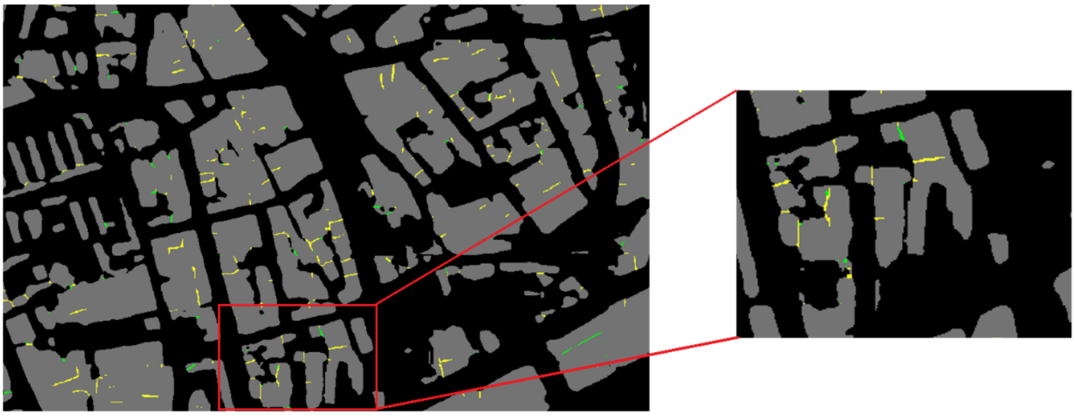


Figure 16: Predicted borders are shown where green color indicates complex building borders and yellow color shows nearby building borders from London.

3.4. Line Vectorization on Border Prediction

The building masks and borders were subjected to the linked component analysis approach at this step. Applying connected component analysis served the objective of precisely fitting lines on building components while ignoring other components. Only the segmented border pixels that had the same position as the building mask were taken into account, and each building component was examined separately.

During the component analysis, certain criteria were used to ensure accurate lines. Both line components with fewer than 10 pixels and building components with fewer than 5 pixels were excluded. These criteria aid in the removal of irrelevant or distracting components.

The RANSAC line fitting method was used with a predetermined number of 5000 iterations, after the component analysis. For fitting mathematical models, RANSAC is an effective technique that can precisely predict the ideal parameters based on the given data points.

The midpoint of each line was determined, and the line was then stretched by 30 pixels in both directions by utilizing the line coefficients. The method for line extension determines if the line extends to 0 pixels in both directions. If the line does not reach zero pixels in both directions, the line fitting process is not executed, ensuring that only complete lines are taken into consideration.

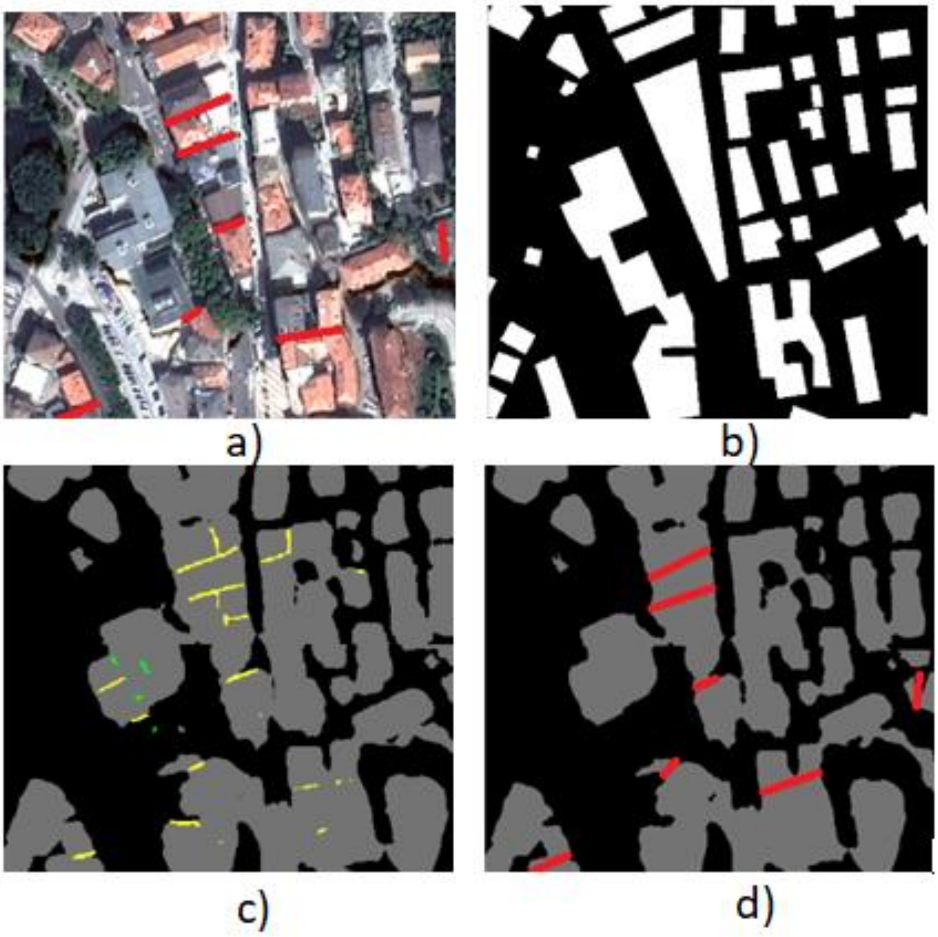


Figure 17: a) Satellite image, b) ground truth building mask, c) building and border segmentation result, d) line vectorization result with building mask segmentation map (green and blue color represent complex, yellow and red color represent nearby building border.) for Trento.

