remote sensing

MDPI

Article

LOD2-Level+ Low-Rise Building Model Extraction Method for Oblique Photography Data Using U-NET and a Multi-Decision RANSAC Segmentation Algorithm

Yufeng He [1,2,3], Xiaobian Wu [4], Weibin Pan [5], Hui Chen [6,*], Songshan Zhou [7], Shaohua Lei [4], Xiaoran Gong [4], Hanzeyu Xu [8] and Yehua Sheng [3]

1 Key Laboratory of Poyang Lake Wetland and Watershed Research of Ministry of Education, Jiangxi Normal University, Nanchang 330022, China; yuf.he@jxnu.edu.cn
2 Key Laboratory of Natural Disaster Monitoring, Early Warning and Assessment of Jiangxi Province, Nanchang 330022, China
3 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Key Laboratory of Virtual Geographical Environment of Ministry of Education, School of Geography, Nanjing Normal University, Nanjing 210023, China; shengyehua@jxnu.edu.cn
4 National Key Laboratory of Water Disaster Prevention, Nanjing Hydraulic Research Institute, Nanjing 210029, China; xbwu@nhri.cn (X.W.); shaohualei@nhri.cn (S.L.); 202212480071@nuist.edu.cn (X.G.)
5 North Information Control Research Academy Group Co., Ltd., Nanjing 211106, China; peterj2117@gmail.com
6 Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China
7 Jiangxi Institute of Land Space Survey and Planning, Technology Innovation Center for Land Spatial Ecological Protection and Restoration in Great Lakes Basin, Ministry of Natural Resources, Nanchang 330029, China; 05zss@163.com
8 Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou 225009, China; xuhanzeyu@yzu.edu.cn
* Correspondence: dg20270003@smail.nju.edu.cn

check for updates

Citation: He, Y.; Wu, X.; Pan, W.; Chen, H.; Zhou, S.; Lei, S.; Gong, X.; Xu, H.; Sheng, Y. LOD2-Level+ Low-Rise Building Model Extraction Method for Oblique Photography Data Using U-NET and a Multi-Decision RANSAC Segmentation Algorithm. *Remote Sens.* **2024**, *16*, 2404. https://doi.org/10.3390/rs16132404

Academic Editor: Giuseppe Casula

Received: 4 June 2024
Revised: 26 June 2024
Accepted: 27 June 2024
Published: 30 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract: Oblique photography is a regional digital surface model generation technique that can be widely used for building 3D model construction. However, due to the lack of geometric and semantic information about the building, these models make it difficult to differentiate more detailed components in the building, such as roofs and balconies. This paper proposes a deep learning-based method (U-NET) for constructing 3D models of low-rise buildings that address the issues. The method ensures complete geometric and semantic information and conforms to the LOD2 level. First, digital orthophotos are used to perform building extraction based on U-NET, and then a contour optimization method based on the main direction of the building and the center of gravity of the contour is used to obtain the regular building contour. Second, the pure building point cloud model representing a single building is extracted from the whole point cloud scene based on the acquired building contour. Finally, the multi-decision RANSAC algorithm is used to segment the building detail point cloud and construct a triangular mesh of building components, followed by a triangular mesh fusion and splicing method to achieve monolithic building components. The paper presents experimental evidence that the building contour extraction algorithm can achieve a 90.3% success rate and that the resulting single building 3D model contains LOD2 building components, which contain detailed geometric and semantic information.

Keywords: oblique photography technology; U-NET; LOD2; MVS cloud point; building contour extraction; monolithic building

1. Introduction

Low-rise buildings are a very common type of structure and serve as primary spatial entities and the main locations for human activities, especially in rural and peri-urban areas. High-precision three-dimensional models of low-rise buildings can accurately represent

real-world building geography and attribute information [1]. These models play a significant role in planning, management, construction, and emergency response. Therefore, constructing true 3D models of low-rise buildings has become a fundamental aspect of information construction. These models can facilitate the development of applications such as comprehensive maps, morphology analysis, and high-precision cartography [2]. The technology for modeling low-rise buildings has significant development prospects and has been applied in various fields, offering detailed insights into their structural and spatial characteristics [3].

In recent years, oblique photography technology has provided a more effective way to quickly obtain scene data and automatically construct 3D models of scenes [4]. However, 3D modeling based on oblique photography is limited to regional modeling, resulting in a continuous and irregular Digital Surface Model (DSM) of the entire area. This model does not differentiate between various features such as buildings, ground, and vegetation, nor does it provide information on the geometry or semantics of these features [5–7]. The management and application of 3D models of buildings in the scene can be improved. Current research on constructing monolithic models (LOD1) based on oblique photography involves physically cutting the overall scene DSM and separating different feature objects from it. However, this method has a long preprocessing time, poor modeling effect, and obvious jaggedness. The level of refinement of these 3D models is insufficient for constructing detailed models of buildings [8,9].

The use of convolutional neural networks, particularly fully convolutional neural networks, has enabled the quick, efficient, and accurate extraction of large-scale, multi-scale, and multi-type buildings with the development of deep learning. However, the recovery of the building structure needs to be strengthened. To maintain the spatial details of buildings, deep learning-based methods typically use multi-feature fusion and edge-attention mechanisms. It has been found that the edges of buildings obtained by previous methods are unstable and can be greatly affected by various factors such as imaging conditions, lighting, shadows, and tree occlusion [8,10]. Therefore, it is necessary to investigate the direct application of building contours extracted from the depth model to the point cloud extraction of building monoliths.

Constructing accurate 3D models of buildings involves extracting building details from triangulation networks or point clouds based on the overall structure of the building. This allows for the creation of building models with varying levels of detail. It is particularly critical to realize the fine-grained extraction of building components at the LOD2+ level (LOD2 is a model with a simplified roof shape and where the object's parts can be modeled in multiple semantic classes, e.g., roof and wall. LOD3 is an architecturally detailed model with windows, doors, balconies, etc., being considerably more complex than its preceding counterpart [11,12]).

Current research focuses on multi-level detailed 3D modeling of buildings based on Unmanned Aerial Vehicle (UAV) oblique photography. The research mainly generates LOD2-level 3D models of buildings and completes the reconstruction of the roof structure. The reconstruction of other fine structures, such as eaves, chimneys, skylights, and balconies, has been less studied [10,13,14]. For instance, Dahlke et al. extracted building contours from Multi-View Stereopsis/Stereo (MVS) point clouds and combined them with DSM to determine the building overhang structure. They then used local regression methods to segment the roof in 3D space and area growth methods to obtain the roof topology. Then, a 3D model of the building at the LOD2 level was finally generated. However, this method relies solely on 3D and 2.5D depth information and is more susceptible to interference from vegetation. Li, Nan, Smith, and Wonka [14] utilized an MVS point cloud as experimental data and classified it into buildings, ground, trees, and other categories. The building's roof structure was extracted using the MRF based on the depth maps generated from the statistical analysis of the building's coverage mesh. The building boundary segments were simplified using the Douglas–Peucker polygon approximation algorithm to construct a more regularized LOD2-level 3D model of the building in the regional scene. In the previous

LOD2-level 3D model construction, the building façade model stretches the building height as the building façade after obtaining the building contour, and then mapping the texture on the eave to represent it. Usually, the building contour is obtained as the building contour outside the eaves, and the previous reconstruction method did not consider the actual positional relationship between the building façade and the eaves, and the generated 3D model did not conform to the real building. The method is simple, but it only generates 2.5D models of buildings (LOD 1), which results in the loss of information from the MVS point cloud on the building façade. Therefore, achieving the extraction of eaves and daughter wall parts is more in line with the needs of LOD2-level 3D model construction, which is of great significance for the next step of refined 3D model construction.

