IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 1, 2022, accepted February 15, 2022, date of publication February 18, 2022, date of current version March 1, 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3152744*

# 3D Urban Buildings Extraction Based on Airborne LiDAR and Photogrammetric Point Cloud Fusion According to U-Net Deep Learning Model Segmentation

**PENGCHENG ZHANG[1,2], HUAGUI HE[1,2], YUN WANG[3], YANG LIU[1,2], HONG LIN[1,2], LIANG GUO[1,2], AND WEIJUN YANG[1,2]**

[1]Guangzhou Urban Planning and Design Survey Research Institute, Guangzhou 510060, China
[2]Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, Guangzhou 510060, China
[3]College of Forestry, Beijing Forestry University, Beijing 100083, China

Corresponding author: Huagui He (hehuagui_updsrig@163.com)

This work was supported in part by the Key Area Research and Development Program of Guangdong Province under Grant 2020B0101130009; in part by the Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, under Grant 2020B121202019; in part by the Smart Guangzhou Spatiotemporal Information Cloud Platform Construction under Grant GZIT2016-A5-147; in part by the Gao Fen Project of China under Grant 30-Y20A34-9010-15/17; and in part by the Construction of Public Service Platform: Building Information Modeling (BIM) and City Information Modeling (CIM)-Based Integrated Perspective under Grant TC19083WA.

**ABSTRACT** The LiDAR and photogrammetric point clouds fusion procedure for building extraction according to U-Net deep learning model segmentation is provided and tested. Firstly, an initial geo-localization process is performed for photogrammetric point clouds generated using structure-from-motion and dense-matching methods. Then, point cloud segmentation is carried out based on U-Net deep learning model. The precision of the U-Net model for buildings extraction reachs 87%, with F-score of 0.89 and IoU of 0.80. It is shown that the U-Net method is effective for high-resolution image extraction. The detailed information can accurately be identified and extracted, such as vegetation located between buildings and roads. After segmentation, each chunk of the LiDAR and photogrammetric point clouds are finely registered and merged based on the iterative closest point algorithm. Finally, the fused point clouds are obtained. It shows that the structure and shape of the buildings could be delineated from the fused point clouds when both enough ground points and a higher point density are available. Furthermore, color information improves both visualization effect and properties identification. The experiments are conducted to extract individual buildings from three types of point clouds in three plots. A DoN (Difference of Normals) approach is used to isolate 3D buildings from other objects in densely built-up areas. It shows that most building extraction results have a Precision $>$ 0.9 and favorable Recall and F-score values. Although the LiDAR extraction results have some advantages over the photogrammetric and fused ones in terms of Precision, the Recall and F-score results appear best for the fused point clouds. It shows that the fused data contains a high point density and RGB color information and could improve the building extraction.

**INDEX TERMS** Building extraction, point clouds, U-Net, deep learning, segmentation, difference of normals.

## I. INTRODUCTION

The extraction and identification of 3D urban buildings have become a crucial issue in many applications, such as urban building database updating, city management, disaster assessment, digital mapping, transportation planning, cadastral, and telecommunication network management [1], [2]. However, extracting 3D building information through field surveys is labor-intensive and time-consuming, yet usually unavoidable. Consequently, relatively little information is obtained on the 3D building information updates compared to the rapid rate of urbanization, especially in developing countries.

The associate editor coordinating the review of this manuscript and approving it for publication was Yong Yang.

Recently, the technologies such as remote sensing, computer vision, and machine learning have provided opportunities and prospects for building automation extraction. Building extraction from remote sensing data has become a valuable alternative to field surveys because of its extensive coverage and regular data updates at a low cost [3]–[5]. However, for a long time, automatic approaches of building detection from remote sensing data have been complex if not impossible due to scene complexity, incomplete extraction, and sensor dependencies, especially in big cities with dense buildings [6]–[8]. Methodologically, *building extraction* refers to dividing a given dataset into non-overlapping homogeneous regions and recognizing the buildings from those regions. Various image-recognition algorithms have been proposed based on pixel, geometric structure, and object-based identification [9]–[11]. Nowadays, deep learning techniques are widely used because they can use the number of layers of the network to represent a multi-layered characteristic [12], [13]. Through multi-layered learning, the original input was mapped to multi-variable labels, thereby achieving accurate classification. Krizhevsky *et al.* [14] proposed the convolutional neural network Alexnet. The accuracy of the network on the ImageNet dataset reached 84.6%. The success of Alexnet enlarged the application of convolutional neural networks in the field of computer vision [14]. Long *et al.* [15] proposed a fully convolutional neural network FCN containing convolutional layers and an end-to-end learning method. And the accuracy rate of the FCN model on the PASCAL VOC data set has reached 90.3% [15]. Ronneberger *et al.* [16] proposed a new network structure, U-Net, which adds a pooling layer after each convolution operation. The model was obtained on the medical image data set. It has been verified that the speed of network convergence was better than FCN, and the accuracy of segmentation has reached 92.03% [16]. He *et al.* [17] proposed Resnet, which could control the ratio of the output of the previous layer through the loss function, rather than training all the results of the previous layer. And, the accuracy on the ImageNet dataset reached 96.43% [17]. Research shows that the deep learning algorithm effectively solves the problems of complex high-resolution image building extraction. It is of great significance for high-precision urban ecological environment monitor. However, it is still challenging to reach a satisfactory effect in the dense building extraction because there are usually obstructions from surrounding buildings and high trees. Nevertheless, it is virtually unavoidable, especially in very high-spatial resolution remote sensing images. Moreover, it is almost impossible to extract 3D building information from images.