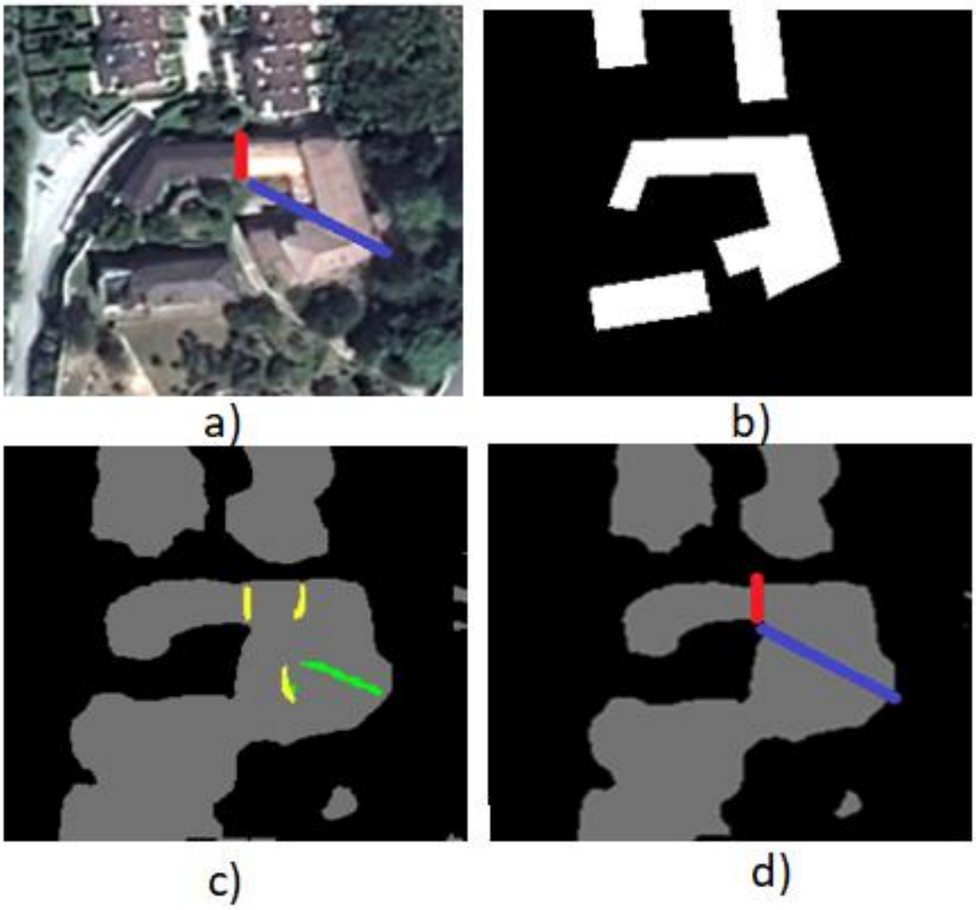


Figure 18: a) Satellite image, b) ground truth building mask, c) building and border segmentation results and d) line vectorization result with building mask segmentation map (green and blue color represent complex, yellow and red color represent nearby building border.) for Trento.

The application of a line fitting criteria presents a constraint where lines are only fitted if they intersect regions with zero-pixel values in both line directions within the building polygon. To accommodate long and narrow polygons, a line length threshold of 30 is adopted, following the exploration of higher parameters. While the line fitting algorithm has demonstrated success for many cases, it encounters limitations when dealing with large building polygons, leading to suboptimal outcomes in such scenarios.

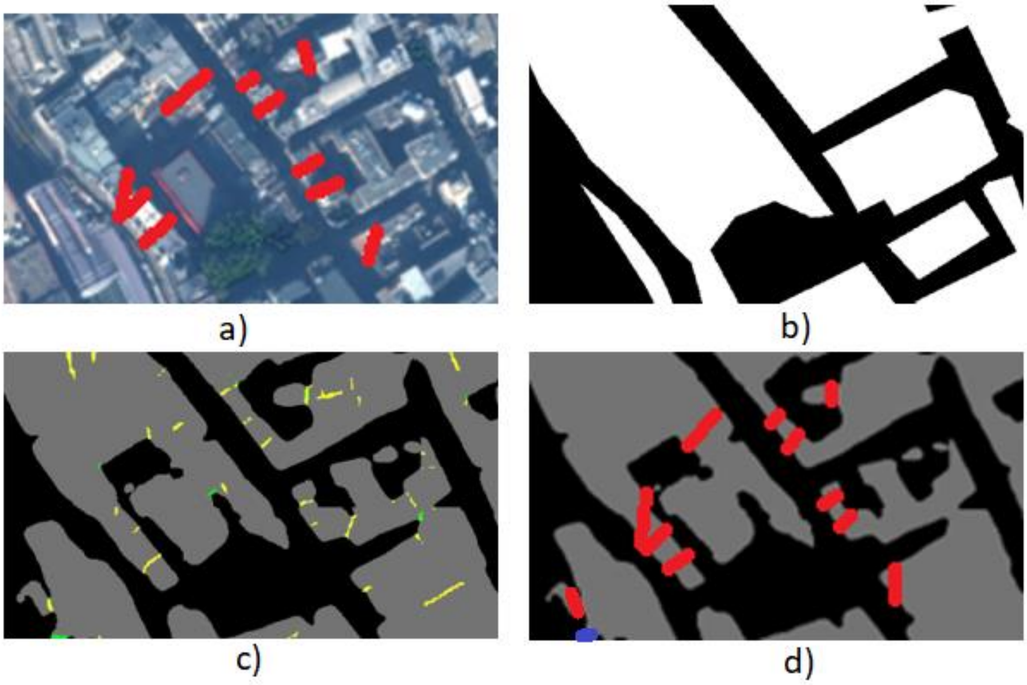


Figure 19: a) Satellite image, b) ground truth building mask, c) building and border segmentation results and d) line vectorization result with building mask segmentation map (green and blue color represent complex, yellow and red color represent nearby building border.) for London.

This process accomplishes accurate line fitting on the building components, enabling the exact delineation of the structures' borders, by applying connected component analysis followed by RANSAC line fitting. This method facilitates the study's goal by ensuring the identification and extraction of the desired lines that serve as the building borders.

3.5. Building Model Reconstruction

A comprehensive approach to building model reconstruction, comprising several key steps. The process commences with the initial extraction of 2D polygons, where building footprints are delineated from the input images. Subsequently, a grid-based rectangle decomposition technique is employed to further refine the building representation, breaking down the extracted polygons into individual building rectangles. To enhance the precision of the model, a graph-cut labeling method is applied to refine the orientation of these rectangles. Next stages involve 3D model fitting, where the extracted building rectangles are transformed into 3D structures based on the input data. Moreover, an essential aspect of the process is the enforcement of building type consistency, ensuring that the reconstructed models align with architectural conventions and regional characteristics. Lastly, complex 3D buildings are recovered through model-level merging, where the fragmented structures are integrated into coherent and detailed representations. By systematically combining these steps, the proposed approach aims to achieve more accurate and comprehensive building model reconstructions, offering valuable insights for urban planning, architectural design, and various geospatial applications.

3D building model reconstruction is processed by an open-source code package of Sat2lod2 produced by (Gui et al., 2022). This open-source code is implemented in Python for a comprehensive building model reconstruction task.

Application of initial line fitting detects straight lines representing building outlines. Subsequently, a line correction process is performed to refine the detected lines, followed by regularization techniques to ensure their consistency and adherence to architectural principles. Once the building outlines are obtained, the extraction of building rectangles is achieved through a building decomposition procedure. This step facilitates the breakdown of complex building structures into simpler rectangular components, which forms a crucial foundation for subsequent reconstruction efforts.