The goal of the image segmentation network is to divide the image into several regions with similar properties [15,16], aligning with the concept of extracting building regions from an overall image. Therefore, adopting image segmentation networks for Digital Orthophoto Map (DOM) building object extraction is a promising method. Image segmentation networks represented by E-Net [17], U-NET [18], Seg-Net [19], and Eef-Net [20] are primarily applied to close-range photographic images, such as those in the open-source COCO dataset, which contains 80 categories, including people, bicycles, dogs, handbags, and sports balls. In contrast, oblique photography images acquired by UAV aerial photography contain diverse ground features, such as buildings, vegetation, rivers, and roads, with more distinctive textures. Currently, the effectiveness of image segmentation networks for building extraction from oblique photography images is not well understood. Thus, it is necessary to study the applicability of image segmentation networks on oblique photography image datasets to identify the most suitable network for this task. Additionally, most building façades and roofs are planar and belong to simple geometries that can be represented by a limited number of parameters. To extract building components at the LOD2 level, it is worthwhile to investigate the use of the RANSAC point cloud segmentation algorithm.

This paper proposes a monolithic modeling method for buildings based on the fusion of orthophotos and point clouds, supported by the aforementioned research. After extracting building objects from orthophotos using U-NET, the contour lines are optimized based on the building's main direction and center of gravity This results in a complete and continuous building boundary, which is then used to extract a pure monolithic object from the point cloud of the entire scene. A multi-decision Random Sample Consensus (RANSAC) segmentation algorithm is then utilized to extract building details from the point cloud, resulting in the creation of a detailed single 3D model of the building. The main contributions of this paper are as follows: it proposes a method to extract detailed building models by combining U-NET and RANSAC. The method first extracts continuous-roofed building models using optimized contours and then classifies the building components based on geometric features. This makes it applicable to low-rise buildings in southern China.

2. Background

The oblique photography technique is a less costly alternative to 3D modeling, using LIDAR point clouds to obtain point clouds for 3D modeling [5–7]. Oblique photography is a technique used to acquire vertical and tilted images simultaneously over a large area, providing texture information of building roofs and façades [21]. In recent years, the construction of multi-level detailed 3D models of buildings based on UAV oblique photography has rapidly developed due to continuous improvements in oblique photography technology and systems. Accurately extracting pure building monolithic point clouds from scene point clouds is a prerequisite for building 3D modeling. Previous studies have addressed the challenge of segmenting building monolithic point clouds by using semantic rules and traditional machine learning or deep learning algorithms to distinguish between building and tree morphology in the presence of large noise in MVS point clouds [21,22]. Semantic rule-based methods typically use prior knowledge to detect buildings by leveraging building semantic information [17,23,24]. For instance, Xiao et al. [25] utilized building

façade information from tilted images to detect buildings. The algorithm selects image pairs oriented to the four directions of the building, extracts the building using line information, and obtains the general location of the building. Nex et al. applied a filtering algorithm to eliminate ground points. Then, they used the Normalized Vegetation Index (NDVI) to remove vegetation points. Finally, they projected the remaining points representing the building onto a two-dimensional plane to obtain the building's outline. Algorithms based on traditional machine learning or deep learning typically begin with feature extraction and subsequently employ classifiers to identify buildings based on the extracted features. Zhang et al. [26] used Conditional Random Field (CRF) to combine point cloud and 2D image data for scene-based semantic segmentation. Gerke and Xiao [27] compared supervised classification using random trees as a classification algorithm for voxelized point cloud data and unsupervised classification using Markov Random Fields (MRFs) for graph-cutting algorithms, and the effectiveness of both methods for classifying oblique photogrammetric point clouds, but mainly for relatively simple scenarios. Although the methods mentioned above have yielded positive results, they may not be sufficient to meet the requirements of detailed building models in more complex scenarios. For instance, noise interference is a significant challenge in segmenting more detailed point clouds in city scenes.

The process of 3D modeling with MVS (Multi-View Stereo) point clouds typically involves segmenting a detailed point cloud from a monolithic point cloud using a segmentation method and then using the detailed point cloud for modeling. Li, Nan, Smith and Wonka [14] reconstructed a building using an MVS point cloud based on the Manhattan assumption, which presumes that the main components of a building consist of planes with axes aligned to them. They approximated the building geometry using a set of boxes. The article describes the use of Regularized Markov Random Fields (MRF) to determine the optimal combination of boxes for generating LOD1-level geometrical models of buildings in large-scale scenes. Nan and Wonka [28] proposed a binary labeling strategy for reconstructing from MVS point clouds. This strategy involves intersecting extracted planar elements to generate a set of candidate planes, selecting the best subset of candidate planes through optimization, and finally obtaining the LOD1-level surface model of the building. This method is appropriate for reconstructing planar objects that have been segmented but is not capable of recovering buildings with unique shapes. Wang et al. [1,29] constructed a 3D building model by extracting roof points using the reversed iterative mathematics morphological method, and then point-based segmentation by using smoothness was used for extracting different roof patches. A horizontal direction layer connection was created for different patches, building patches were connected, and the model was built for each building. This approach can be effective in creating LoD2 for complex building layouts but requires high-quality data and no vegetation. Dahlke, Linkiewicz and Meissner [13] presented a workflow of extracting and modeling 3D structures using local regression window, roof segmentation, and reconstruction. Malihi et al. [30] presented a workflow for modeling structures by applying the RANSAC algorithm to detect planes. Symmetry was assumed for building roof planes, and geometrical constraints were used. The 3D models of buildings created using the aforementioned methods typically only meet the detail level of LOD1~2. These models primarily focus on reconstructing the roof portion of the building, with less attention given to other façades and detailed structures such as eaves, balconies, and chimneys. The geometric information obtained directly from the MVS point cloud TIN can result in blurred edges and discontinuities in individual buildings. The building's geometric information can be easily recovered from the MVS point cloud, TIN, and other data. However, this can result in blurred edges and discontinuity within the building.

In previous LOD2-level 3D model constructions, the building façade was represented by stretching the building height after obtaining the building silhouette and then mapping the texture onto the eaves. This reconstruction method does not consider the actual positional relationship between the building façade and eaves, resulting in a generated 3D model that does not accurately reflect the real building. Currently, advancements in

image segmentation provide new opportunities for building component model extraction. Image segmentation networks, which are a type of convolutional neural network (CNN), are usually modified and adjusted based on classification CNNs such as VGGNet [15] and ResNet [16], and play an essential role in image semantic understanding. The goal of image segmentation networks is to classify each pixel of an image and divide the image into several regions with similar properties. This aligns with our goal of extracting building regions from the overall image, making image segmentation networks suitable for Digital Orthophoto Map (DOM) building object extraction. Long et al. [31] proposed the Fully Convolutional Network (FCN), a pioneering work in using CNNs for image segmentation tasks, extending image-level classification to pixel-level classification. Its main contributions include: 1) Replacing fully connected layers with convolutional layers, allowing the network to accept images of arbitrary sizes, achieve pixel-by-pixel classification, and output segmentation maps consistent with the original image size. 2) Using a deconvolutional layer (transposed convolution) to map the generated feature maps back to the original image size. In their paper, bilinear interpolation is used to up-sample the feature maps, and a trained CNN can also be used for the deconvolution operation. Following FCN, a series of image segmentation networks with encoding-decoding structures have emerged. The encoder part extracts features from the input image, resulting in a feature map, and is typically derived from classification CNNs such as VGGNet and ResNet. The decoder part performs pixel-by-pixel classification of the obtained feature map, i.e., image segmentation, and its structure is primarily designed according to the segmentation task, which significantly affects the results of the segmentation network. Currently, four convolutional networks that are commonly used and perform well in the field of image segmentation are U-NET [18], ENet [17], SegNet [19], and ERFNet [20]. However, the accuracy of building segmentation by these four networks for UAV photography scene data needs to be further explored.