With the development of 3D scanners and other point clouds generation techniques, 3D measurements are generally performed for building extraction. Light detection and ranging (LiDAR) is a common method of obtaining point cloud datasets due to its accuracy, speed, and ability to capture millions of points in a very short time. Now, it is possible to calculate digital terrain models (DTM), digital surface models (DSM), and three-dimensional (3D) models of buildings from a georeferenced point cloud. Other building modeling metrics, such as building shape and voxelization, can also be effectively analyzed. However, airborne LiDAR acquisitions remain very costly, especially in big cities with complex surroundings[18]. So it is a significant barrier to its widespread application, especially for local city management and building modeling studies based on annual or more frequent observations with numerous points at small sites or sampling plots [19]. Moreover, various building types in urban areas make it difficult to detect buildings in complex scenes automatically. Many existing algorithms are intricate and often fail in complex inner-city environments without enough points [20]. Aerial photogrammetry is used to decrease the cost and 3D point clouds can be produced by applying the SfM (Structure-From-Motion) method over large areas [21]–[24]. The techniques have also been regard as a viable alternative to LiDAR for 3D forestry applications and already proven successful in extracting tree height, individual plant structure, and other 3D modeling metrics in forest surveys [25], [26].

Nevertheless, the advantages for 3D measurements and modeling exists in metrological and reliability. For example, the reliability of photogrammetric point clouds for building extraction needs to be evaluated because of noisy points [27]. To this end, clear accuracy statements and evaluations must be carried out before it is applied. Several recent publications have compared LiDAR and imaging techniques regarding accuracy, resolution, and dense 3D reconstructions of small scenes [19], [28]. Combining LiDAR and photogrammetric point clouds may improve building metrics extraction accuracy [29]. However, few experiments were conducted when LiDAR and photogrammetric fused point clouds were applied to extract accurate dense urban 3D buildings.

Here, we demonstrate and evaluate a practical urban 3D building extraction method by fusing two different point clouds according to U-Net deep learning model segmentation. The study focuses on the practical procedure to extract 3D buildings with reliable accuracy. So the UNET model was used to extract building polygons from images. And the point clouds fusion strategy was applied in each polygon to add the point cloud density for future extraction. Firstly, the study area and remote sensing dataset, including high resolution image and 3D point cloud datasets from both LiDAR scanning and image-based matching techniques are introduced in Section 2 (STUDY AREA AND DATA). Then, the U-Net convolutional neural network is used for image segmentation. After that, the point cloud fusion method is demonstrated in Section 3 (METHOD). In Section 4 (RESULT), the image classification and point clouds fusion results are presented, and the performance of buildings extraction is evaluated. Finally, the procedure is discussed in section 5 (DISCUSSION), and some initial conclusions are made regarding applying fused point clouds to urban 3D building extraction in section 6 (CONCLUSION).