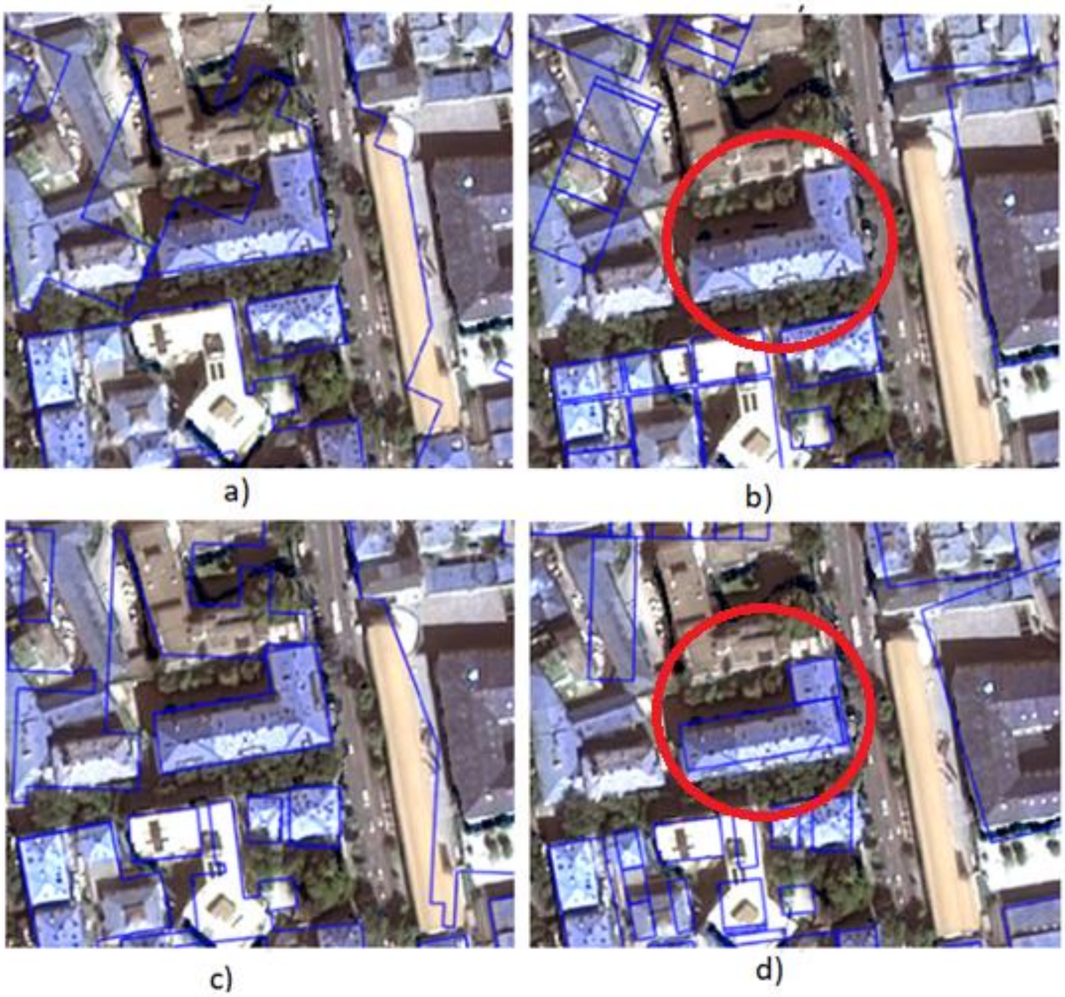


Figure 20: Figures a) and b) show initial building outline extraction and decomposition from building mask, c) and d) initial building outline extraction and decomposition from our building border detection.

The outline extraction from building mask segmentation is slightly different for the two cases. In Figure 20a), it successfully detects the initial building outline, but in Figure 20b), it fails during decomposition. Building in red circle shows the decomposition result. Conversely, in Figure20c), the building mask segmentation with our method extracts the initial building outline, which is then successfully regularized into rectangles in Figure 20d). The building is shown in a red circle with two decomposed rectangles.

Notably, certain nearby borders have been effectively detected, offering significant advantages in automatically identifying nearby buildings and treating them as distinct entities during the modeling process. The integration of these detected borders facilitates more precise and efficient building model reconstructions, particularly in densely populated urban environments. However, it is noteworthy that the current approach, relying solely on building masks, considers these adjacent buildings as a single entity and fits a singular suitable model. Further refinement and consideration of the detected borders may be essential to accurately capture the individuality of neighboring buildings within the reconstruction process.