The construction of a 3D model of a building at the LOD2 level requires separate modeling of the façade and roof. The point cloud segmentation method involves dividing the entire point cloud into multiple natural surfaces of a single plane. Typical point cloud segmentation methods include the Euclidean clustering-based point cloud segmentation method (He Yifeng et al., 2019), which uses Euclidean distance as a reference for clustering and is mainly suitable for segmenting separated objects; the region-growth-based point cloud segmentation method, which is suitable for segmenting surfaces with small curvature variations [32], and the Random Sample Consensus (RANSAC)-based point cloud segmentation algorithm, which is mainly suitable for layering point cloud data at different levels and is well-suited for detecting objects with special shapes from the overall point cloud. The façade and roof of a building are mostly planar and belong to simple geometric shapes that can be represented by a limited number of parameters. Therefore, the RANSAC point cloud segmentation algorithm is well-suited for extracting the façade and roof surfaces of a building.

To summarize, based on the acquired optimized building contour lines of the scene, a pure single building point cloud was obtained. Using RANSAC point cloud segmentation, the building point cloud was divided into main parts, such as the façade and roof. Subsequently, based on the geometrical characteristics of the wall façade and roof, detailed structures such as eaves and balconies were further extracted. Finally, the LOD2-level building model was constructed.

3. Materials and Methods

3.1. Study Area and Auxiliary Data

This article is about an aerial photography project using drone oblique photography located in Futang Xincun, Ledong Lizu Autonomous County, Hainan Province. Ledong Lizu Autonomous County is a directly governed autonomous county southwest of Hainan Island. The terrain is high in the north and low in the south, with a backdrop of mountains and the sea. The area adjacent to Wuzhishan City and Baisha County in the east and northeast, Sanya City in the southeast, Dongfang City and Changjiang Lizu Autonomous

County in the north, and the South China Sea in the southwest is depicted in Figure 1. The figure shows the approximate location of the survey area, indicated by the blue box. As of the end of 2019, Ledong Lizu Autonomous County had jurisdiction over 11 towns and 188 villages (communities), covering an area of 2765.5 square kilometers and experiencing a tropical monsoon climate.



Figure 1. Diagram of the survey area.

3.2. Scenario Data Collection

The experimental topographic data was acquired in 2020 by DJI's consumer-grade quadcopter Elf 4 Pro, with a Preset flight height of 120m and 80% for Heading overlap (Detailed parameters are shown in Tables 1 and 2). The Context Capture 10.18 oblique photography software was used to generate MVS point cloud and Digital Orthophoto Map (DOM) image data of the scene for subsequent experiments, as shown in Figure 2. The survey area includes various features such as buildings, farmland, and vegetation, with most buildings being 1–4 stories high. To produce the building extraction dataset, the DOM data was cropped into 512 × 512 pixel-sized images and manually labeled using LabelMe, as shown in Figure 3. After data enhancement, which included horizontal flipping, brightness reduction by 50%, contrast increase by 1.5 times, rotation by 180°, and counterclockwise rotation, a total of 3400 images were obtained. The training, validation, and test sets each consist of 2140, 630, and 630 images, respectively. The subsequent 3D modeling data utilizes the last point cloud data from the area covered by the test set.

Table 1. UAV parameters.

Parameter Name	Value
UAV type	Consumer quadcopter Elf 4 Pro
Maximum flight time	About 30min
Gimbal	Gimbal: 3-axis gimbal (pitch, roll, yaw)
Pitch Angle:	−90° to 30
Camera	FC6310S
Lens parameters	FOV 84°8.8 mm/24 mm (35 mm)
Camera focal length	9 mm
Photometry mode	Off-centre averaging
Satellite positioning module:	GPS/GLONASS dual mode
Image sensor	1-inch CMOS, effective pixels 20 million

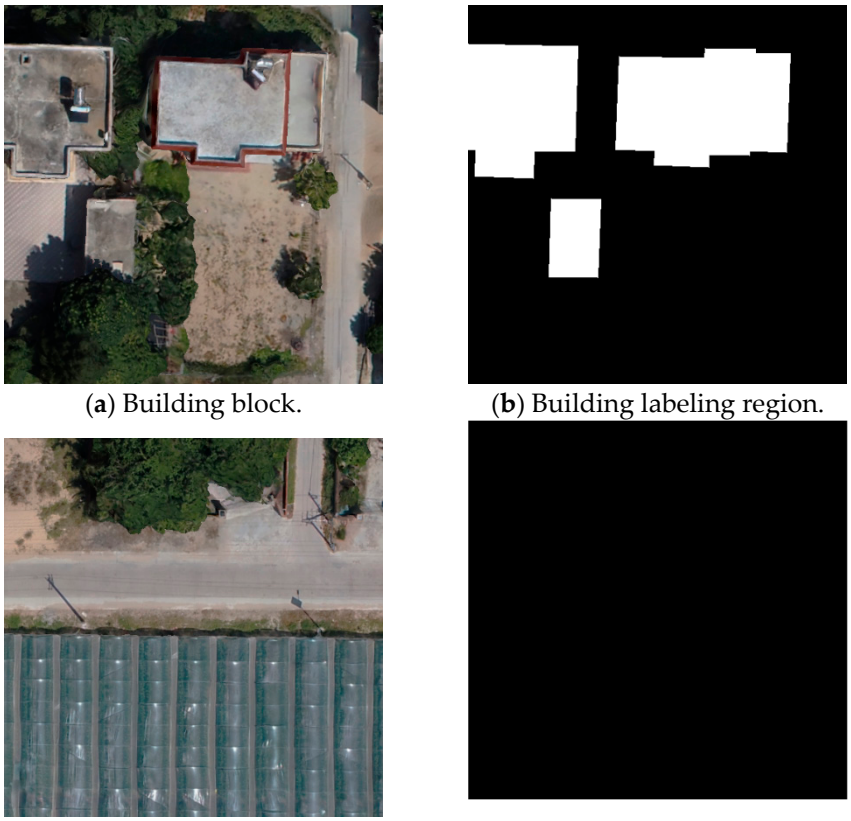
Table 2. UAV Operation parameters.

Operation	
Aerial photography time	2019.10.7
Operation Software	DJI GO
Aerial Route Software	DJI GSpro
Preset Flight Hight	120 m
Number of Strips	8
Heading Overlap	80%
Overlap inside direction	75%
Operation time	3 h



(**a**) Scene MVS point cloud (**b**) Scene DOM Images

Figure 2. MVS point cloud and Digital Orthophoto Map (DOM) image in the scene.



(**a**) Building block. (**b**) Building labeling region.