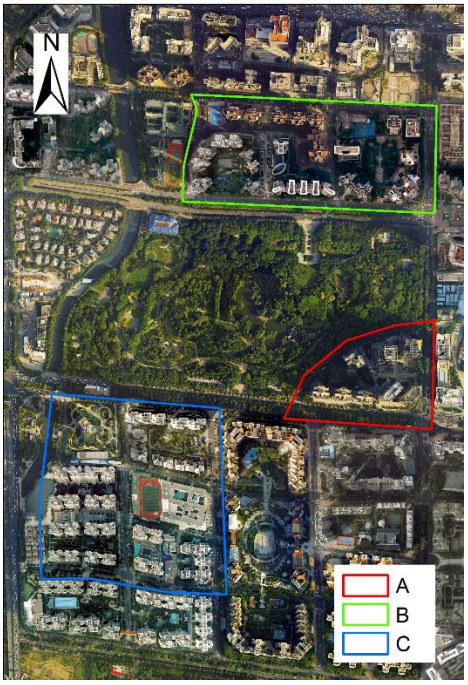


**FIGURE 1.** Aerial photograph of the three study areas around Zhujiang Park.

## II. STUDY AREA AND DATA

The study area, Zhujiang New Town (ZNT), covering an area of 6.44 km$^2$, is located in a central business district in the city of Guangzhou, southern China (23° 06' N, 113° 45' E) (Figure 1). Guangzhou presented a high urban density and complex buildings, especially in ZNT, where many high-end residential complexes and Zhujiang Park exist. It also contains a continuous open plaza that extends approximately 1.5 km from Huangpu Avenue to the Pearl River. The plaza incorporates underground shopping malls, vehicular tunnels, and a people mover system. Other city landmarks, like Opera House, Children's Palace, Library, Museum, supertall Twin Towers, and Canton Tower (the tallest structure in Guangzhou), also lie in this area.

In ZNT, LiDAR and aerial oblique photogrammetric image datasets were obtained. Due to the cost of LiDAR and oblique photogrammetry data acquisition, only a few data with the same coverage were tested in the experiments. Three plots labeled A, B, and C were selected in this area for 3D building-extraction purposes. Aerial images and a LiDAR dataset both covering the ZNT area were available. The specifications of the camera and LiDAR settings could be found in [5].

## III. METHOD

In this paper, a practical procedure was proposed to fuse the LiDAR data and the photogrammetric point clouds generated from the aerial images to improve the extraction of dense urban 3D buildings. Firstly, the orthophoto was generated by the photos from five aerial cameras. The orthophoto dataset

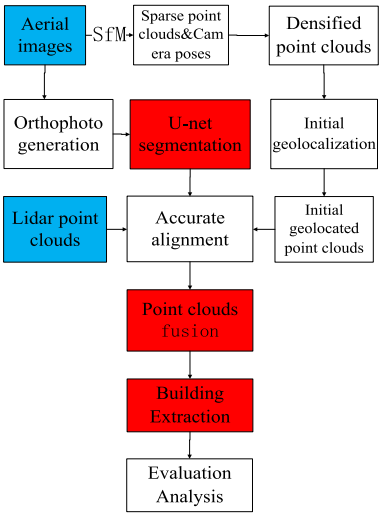


**FIGURE 2.** Workflow of the study.

is divided into the training set, validation set, and test set. The training set and the validation set are used to train the U-Net deep learning model, and the test set is used to test the image segmentation ability of the trained model. To increase the number of the training dataset, the training set and the validation set are processed for data enhancement. The training data and corresponding labels are put into the U-Net model during the training step, and the network parameters are constantly updated to reduce the network loss value to the convergence value. The test image is input to the trained U-Net convolutional neural network model for pixel-by-pixel prediction during the prediction process. The remote sensing image classification and extraction result map are obtained, and its accuracy is analyzed. Then, the fused point clouds are generated from the LiDAR and photogrammetric point cloud in each segmentation polygon. In this step, the camera pose for each picture (indicating motion) and 3D photogrammetric point cloud (indicating structure) were generated from aerial images using the SfM algorithm [30], [31]. The depth of each pixel in the picture was computed to densify the point clouds using the patch-match dense matching method [27]. After that, an initial geo-localization of the dense point cloud was performed by aligning the SfM camera positions to the imaging metadata from GPS to reduce significant differences in rotation, scale, and translation between the two kinds of point clouds. Using segmentation results from the U-net prediction in this area, both LiDAR and photogrammetry point clouds were segmented. The entire area was separated into several chunks by segmentation polygons. For each piece, accurate registration of two 3D point clouds was carried out based on the ICP (Iterative Closest Point) algorithm. Finally, the fused point clouds were carried out to extract the building using the difference of normals (DoN) approach, and the results were quantitatively and qualitatively compared. An overview of the proposed method is illustrated in Figure 2.