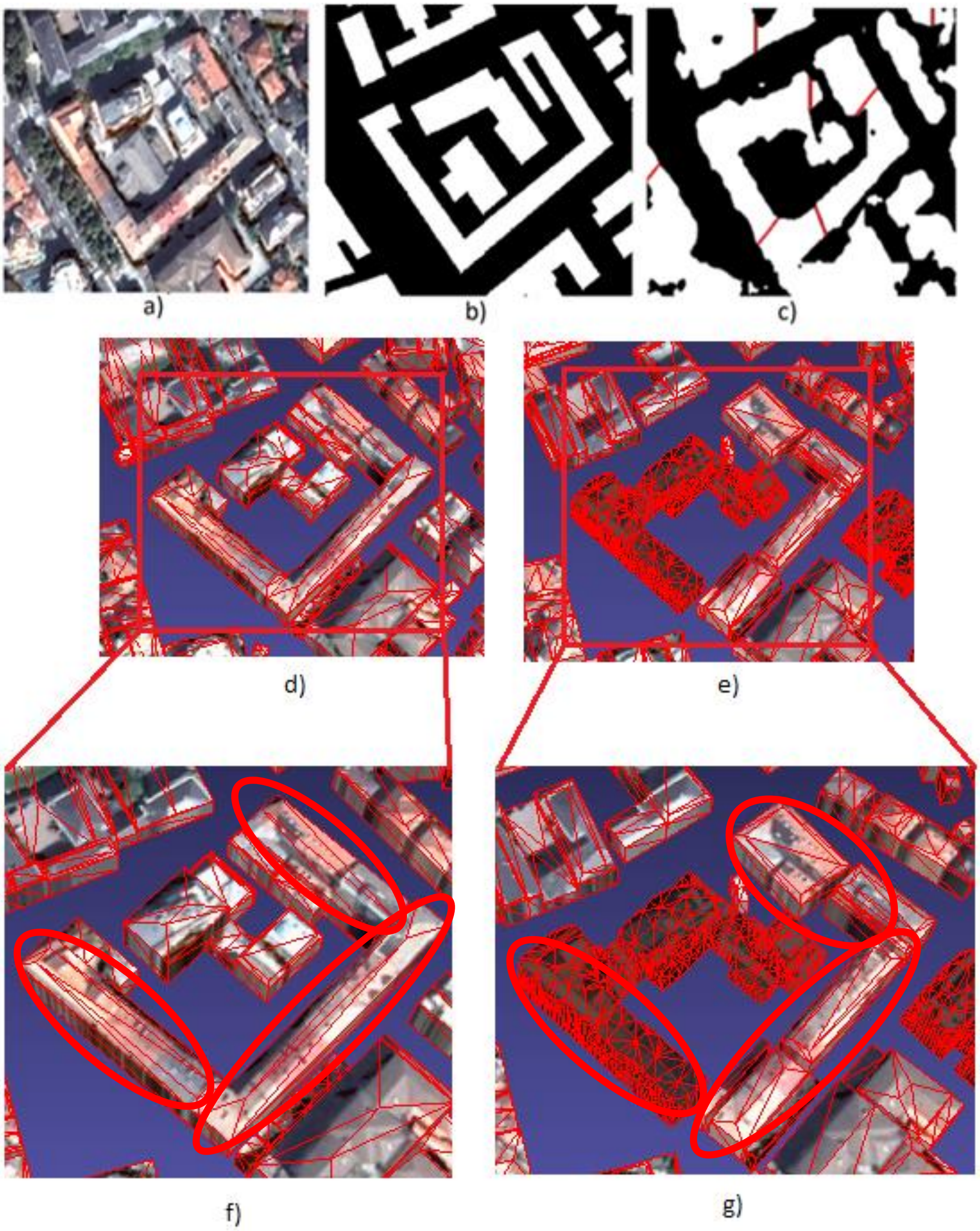


Figure 21: Figures present a) satellite image, b) ground truth building mask, c) segmentation map of building mask with vectorized border prediction, d) and f) building model from building mask, and e) and g) building model from our building border detection.

The vectorized borders intricately divide the building mask into multiple polygons, as depicted in Figure 21c) below, owing to the presence of distinct adjacent buildings within the area shown in Figure 21a). Following the reconstruction of the model, Figure 21d) and Figure 21c) represent the region of interest. The same area within the rectangles is further magnified, and disparities are illustrated using ellipses in Figure 21f) and Figure 21g).

The results reveal that, in the case of Figure 21f), each of the three building blocks have been assigned a unique model for reconstruction, considering that the buildings are interconnected. Conversely, in Figure 21g), different models are employed for different building components, even though not all buildings are precisely detected as separate entities. The number of modeled buildings is more than other.

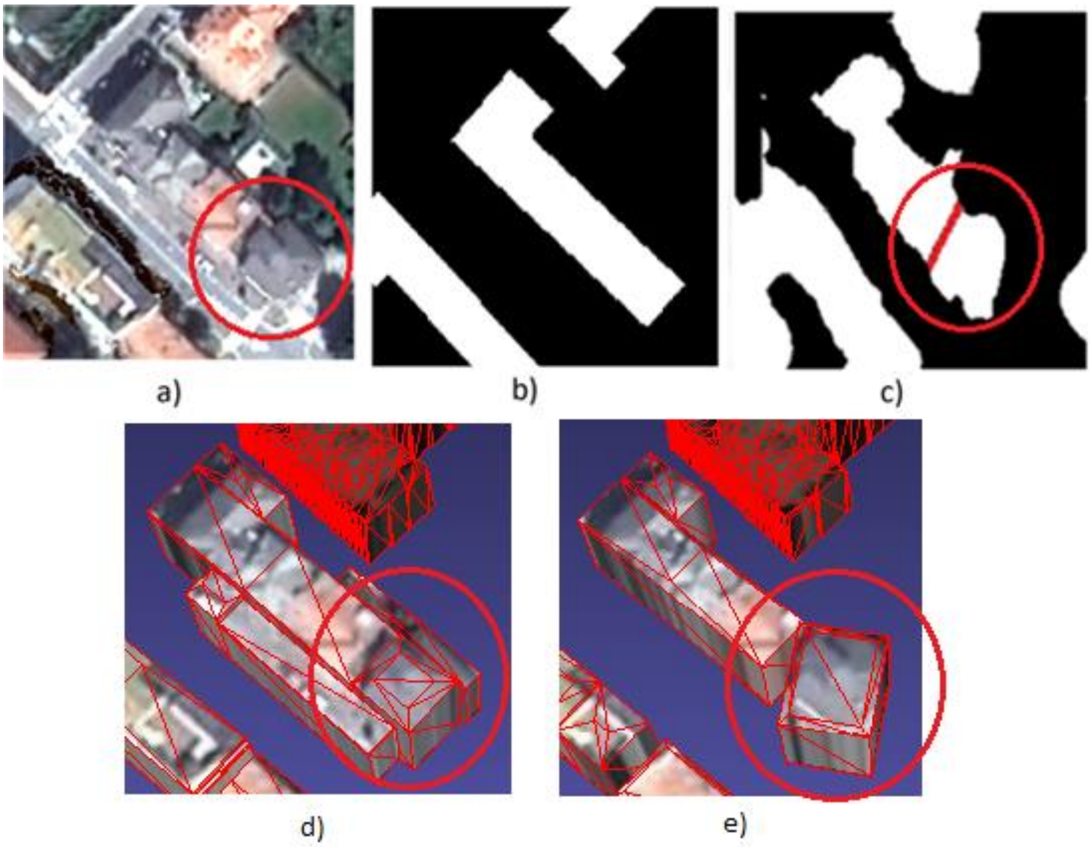


Figure 22: Figures present a) satellite image of buildings, b) ground truth building mask, c) segmentation map of building mask with vectorized border prediction, d) building model from building mask and e) building model from our building border detection.

The individual modeling of buildings is made easier by the detection of specific boundaries made possible by the reconstruction. However, since certain structures need merging with neighboring components, simple separation by boundary detection is insufficient for complete building modeling. Figure 22a) shows a distinct adjacent building in the red circle and Figure 22c) shows that border segmentation and vectorization successfully separate the building from the mask. Building mask modeling is shown to be successful in Figure 22d) where building mask segmentation directly used to model fitting, but using our building border detection enables to separate building modeling, though it requires merging some components later to produce a more accurate representation Figure 22e).