(**c**) Non-building blocks. (**d**) Non-building blocks label region.

Figure 3. DOM labeled blocks.

3.3. Methods

The paper presents a method consisting of two parts: (1) accurate monolithic point cloud extraction of the building using U-NET, based on the main direction of the building and the center of gravity of the contour, and (2) monolithic building detail point cloud extraction and fine modeling (Figure 4).

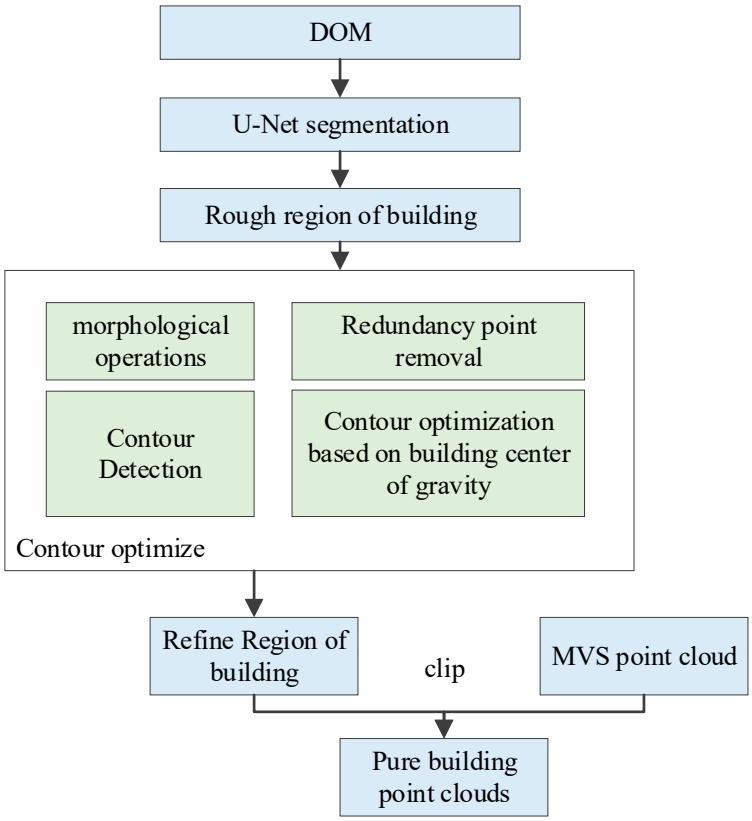


Figure 4. Building monolith point cloud extraction flowchart.

3.3.1. Building Monolith Point Cloud Extraction

(1) U-NET-based building extraction

The UNet network was proposed by Ronneberger et al. (2015) at the University of Fitzbauer, initially as an image segmentation network for biomedical image segmentation. It is an encoding-decoding structure, and the network as a whole is in the shape of a "U", with a concise and stable overall structure. The encoder part of U-NET consists of two 3×3 convolutional layers (ReLU) + 2×2 maximal pooling layers iteratively, with a doubling of the number of channels for each down sampling; the decoder part consists of a 2×2 upsampling convolutional layer (ReLU) + the feature maps output from each layer of the encoder summed with the results of the up-sampling of the decoder layer + two 3×3 convolutional layers (ReLU) iteratively. The main feature of the network is to fuse the feature maps of the previous lower layers by jump-joining them during the up-sampling process performed by the inverse convolution and then up-sampling them again. This process is repeated until the output is obtained. And every pooling layer will generate a new scale so that the structural design can obtain richer contextual information and improve the accuracy of the segmentation network through multi-scale fusion.

For building extraction, images obtained through UAV oblique photography are processed using Context Capture 10.18 software to generate MVS point clouds and DOM data. These data are cropped to produce a building outline extraction dataset, with negative samples added to balance the dataset. The proportion of non-building samples reaches

about 20%. The trained U-NET model is then used to extract building contours from the dataset.

(2) Precisely extract building outlines using the main building orientation and centroid

The building boundaries obtained through U-NET often consist of irregular zigzag lines that do not conform to traditional perceptions of real-world building contours. To obtain an accurate silhouette of the building, an optimization method based on the building's main direction and center of the silhouette is employed.

First, morphological operations are applied to address small voids and unsmooth edges in the building extraction results obtained from U-NET. Next, the edges are extracted, and contours are encoded using the Canny operator [33]. Finally, background patches are then removed based on aspect ratio and degree of rectangularity. However, direct contour detection often generates redundant points in the building contour curve, resulting in large data volumes. To address this issue, the Douglas–Peucker algorithm [33] is utilized to perform polygonal approximation and reduce the number of points. In real geographic scenes, buildings typically consist of several perpendicular walls, requiring only a few right-angled points to represent the building contours accurately. To remove redundancy and straighten the building contour edges, a method based on the principal direction and the contour's center of gravity is proposed. The algorithm consists of two main parts: calculating the building's principal direction and contour center of gravity and optimizing the contour lines.

(1) Calculation of the center of gravity of the main direction and profile of the building [34]

In calculating the main direction of various building contours, a statistical weighting method based on a series of candidate main directions between 0 and 90 degrees is employed, taking into account the orientations of buildings within the study area. The candidate with the highest weight is selected as the main direction of the building contour. This approach comprehensively considers the edges of each building contour while avoiding the impact of long edge errors, thereby providing the necessary support for the subsequent optimization of building contours. The contribution value for candidate direction OA (Figure 5) of the L edge can be calculated as follows:

$$M_L = \left(\frac{\alpha - \beta}{\alpha}\right) \times \mathrm{D}_L \tag{1}$$

where α is the candidate direction angle threshold, β is the difference between L and the candidate direction, and D_L is the length of L. When the difference between the direction of the building edge L and the candidate direction is less than α, this boundary is considered to be contributing to the candidate direction. The contribution values of all the edges of the polygon in the OA candidate directions are accumulated, and the direction with the largest contribution value is selected as the main direction.

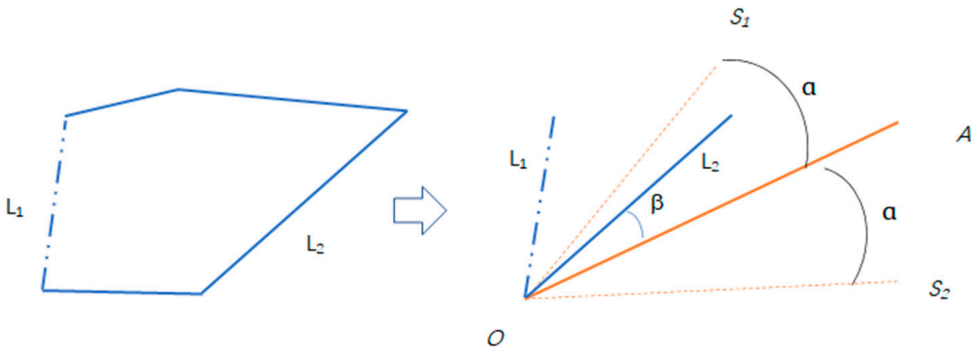


Figure 5. Statistical weighting method to identify the main direction of the building.