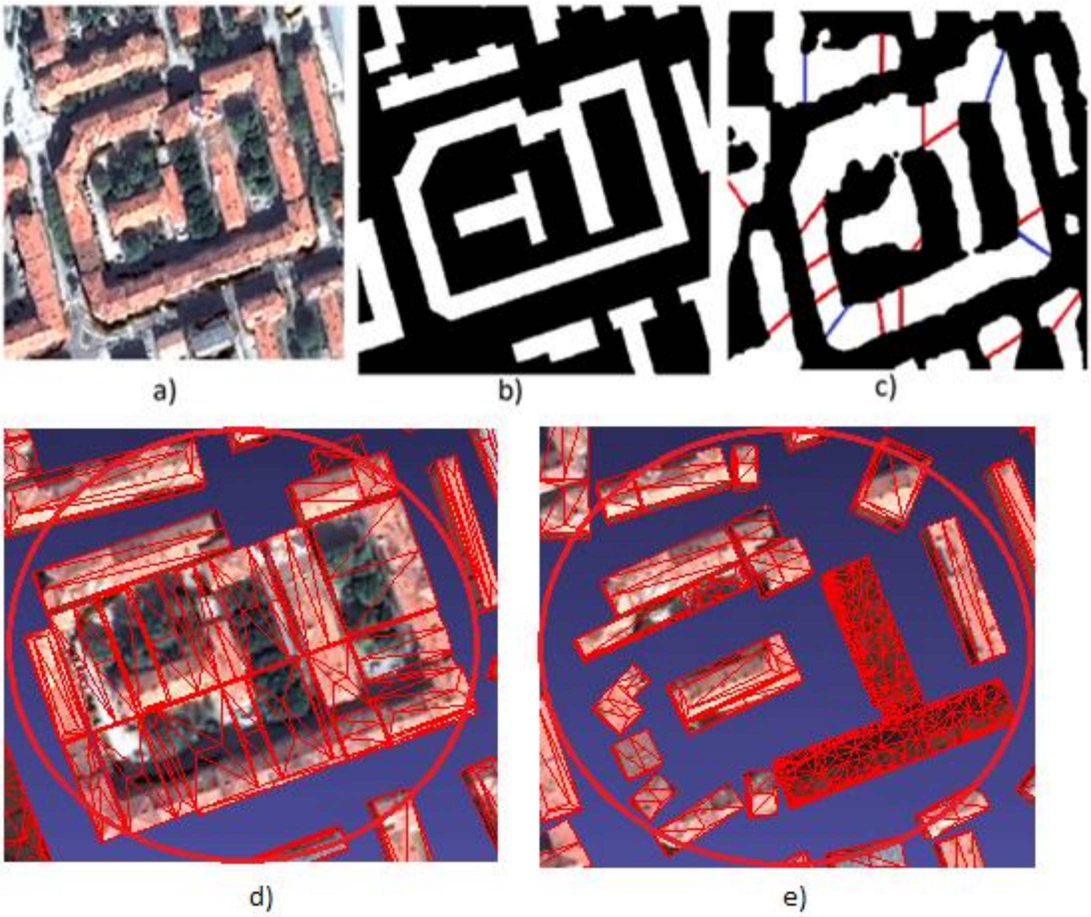


Figure 23: Figures present a) satellite image of buildings, b) ground truth building mask, c) segmentation map of building mask with vectorized border prediction, d) building model from building mask and e) building model from our building border detection.

The inclusion of borders in the building mask is beneficial in the case of complex structures such as those shown in Figure 23a) since it helps to divide the mask into distinct components. This method ensures an accurate representation of the buildings of interest by preventing the unintended inclusion of non-building regions during the modeling stage.

When utilizing the building mask directly for reconstruction (as depicted in Figure 23d), the buildings are considered identical in the red circle, resulting in a unified model. However, our building border detection method allows for partial separation of some buildings, but with suboptimal performance in border segmentation. Consequently, the reconstruction outcome in Figure 23e) exhibits diminished quality owing to the limitations of border segmentation and separately reconstructed buildings are shown in the red circle.

Complex building borders are currently utilized exclusively for the purpose of separating building components. However, it is important to note that a more refined and improved procedure can allow for the merging of these components when appropriate. This enhancement in the methodology would enable a more comprehensive and accurate representation of the complex building structures.

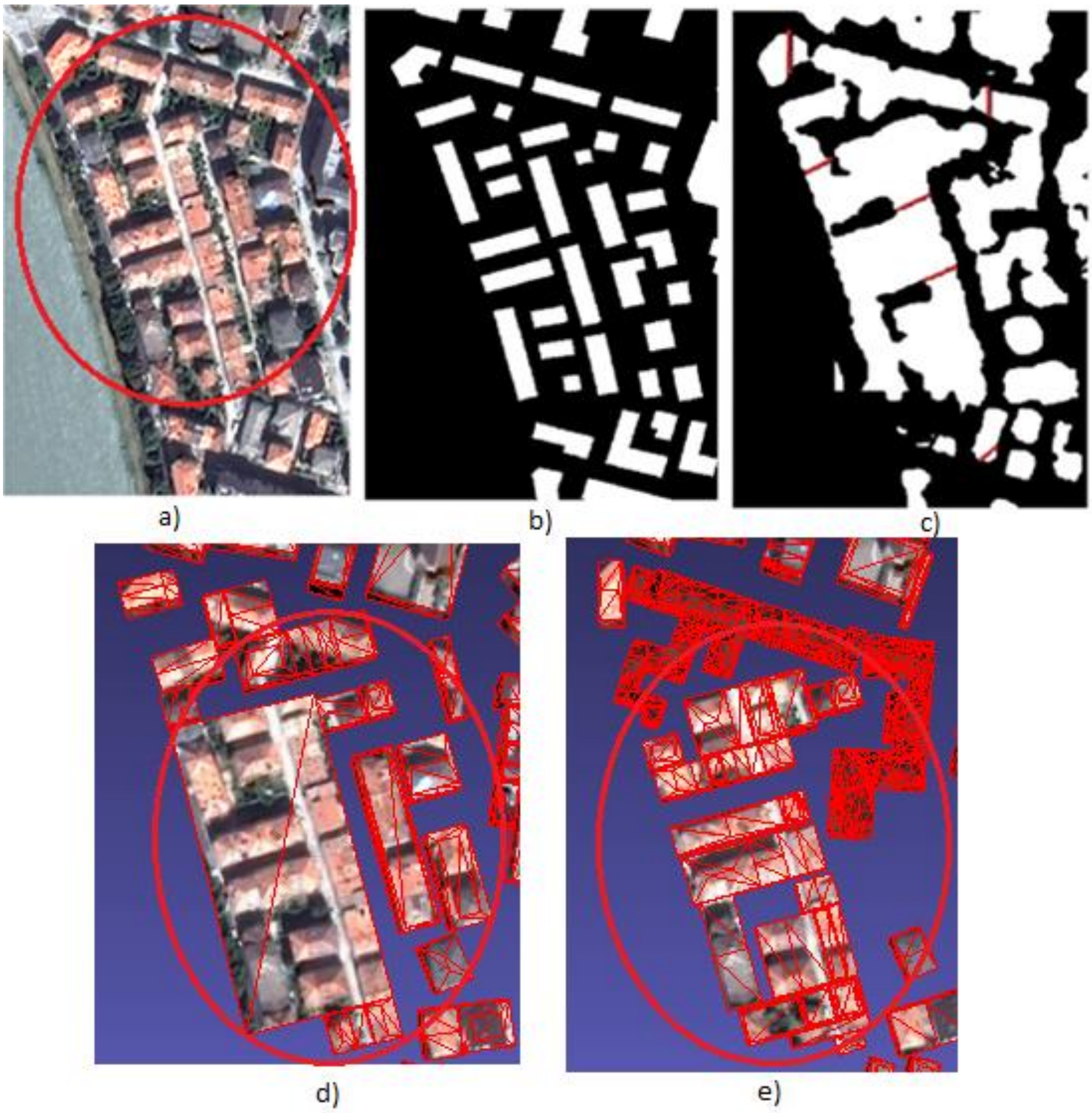


Figure 24: Figures present a) satellite image of buildings, b) ground truth building mask, c) segmentation map of building mask with vectorized border prediction, d) building model from building mask and e) building model from our building border detection.