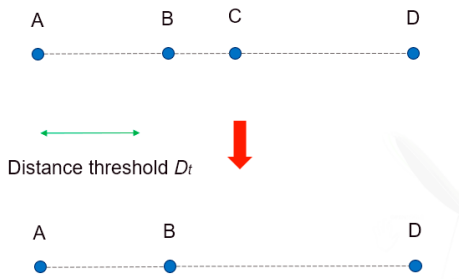
The center of gravity of the building contour is obtained by loading a single building contour line point set and calculating its zero-order and first-order moments. The mathematical expression for the zero-order moment of the binary image (M_{00}) is: $M_{00} = \sum_I \sum_J V(i,j)$. Where $V(i, j)$ refers to the pixel value at (i, j). The first-order moments M_{10} and M_{01} are given by: $M_{10} = \sum_I \sum_J i \times V(i,j)$; $M_{01} = \sum_I \sum_J j \times V(i,j)$.

The center of gravity coordinates (X_C, Y_C) of the binary image are computed from the first-order moments as follows:

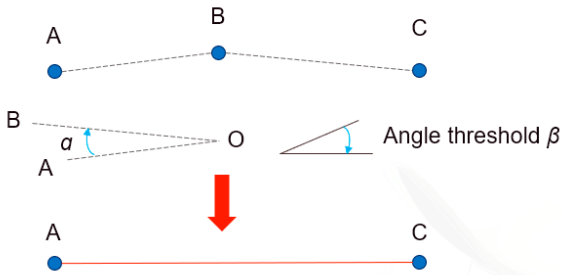
$$\begin{cases} X_C = \dfrac{M_{10}}{M_{00}} \\ Y_C = \dfrac{M_{01}}{M_{00}} \end{cases} \quad (2)$$

(2) Contour line optimization

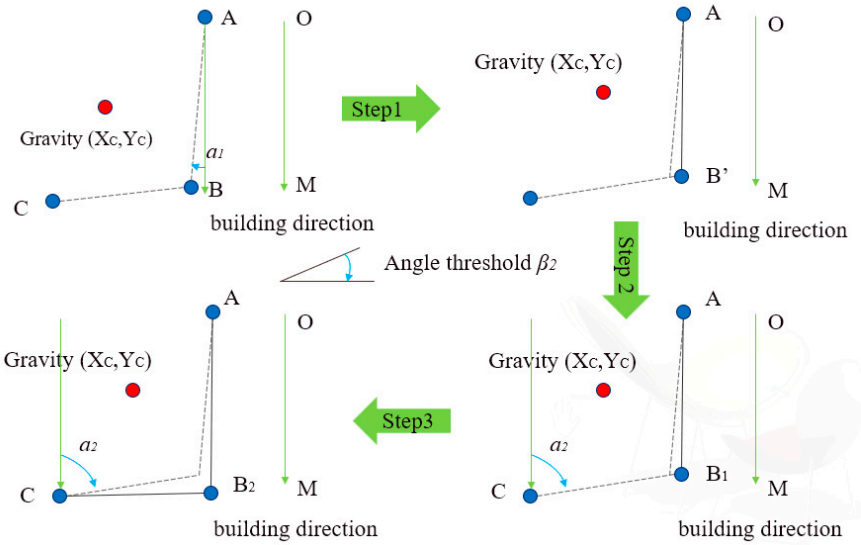
Contour line optimization involves merging proximity points and parallel edges, followed by right-angling feature edges based on the building's principal direction and the center of gravity of the building contour (Figure 6). In Figure 6, points A, B, C, and D represent points on the contour line, while points O and M indicate the direction of the building. The process begins by merging overly close contour points and near-parallel adjacent edges. Subsequently, characteristic edges are right-angled according to the building's principal direction and centroid. These steps are iterated until the number of points in the building contour stabilizes. Finally, the optimized building outline is obtained by connecting the point sequences of the contour lines in a clockwise direction.



(**a**) Merging the over-near contour points.



(**b**) Merging the near-parallel adjacent edges.



(**c**) Right-angle based on the direction of the building and gravity of the building contour.

Figure 6. Contour line optimization method.

(3) Point cloud segmentation of single building

Once the optimized building contour has been obtained, it is used as a mask to extract the point cloud within this range from the overall MVS point cloud of the scene. This process isolates the point cloud of a single building.

3.3.2. Building Fine Modeling

In previous LOD2-level 3D model constructions, the building façade model was obtained by stretching the building contour to the building height and mapping textures onto the eaves. Typically, the building contour obtained includes the eaves and the reconstruction method often fails to consider the actual positional relationship between the building façade and the eaves, frequently overlooking balcony structural information. As a result, the generated 3D model does not accurately reflect the real building. This paper addresses these issues by achieving accurate extraction of eaves, daughter walls, balconies, and façades, which aligns more closely with the requirements for LOD2-level 3D model construction. The overall process is shown in Figure 7.

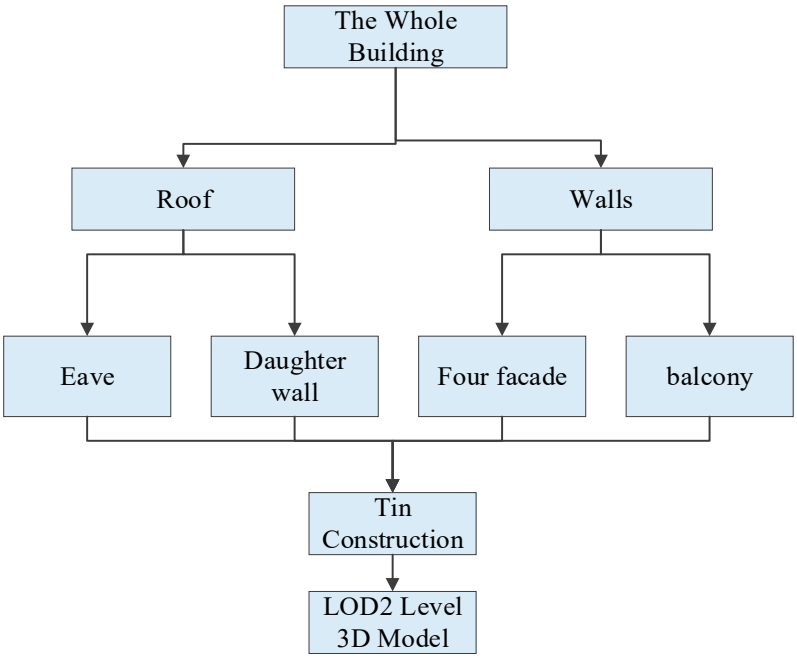


Figure 7. Fine building modeling RANSAC segmentation flowchart.

(1) Roof Extraction

The point cloud segmentation algorithm based on Random Sampling Consistency RANSAC [33], is particularly suitable for layering point cloud data at different levels and detecting objects with special shapes within the entire point cloud. Since the façade and roof of a building are mostly planar and can be represented by simple geometric shapes with a limited number of parameters, the RANSAC algorithm is ideal for extracting these surfaces. Due to the characteristics of UAV oblique camera aerial photography, which captures buildings from various angles, the dense image-matching process can generate point clouds for internal building features such as balconies. Consequently, the RANSAC segmentation method may produce internal point clouds and complex segmentation results in addition to the main plane segments.

(2) Eaves and pat wall extraction

Eaves are the edges of roofs that extend in all directions, and daughter walls are short walls around the roof's edge. The extraction process involves calculating the average height Z_{roof} of the segmented roof point cloud, projecting all point cloud heights to this average height Z_{roof}, and then traversing the eave/daughter wall portion of the MVS point cloud to classify points. Points with heights less than or equal to Z_{roof} are identified as eave

point clouds, while those with heights greater than Z_{roof} are classified as daughter wall point clouds.

(3) Balcony Extraction

First, calculate the minimum and maximum coordinates of the eaves point cloud data: X_{house_Min}, Y_{house_Min}, Y_{house_Max}, Z_{house_Max}. Then, calculate the average x-coordinate of the left façade points and the rear façade points X_{left_avg}, Y_{behind_avg}, as well as the minimum coordinates of these points (Z_{left_min}, Z_{behind_min}). Due to the existence of internal points on the front façade, not all points lie in a single plane. Therefore, the point cloud data of the front façade is read sequentially, and the minimum y-coordinate of the front façade is recorded as Y_{befor_min};

The RANSAC-based point cloud segmentation does not attribute balconies outside the front façade to the front façade but instead includes them with other surfaces, resulting in disruptive segmentation. To isolate the balcony points, use the minimum y-coordinate Y_{befor_min} and the lowest z-coordinate Z_{house_Min} from the eaves as constraints. Traverse the original point cloud of the building and store points whose y-coordinate is less than Y_{befor_min} and whose z-coordinate is less than Z_{house_Min} separately. This process results in the point set of the balcony on the front façade of the building. In the figure, the red area represents the point cloud of the front façade, with the y-coordinate adjusted to Y_{befor_min} (the result of projecting onto the same plane), and the brown part represents the point cloud of the balcony on the front façade.

(4) Façade Extraction

Building façades can be classified into two types: flat façades and complex façades, which may include balconies. To ensure consistency, the façade point clouds can be projected onto the same plane using the average x or y coordinates of the façade point cloud and the lowest z-coordinate of the eaves as references. When a building has structures such as indoor balconies, the UAV will capture the interior from multiple angles during aerial photography. This results in a large number of internal point clouds when generating the MVS point cloud, leading to voids due to the lack of points on certain façades during RANSAC point cloud segmentation. Additionally, when a building's façade consists of multiple smaller façades, balconies, and other detailed structures, the RANSAC segmentation of the entire building cannot distinguish between individual façades.

To address this issue, this paper proposes a multi-decision façade segmentation algorithm for complex façade segmentation. If there is a gap in the exterior of a building, the lowest eaves coordinate is used as a constraint. The interior point cloud is then segmented using RANSAC to obtain point clouds that are close to different parts of the exterior, and these point clouds are projected onto their respective parts of the exterior to fill the gaps. When a building's façade consists of multiple smaller façades, balconies, and other detailed structures, a multi-level façade segmentation algorithm is utilized. The same wall is usually the same color but different from windows, doors, and other detailed structures. Using color differences can extract windows and doors with a different depth than the walls.

The algorithm first traverses the façade point cloud to obtain the maximum and minimum depth values along the normal line direction of the façade, denoted as $depth_{max}$ and $depth_{min}$, respectively. The hierarchical interval is calculated as $D_{lay} = (depth_{max} - depth_{min})/n$, where n is the number of layers (typically 1–3 for low-rise buildings). To obtain the façade point cloud, the algorithm traverses the point cloud to collect different layers, denoted as p (p_1, p_2, …, p_n). Then, it uses the RGB information of the point cloud mapping to construct the spatial-grayscale similarity quantization value M for each layer, dividing the sub-façades of the wall based on both color and wall depth differences. If the M value of a point cloud layer is less than or equal to Dm, the point cloud sets are merged. After traversing all the sets, the point cloud with different façade segmentations can be obtained.

$$M = \frac{D_{lay}}{AVGDRGB}\left(\sum_0^t \frac{p(r,g,b)}{t} - \sum_0^o \frac{q(r,g,b)}{o}\right)^2 + \left(\sum_0^t \frac{p(x,y,z)}{t} - \sum_0^o \frac{q(x,y,z)}{o}\right)^2 \quad (3)$$

Here, t and o represent the number of point clouds in different point cloud layers, p(r, g, b), q(r, g, b) are the RGB values of the point clouds in different layers, p(x, y, z) and q(x, y, z) represent the spatial locations of the point clouds in different layers, and $AVGDRGB$ represent the average RGB color distance between point cloud layer t and o.

(5) Fine-grained 3D modeling of buildings

The point clouds extracted from the roof surface, eaves, daughter walls, balconies, and façades are utilized to generate TIN 3D models. Those TIN models were then merged and integrated to create a refined 3D model of a single building with component information at the LOD2 level.

4. Results and Discussion

The computer environment used for the experimental 3D modeling includes a Windows 10 operating system, Intel Core i7-8700 3.2 GHz CPU, and 16 GB RAM. For U-NET model training, a Linux operating system was utilized, equipped with an NVIDIA GeForce GTX 1080Ti graphics card and Pytorch 1.1.0. The building contour extraction results are evaluated using three common metrics that accurately reflect the segmentation performance: average Intersection over Union (IoU) [35], precision [36], recall [37], and $F1$ are used as the evaluation basis, calculated as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \quad (4)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} \times \text{recall}} \quad (7)$$

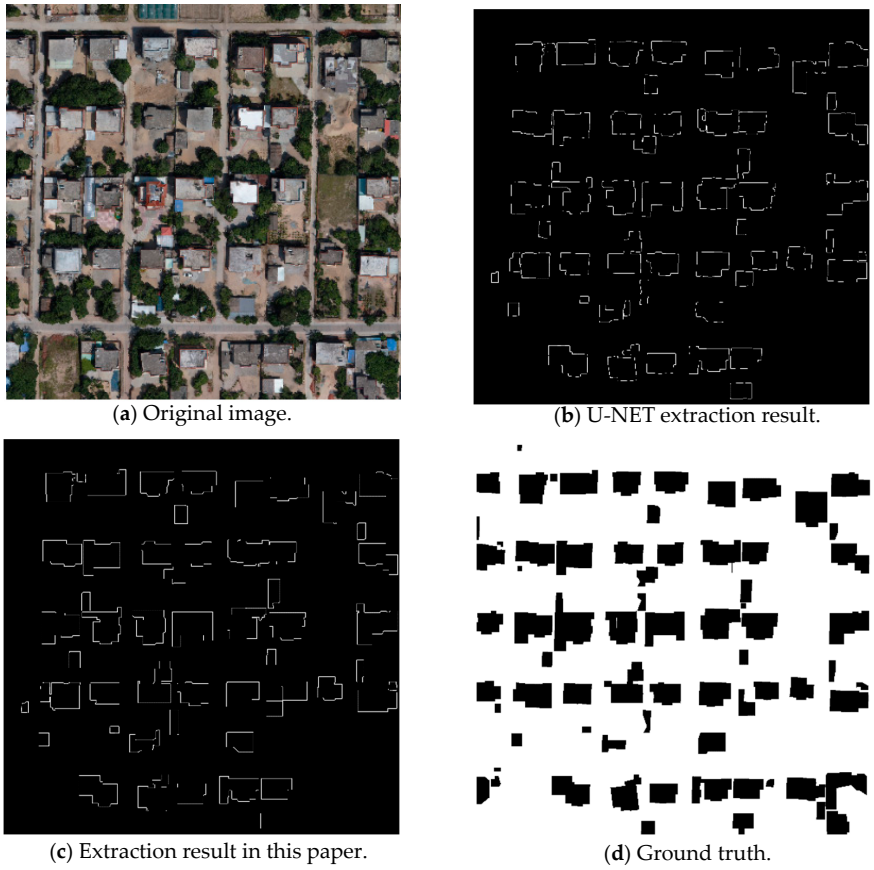
where TP (True Positive) is a pixel of a building correctly identified as a building by the network model; FP (False Positive) is a pixel of other features incorrectly identified as a building by the network model; FN (False Negative) is a pixel of a building incorrectly identified as other features by the network model; and TN (True Negative) is a pixel of other features correctly identified as other features by the network model.