Building modeling utilizing building segmentation maps enriched with borders proves beneficial by enabling the separation of certain buildings, as exemplified in Figure 24. There are many discrete buildings in the area of interest shown in Figure 24a). Building modeling with building mask segmentation yields limited models seen in Figure 24d) with a red circle. It identifies all buildings the same and fits a single model on distinct buildings. However, our building border detection seen in Figure 24e) allows for the successful separation of a subset of buildings, enhancing the reliability and accuracy of the modeling process.

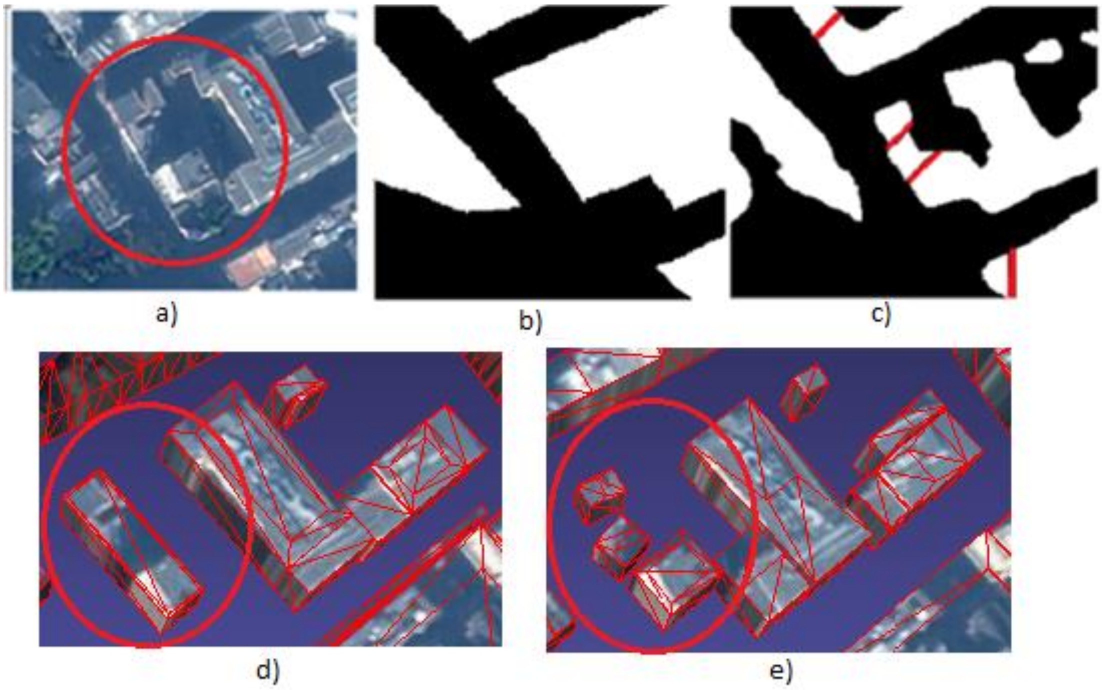


Figure 25: Figures present a) satellite image of buildings, b) ground truth building mask, c) segmentation map of building mask with vectorized border prediction, d) building model from building mask and e) building model from our building border detection.

Similarly, there are adjacent buildings seen in Figure 25a) for London area and during reconstruction with building mask segmentation, it fits a single model on adjacent buildings shown in Figure 25d). Conversely, our building border detection method fits multiple models on the distinct buildings shown in Figure 25e).