4.1. Building Contour Accurate Extraction Results

The accuracy metrics of the building extraction method used in this paper are shown in Table 3. It can be seen that IoU is improved by 5.84% after applying the contour optimization algorithm compared to just using U-NET. Compared to ENET, SegNet, and ERFNet, the method presented in this paper improves IoU accuracy by 6.79%, 10.21%, and 10.24%, respectively. Figure 8 demonstrates that after contour optimization, the integrity of the extracted building contours is better and more regularized.

Table 3. Building extraction results.

Method	IoU (%)	Precision (%)	Recall (%)	F1 (%)
ENet	80.10	83.73	94.13	88.63
SegNet	80.13	78.77	94.00	85.71
ERFNet	83.55	85.59	94.42	89.79
U-NET	86.50	87.69	94.56	91.00
U-NET + contour optimized (our method)	**90.34**	**95.04**	**94.81**	**92.63**



(**a**) Original image. (**b**) U-NET extraction result.

(**c**) Extraction result in this paper. (**d**) Ground truth.

Figure 8. Extraction results of this paper's method.

4.2. Fine Modeling Results of Single Building Point Cloud

4.2.1. Single Building Point Cloud Segmentation Results

The building contour lines obtained can eliminate vegetation interference and preserve the distinct edges of the buildings. This allows for effective handling of vegetation interference adjacent to the buildings during the extraction of the point cloud of a single building, resulting in a pure single building point cloud, as depicted in Figure 9.



(**a**) Scene MVS point cloud. (**b**) Single building point cloud.

Figure 9. Segmentation result of the single building point cloud.

4.2.2. Detailed Extraction

The point cloud extraction results for the roof, daughter wall and eaves of the building are shown in Figure 10 (the roof point cloud is shown in yellow in the figure, and the daughter wall and eaves point clouds are shown in pink and red, respectively). In order to prove the effectiveness of this paper's method, the widely used Ransac algorithm is used in the façade and balcony extraction phase to compare with this paper's algorithm. *Dm* and layers *n* threshold in this case are set as 0.5 m, 2 separately, meaning that the max count of sub-façades is 2 and the façade depth less than 0.5 cannot be distinguished. The balcony segmentation results are shown in Figure 11, and the façade point cloud segmentation results are shown in Figure 12.



(**a**) The whole building.



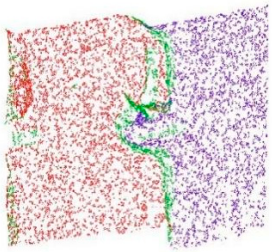
(**b**) Roof point cloud.



(**c**) Daughter wall (red) and eaves (pink).

Figure 10. Roof, pat, eaves point cloud segmentation result.

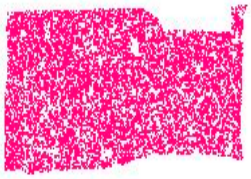






(**a**) Point cloud of building with balcony. (**b**) RANSAC segmentation result. (**c**) Balcony extraction.

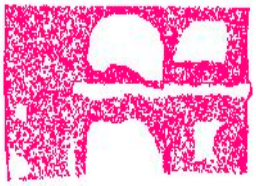
Figure 11. Balcony point cloud segmentation results.



(**a**) Flat façade RANSAC segmentation.



(**b**) Flat façade segmentation results.



(**c**) Complex façade RANSAC segmentation.



(**d**) Complex façade RANSAC segmentation result

Figure 12. Façade point cloud segmentation results.

Figure 10 shows that the segmentation algorithm proposed in this paper can more completely segment the point clouds of the roof, eaves, and daughter walls of the building. Figures 11 and 12 demonstrate that the RANSAC point cloud segmentation method does not perform as well on complex façades, especially when the building façade includes balconies and uneven structures, compared to flat façades. The extraction of these façades results in a large number of voids, making it impossible to obtain a complete façade point cloud. Additionally, RANSAC segmentation does not utilize the color information of the MVS point cloud.

To avoid these issues, it is recommended that a different segmentation method be used. Constructing the 3D model directly from the façades obtained by point cloud segmentation may result in gaps and non-closure between façades. This paper presents a method to prevent the loss of detail points in point cloud segmentation results. The proposed method ensures that the segmentation results are smoother and retain the color information of the MVS point cloud, better reflecting the real positional relationships between the building façades.

4.2.3. Building Refinement 3D Model Construction

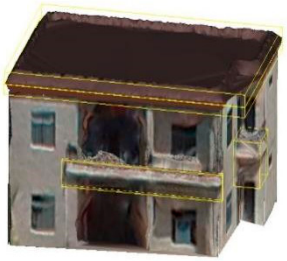
Figure 13 illustrates the comparison between the 3D model of the building intercepted from the scene 3D model generated directly by the Context Capture 10.18 oblique camera software and the 3D model constructed based on the method proposed in this paper. The proposed method can extract and reconstruct a fine 3D model of the building at the LOD2 level using the overall scene data. The resulting model has a flat and clear structure, accurately representing the roof surface and several façades of the building. In contrast, the 3D model of the building generated from the scene 3D model has uneven façades, is fixed to the ground, and cannot differentiate between various façades and roof surfaces.



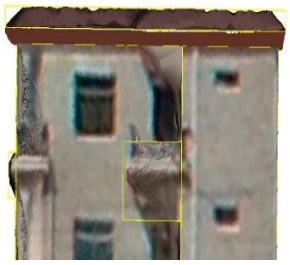
(**a**) 3D building model cutting from a big scene.



(**b**) LOD2-level 3D model in this paper



(**c**) Front view of LOD2-level 3D model.



(**d**) Right view of LOD2-level 3D model.

Figure 13. Building 3D model construction in the paper.

5. Conclusions

This paper proposes a deep learning-based method for extracting building monolithic point clouds to obtain a fine 3D model. The method utilizes a multi-decision point cloud segmentation algorithm to capture detailed structural information. The aim is to address the challenge of obtaining high-precision fine 3D models from scene point clouds. To improve the precision of the neural network extraction of the building contour, we propose a contour

line optimization method based on the main direction of the building and the center of gravity of the contour. This method achieves an IoU of 90.34% for the building contour extraction result. Using the optimized contour lines, we obtain a pure single building point cloud from the overall point cloud of the scene. The building is then segmented into detailed structure point clouds, including the roof, façade, eaves, daughter wall, and other components, using the multi-decision RANSAC point cloud segmentation algorithm. These structures are converted into triangular mesh models and combined to create a single, high-quality LOD2-level 3D model of the building.

However, the method proposed in this paper has some shortcomings. For instance, the parameter threshold settings largely rely on human a priori knowledge. Future research should explore how to adaptively set experimental parameter thresholds by combining experimental data and scenarios to obtain optimal parameters. Additionally, the current approach segments the building point cloud model by continuous roofs, making it particularly suitable for low-rise building model extraction. Further research and development in remote sensing large models and a priori knowledge mapping methods are expected to enhance the accuracy and applicability of the building model and component model extraction, extending beyond low-rise houses. Reducing human involvement to enhance automation will be the focus of future research.