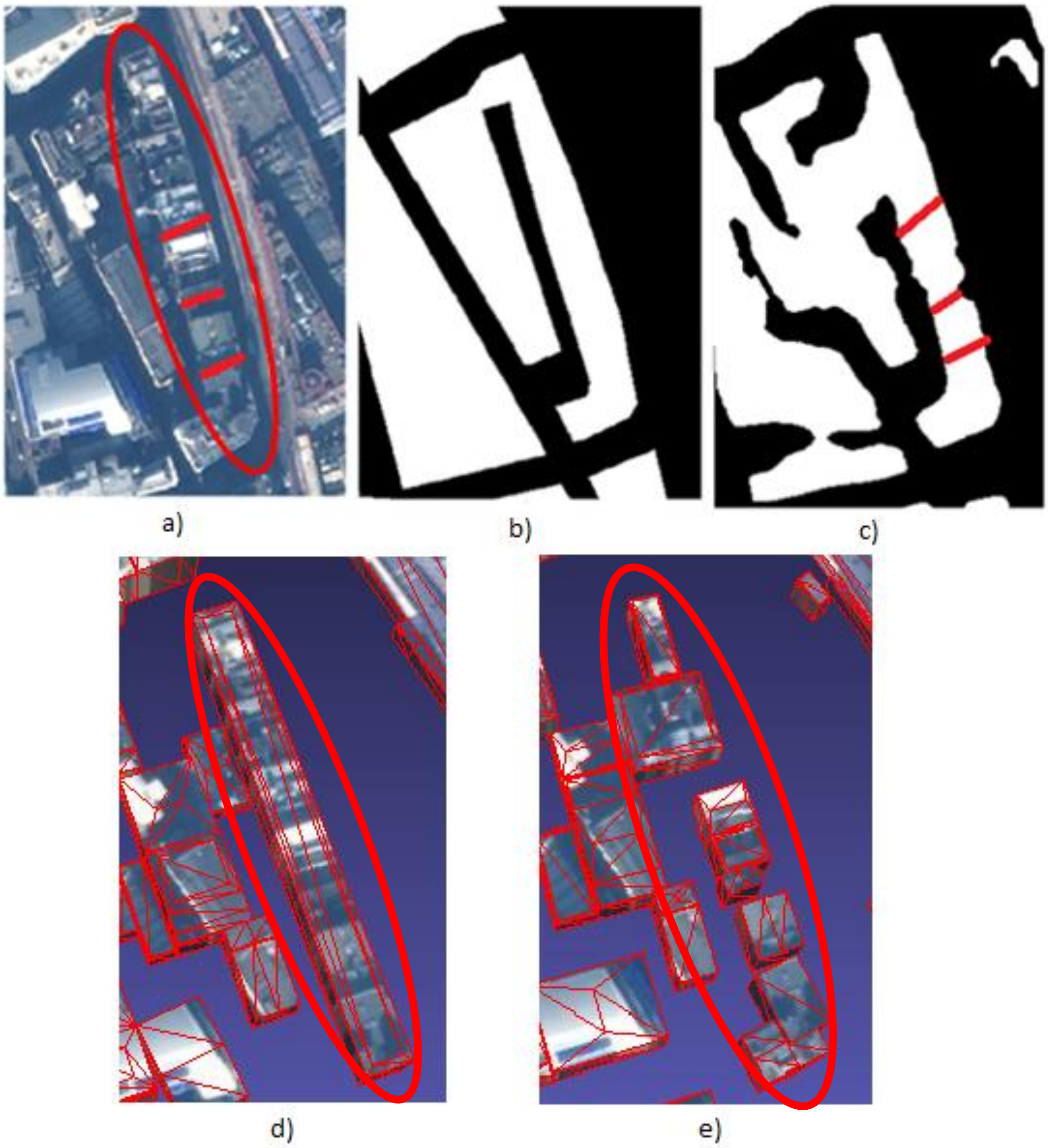


Figure 26: Figures present a) satellite image of buildings, b) ground truth building mask, c) segmentation map of building mask with vectorized border prediction, d) building model from building mask and e) building model from our building border detection.

A reconstruction result for the London area is shown in Figure 26. It is evident that there are several adjacent buildings, as highlighted within the red ellipse seen in Figure 26a). When employing the building mask for modeling these structures, the fitting process yields a single model for the combined buildings, as depicted in Figure 26d). Conversely, our building border detection method results in the emergence of multiple models, as shown in Figure 26e). Despite their adjacency in reality these models are treated separately during the reconstruction process.

The model's performance in London is observed to be lower compared to Trento, primarily due to the fact that the model was trained using the Trento dataset. As a consequence, the model's ability to generalize to the different architectural characteristics and urban landscapes of London is limited. Nevertheless, when comparing the building mask alone to the utilization of building borders, the latter exhibits improved performance in reconstructing additional building sections. The incorporation of building borders proves beneficial in enhancing the precision of the reconstruction process, resulting in a more comprehensive representation of the complex building structures in both cities.

3.6. Reconstruction Result Evaluation

A key factor in deciding the overall precision of LoD-2 building model reconstruction is the exact segmentation of buildings, together with accurate outline extraction and decomposition. To explore the quantitative correlation between ground truth data, building masks, and building masks generated through our building border detection method, we utilized the Columbus and London datasets. The evaluation of the resulting models' accuracy is conducted using the IOU_2 metric, which assesses the precision of

2D building footprint decomposition, and the IOU_3 metric, which measures the accuracy of 3D building model fitting.

To evaluate the accuracy in 2D;

$$IoU_2 = \frac{TP}{TP+FP+FN}$$

where TP refers to the quantity of true positive pixels, signifying those pixels that are accurately identified as building footprints through both automated extraction and manual labeling. FP represents the count of false positive pixels, indicating pixels incorrectly identified as building footprints by the automated process. Conversely, FN corresponds to the number of false negative pixels, representing pixels that are part of actual building footprints but are erroneously missed by the automated extraction.

Following equation is used to evaluate accuracy in 3D;

$$IoU_3 = \frac{\mathrm{TP3D}}{TP3D+FP+FN}$$

Additionally, TP_{3D} is a specific subset of TP pixels. It refers to those true positive pixels whose vertical difference, measured in 3D, from the ground-truth LiDAR data falls within a margin of 2 meter (Gui & Qin, 2021).

Table 2: Evaluation result of building reconstruction.

Region	**Accuracy**	**Original Mask**	**Border Mask**
Trento	IoU_2	0.3467	0.4508
	IoU_3	0.2190	0.3039
London	IoU_2	0. 5092	0. 5013
	IoU_3	0. 2149	0. 2163

Table 2 presents the performance comparison between our building border detection and original building masks in a 2D and 3D context. The results indicate that our method outperform the original masks in terms of 2D performance. Additionally, the same level of improvement is observed for IoU_3 in the training area of Trento when using our building border detection strategy. However, when comparing the results for London, both types of masks show approximately the same level of accuracy.

3.7. Discussion

Automated reconstruction of building models remains a challenging task, involving various aspects such as identifying building mask and borders, border vectorization, extracting outlines, decomposing complex structures, and fitting the models. This research introduces a novel method that utilizes the borders of distinct and complex buildings to detect buildings and their components separately. While the method shows promising results, the model's performance goes beyond relying solely on the segmentation map. The study identifies that adjacent buildings may be erroneously separated, necessitating the merging of such buildings at their edges. Additionally, the current approach struggles to realistically reconstruct complex buildings since it tends to model their components independently rather than as an integrated whole.

Conversely, the performance of the segmentation model during training exhibits limitations attributed to the lack of the border class in comparison to the background and building classes. Enhancing the model's comprehensiveness could potentially lead to improved accuracy by enabling the detection of additional borders within intricate urban regions. The introduction of a novel segmentation model has the potential to yield more favorable outcome results, thereby advancing state-of-the-art in this area of research.

Furthermore, this investigation suggests that employing an enhanced border vectorization method could lead to higher accuracy. The study involves the detection of each building component, followed by the application of border vectorization within the corresponding polygon segment. However, it is noted that certain building polygons encompass a wide building area, which may result in the failure of line fitting when using a threshold approach. These observations highlight the significance of refining the border vectorization technique to address such challenges and potentially yield improved results in this context.

Chapter 4. Conclusion

This research endeavors to address the challenges of accurately identifying and separating complex building parts and distinct buildings through the utilization of advanced deep learning techniques. The study leverages valuable information derived from building borders to effectively detect and distinguish individual buildings and their respective sections. Additionally, the research aims to enhance 3D building model reconstruction by integrating Digital Surface Models (DSMs), thereby refining the reconstruction process, and producing more precise and comprehensive representations. The resulting reconstructed building models offer valuable insights into the spatial arrangement, architectural characteristics, and composition of intricate urban structures, making them a valuable asset for urban planning, architectural design, and diverse geospatial applications.

In the segmentation task, the performance of the complex building border class was observed to be lower compared to the nearby building border class. This disparity can be attributed to the fact that the complex building border class did not benefit from the RGB channels. The complex building borders were encompassed within the same building polygon, leading to a lack of distinctive information from the RGB channels. Consequently, this limited availability of RGB information adversely affected the segmentation accuracy of the complex building border class.

The utilization of building borders in the current approach exhibits suboptimal performance, as it disregards a significant portion of the predicted border pixels.

Consequently, the incorporation of building borders does not substantially impact the overall results of the task. However, it is worth noting that despite its limited effect on the overall performance, reconstruction result seems to be effective on our building border detection compared to building mask only. It gives an idea of where buildings are distinct or complex.

Future research endeavors involve enhancing the model's performance by augmenting the training process with a more extensive set of building borders derived from the building mask. Accurate prediction of building borders holds the potential to facilitate precise detection of individual buildings. In addition, the study faces challenges in achieving satisfactory line fitting results on the predicted building borders, as it selectively ignores borders based on certain parameters, limiting its efficiency. There is scope for more efficient utilization of these borders to enhance the reconstruction process.

Furthermore, an alternative approach to improve the study's performance involves leveraging additional data sources, such as Digital Surface Models (DSM). Prior research has demonstrated successful results by incorporating nDSM data as an additional parameter in the border segmentation process (Schuegraf et al., 2022). Integrating information from DSM differences may also prove advantageous in enhancing line fitting accuracy. Consideration of the geometric properties of buildings with DSM, can collectively contribute to improved reconstruction accuracy.

Another promising alternative is to explore instance-level building outline extraction techniques, which could potentially yield higher accuracy in the overall reconstruction process. By investigating these alternative approaches, the study aims to achieve more robust and precise building model reconstruction, thus advancing the effectiveness and utility of the methodology.

Bibliography

Buyukdemircioglu, M., Kocaman, S., & Isikdag, U. (2018). Semi-Automatic 3D City Model Generation from Large-Format Aerial Images. *ISPRS International Journal of Geo-Information*, *7*(9), 339. https://doi.org/10.3390/ijgi7090339

Chang, F., Chen, C.-J., & Lu, C.-J. (2004). A Linear-Time Component-Labeling Algorithm Using Contour Tracing Technique. *Computer Vision and Image Understanding*, *93*(2), 206–220. https://doi.org/10.1016/j.cviu.2003.09.002

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv*. http://arxiv.org/abs/2102.04306

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation With Deep Convolutional Nets, Atrous Convolution, And Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*. http://arxiv.org/abs/2010.11929

Douglas, D. H., & Peucker, T. K. (1973). Algorithms For the Reduction of the Number of Points Required to Represent a Digitized Line or Its Caricature.

Cartographica: The International Journal for Geographic Information and Geovisualization, *10*(2), 112–122. https://doi.org/10.3138/FM57-6770-U75U-7727

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, *59*(2), 167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77

Fischler, M. A., & Bolles, R. C. (1981). Random Sample Consensus. *Communications of the ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Grau, V., Mewes, A. U. J., Alcaniz, M., Kikinis, R., & Warfield, S. K. (2004). Improved Watershed Transform for Medical Image Segmentation Using Prior Information. *IEEE Transactions on Medical Imaging*, *23*(4), 447–458. https://doi.org/10.1109/TMI.2004.824224

Gui, S., & Qin, R. (2021). Automated Lod-2 Model Reconstruction from Very-High-Resolution Satellite-Derived Digital Surface Model and Orthophoto. *ISPRS Journal of Photogrammetry and Remote Sensing*, *181*, 1–19. https://doi.org/10.1016/j.isprsjprs.2021.08.025

Gui, S., Qin, R., & Tang, Y. (2022). Sat2lod2: A Software for Automated LoD-2 Modeling from Satellite-Derived Orthophoto and Digital Surface Model. *arXiv*. http://arxiv.org/abs/2204.04139.

Henn, A., Gröger, G., Stroh, V., & Plümer, L. (2013). Model Driven Reconstruction of Roofs from Sparse LIDAR Point Clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *76*, 17–29. https://doi.org/10.1016/j.isprsjprs.2012.11.004

Label and Measure Connected Components in a Binary Image - MATLAB & Simulink. (2023). *Matlab.* https://www.mathworks.com/help/images/label-and-measure-objects-in-a-binary-image.html

Li, W., He, C., Fang, J., Zheng, J., Fu, H., & Yu, L. (2019). Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data. *Remote Sensing*, *11*(4), 403. https://doi.org/10.3390/rs11040403

Li, Z., & Shan, J. (2022). RANSAC-Based Multi Primitive Building Reconstruction From 3D Point Clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *185*, 247–260. https://doi.org/10.1016/j.isprsjprs.2021.12.012

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Nex, F., & Remondino, F. (2012). Automatic Roof Outlines Reconstruction from Photogrammetric DSM. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *I–3*, 257–262. https://doi.org/10.5194/isprsannals-I-3-257-2012

OpenMMLab. (2023). MMSegmentation. *GitHub*. https://github.com/open-mmlab/mmsegmentation

Partovi, Fraundorfer, Bahmanyar, Huang, & Reinartz. (2019). Automatic 3-D Building Model Reconstruction from Very High-Resolution Stereo Satellite Imagery. *Remote Sensing*, 11(14), 1660. https://doi.org/10.3390/rs11141660

Qin, R., Ling, X., Farella, E. M., & Remondino, F. (2022). Uncertainty-Guided Depth Fusion from Multi-View Satellite Images to Improve the Accuracy in Large-Scale DSM Generation. *Remote Sensing, 14*(6), 1309. https://doi.org/10.3390/rs14061309

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241).* Springer International Publishing.

Schuegraf, P., Schnell, J., Henry, C., & Bittner, K. (2022). Building Section Instance Segmentation with Combined Classical and Deep Learning Methods. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *V-2–2022*, 407–414. https://doi.org/10.5194/isprs-annals-v-2-2022-407-2022

Schuegraf, P., Zorzi, S., Fraundorfer, F., & Bittner, K. (2023). Deep Learning for The Automatic Division of Building Constructions into Sections on Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 7186–7200. https://doi.org/10.1109/JSTARS.2023.3296449

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., & Wang, J. (2019). High-Resolution Representations for Labeling Pixels and Regions. *arXiv*. http://arxiv.org/abs/1904.04514.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). Unetformer: A Unet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., ... & Qiao, Y. (2023). Internimage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14408-14419).

Zhang, W., Li, Z., & Shan, J. (2021). Optimal Model Fitting for Building Reconstruction from Point Clouds. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 9636–9650. https://doi.org/10.1109/JSTARS.2021.3110429

Zhou, Z., Fu, C., & Weibel, R. (2023). Move And Remove: Multi-Task Learning for Building Simplification in Vector Maps with A Graph Convolutional Neural Network. *ISPRS Journal of Photogrammetry and Remote Sensing*, *202*, 205–218. https://doi.org/10.1016/j.isprsjprs.2023.06.004