Author Contributions: Conceptualization, Y.H., W.P., S.L. and Y.S.; methodology, Y.H., W.P. and S.L.; validation, X.W. and X.G.; writing—original draft, Y.H. and H.C.; writing—review and editing, S.Z. and H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. 42301533, 42101384), the Natural Science Foundation of Jiangsu Province (BK20210043), and the Open Research Fund of Key Laboratory of Reservoir and Dam Safety Ministry of Water Resources (YK323011).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Florent, L. Some New Research Directions to Explore in Urban Reconstruction. In Proceedings of the 2015 Joint Urban Remote Sensing Event (JURSE), Lausanne, Switzerland, 30 March–1 April 2015.
2. Bo, M.; Ban, Y.; Harrie, L. A Multiple Representation Data Structure for Dynamic Visualisation of Generalised 3d City Models. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 198–208.
3. Gao, M.; Xu, X.; Klinger, Y.; Van Der Woerd, J.; Tapponnier, P. High-Resolution Mapping Based on an Unmanned Aerial Vehicle (Uav) to Capture Paleoseismic Offsets Along the Altyn-Tagh Fault, China. *Sci. Rep.* **2017**, *7*, 8281. [CrossRef] [PubMed]
4. Agarwal, S.; Furukawa, Y.; Snavely, N.; Simon, I.; Curless, B.; Seitz, S.M.; Szeliski, R. Building Rome in a Day. *Commun. ACM* **2011**, *54*, 105–112. [CrossRef]
5. Kristian, S.; Guarnieri, P.; Stemmerik, L. From Oblique Photogrammetry to a 3d Model–Structural Modeling of Kilen, Eastern North Greenland. *Comput. Geosci.* **2015**, *83*, 120–126.
6. Guler, Y.; Selcuk, O. 3d City Modelling with Oblique Photogrammetry Method. *Procedia Technol.* **2015**, *19*, 424–431.
7. Sun, Y.; Sun, H.; Yan, L.; Fan, S.; Chen, R. Rba: Reduced Bundle Adjustment for Oblique Aerial Photogrammetry. *ISPRS J. Photogramm. Remote Sens.* **2016**, *121*, 128–142. [CrossRef]
8. Xiao, J.; Gerke, M.; Vosselman, G. Automatic Detection of Buildings with Rectangular Flat Roofs from Multi-View Oblique Imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2010**, *38*, 251–256.
9. Xie, F.; Lin, Z.; Gui, D.; Lin, H. Study on Construction of 3d Building Based on Uav Images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *39*, 469–473. [CrossRef]
10. Lin, J.; Jing, W.; Song, H.; Chen, G. Esfnet: Efficient Network for Building Extraction from High-Resolution Aerial Images. *IEEE Access* **2019**, *7*, 54285–54294. [CrossRef]
11. Kutzner, T.; Chaturvedi, K.; Kolbe, T.H. Citygml 3.0: New Functions Open up New Applications. *PFG–J. Photogramm. Remote Sens. Geoinf. Sci.* **2020**, *88*, 43–61. [CrossRef]
12. Filip, B.; Ledoux, H.; Stoter, J. An Improved Lod Specification for 3d Building Models. *Comput. Environ. Urban Syst.* **2016**, *59*, 25–37.
13. Dahlke, D.; Linkiewicz, M.; Meissner, H. True 3d Building Reconstruction: Façade, Roof and Overhang Modelling from Oblique and Vertical Aerial Imagery. *Int. J. Image Data Fusion* **2015**, *6*, 314–329. [CrossRef]
14. Li, M.; Nan, L.; Smith, N.; Wonka, P. Reconstructing Building Mass Models from Uav Images. *Comput. Graph.* **2016**, *54*, 84–93. [CrossRef]

15. Simonyan, K. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
17. Adam, P.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
18. Olaf, R.; Fischer, P.; Brox, T. U-NET: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015.
19. Vijay, B.; Kendall, A.; Cipolla, R. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
20. Eduardo, R.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. Erfnet: Efficient Residual Factorized Convnet for Real-Time Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272.
21. Zhang, X.; Sun, J.; Gao, J. An Algorithm for Building Exterior Facade Corner Point Extraction Based on Uav Images and Point Clouds. *Remote Sens.* **2023**, *15*, 4166. [CrossRef]
22. Liang, H.; Lee, S.-C.; Bae, W.; Kim, J.; Seo, S. Towards Uavs in Construction: Advancements, Challenges, and Future Directions for Monitoring and Inspection. *Drones* **2023**, *7*, 202. [CrossRef]
23. De Farias, T.M.; Roxin, A.; Nicolle, C. A Rule-Based Methodology to Extract Building Model Views. *Autom. Constr.* **2018**, *92*, 214–229. [CrossRef]
24. Claudio, M.; Gholamzadehmir, M.; Daniotti, B.; Pavan, A. Semantic Enrichment of Bim: The Role of Machine Learning-Based Image Recognition. *Buildings* **2024**, *14*, 1122. [CrossRef]
25. Xiao, J.; Gerke, M.; Vosselman, G. Building Extraction from Oblique Airborne Imagery Based on Robust Façade Detection. *ISPRS J. Photogramm. Remote Sens.* **2012**, *68*, 56–68. [CrossRef]
26. Zhang, R.; Candra, S.A.; Vetter, K.; Zakhor, A. Sensor Fusion for Semantic Segmentation of Urban Scenes. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015.
27. Gerke, M.; Xiao, J. Supervised and Unsupervised Mrf Based 3d Scene Classification in Multiple View Airborne Oblique Images. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *2*, 25–30. [CrossRef]
28. Nan, L.; Wonka, P. Polyfit: Polygonal Surface Reconstruction from Point Clouds. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
29. Wang, Y.; Xu, H.; Cheng, L.; Li, M.; Wang, Y.; Xia, N.; Chen, Y.; Tang, Y. Three-Dimensional Reconstruction of Building Roofs from Airborne Lidar Data Based on a Layer Connection and Smoothness Strategy. *Remote Sens.* **2016**, *8*, 415. [CrossRef]
30. Malihi, S.; Valadan Zoej, M.J.; Hahn, M.; Mokhtarzade, M.; Arefi, H. 3d Building Reconstruction Using Dense Photogrammetric Point Cloud. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 71–74. [CrossRef]
31. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
32. George, V.; Gorte, B.G.H.; Sithole, G.; Rabbani, T. Recognising Structure in Laser Scanner Point Clouds. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2004**, *46*, 33–38.
33. John, C. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *6*, 679–698.
34. Yin, S.; Yan, X.; Yan, X. Simplification method of building polygon based on feature edges reconstruction. *Acta Geod. Cartogr. Sin.* **2020**, *49*, 703–710. [CrossRef]
35. Alberto, G.-G.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
36. Liu, B.; Wang, X.; Dixit, M.; Kwitt, R.; Vasconcelos, N. Feature Space Transfer for Data Augmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
37. Cheng, J.; Deng, C.; Su, Y.; An, Z.; Wang, Q. Methods and Datasets on Semantic Segmentation for Unmanned Aerial Vehicle Remote Sensing Images: A Review. *ISPRS J. Photogramm. Remote Sens.* **2024**, *211*, 1–34. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.