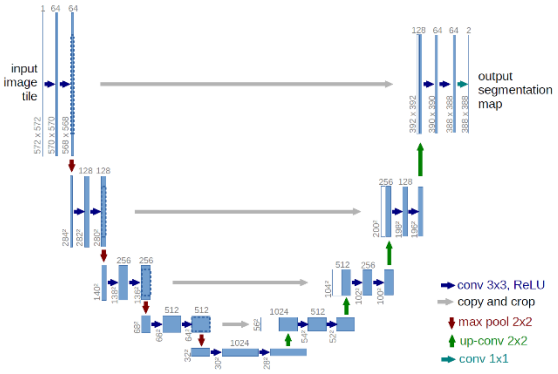


**FIGURE 3.** U-Net model structure diagram [16].

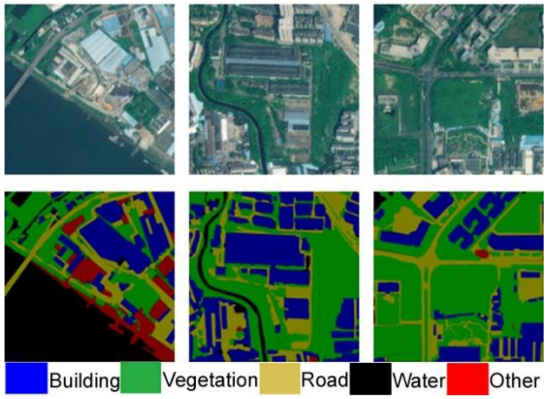


**FIGURE 4.** Data set images and their labels.

### A. U-NET TRAINING AND SEGMENTATION

The U-Net convolutional neural network is U-shaped, as shown in Figure 3, including left and right parts. The left part is down-sampling to extract shallow features of the input image, obtain context information and reduce the image size. The right part is symmetrical up-sampling to get the deep feature information of the picture and achieve precise positioning. The middle part is the jump connection. The feature map generated during the downsampling process is saved and spliced with the feature map of the corresponding upsampling layer in a jump connection mode to reduce the resolution reduction caused by the maximum pooling layer. The design can improve the accuracy of segmentation.

The visual interpretation method is applied for the dataset tiles to establish a labeled dataset using the labelme (Image Polygonal Annotation with Python). The image is divided into five categories: buildings, vegetation, road, water, and others (Figure 4), which is the ground truth label dataset. Due to the accuracy and generalization requirements of the deep learning model, the training dataset and corresponding labels are processed for data enhancement, including image translation, flipping, color transformation, and adding noise.

The network model of this research is implemented using the Pytorch1.6 deep learning framework and the computational platform used in the experiment with a Ubuntu 18.04 (x86_64) operating system, Intel(R) Xeon(R) CPU E5-2603 v3 @ 1.60GHz (12 cores), 125G RAM and NVIDIA TITAN X (Pascal). After many experiments, the model batch size is finally set to 8, the number of iterations is 400. Adam is used as the algorithm optimizer, and ReLu is used as the non-linear activation function with an initial learning rate is 0.0001.

To quantitatively analyze the accuracy of model segmentation, IoU (Intersection over Union), Precision, Recall, and F-score are taken as evaluation indicators. IoU is the most commonly used in the quantitative evaluation of image segmentation, which indicates the difference between the prediction result and the true value. The higher the value, the more accurate the prediction. The formula is as below:

$$\text{IoU} = |A \cup B| / |A \cap B| \tag{1}$$

where A represents the prediction result, and B represents the true value. Precision refers to the proportion of correctly classified pixels to the total number of pixels that are predicted to be true. Recall refers to the ratio of the number of correctly classified pixels to the total number of actually correct pixels. F-score is an important indicator used to measure the accuracy, which considers Precision and Recall. Its maximum value is one, and its minimum value is 0. The higher the F-score, the better the accuracy of the obtained prediction map. The formula is as below:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

### B. POINT CLOUDS FUSION AND 3D BUILDING EXTRACTION

After segmentation, the classification raster was converted to vector polygons based on GDAL (Geospatial Data Abstraction Library). In each polygon, registration is performed to find the relative positions and orientations between LiDAR and photogrammetric point clouds and merge them to extract subsequent buildings. The ICP algorithm was carried out to finely align the photogrammetric point cloud to LiDAR data in each chunk. The ICP algorithm iteratively assigns correspondence based on a closest-distance criterion and finds the rigid transformation using a least-squares approach. After registration, the photogrammetric point cloud was merged into the LiDAR dataset in each chunk. The ground points and non-ground points were distinguished by using cloth simulation filter (CSF) algorithm[32].

To compare the building extraction among LiDAR, photogrammetric and fused point clouds, a point cloud segmentation strategy called the DoN was tested [33]. In the extraction of building roofs and facades, the response of the normal across two different radii: $r_1 < r_2$ are compared. Firstly, the

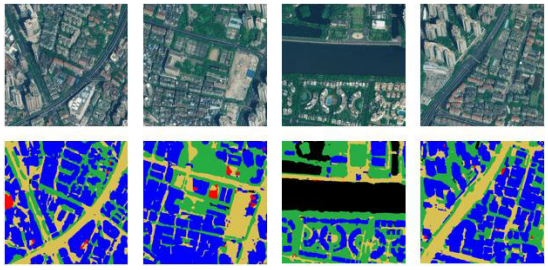

**FIGURE 5.** U-Net model segmentation result (Legend as figure 4).

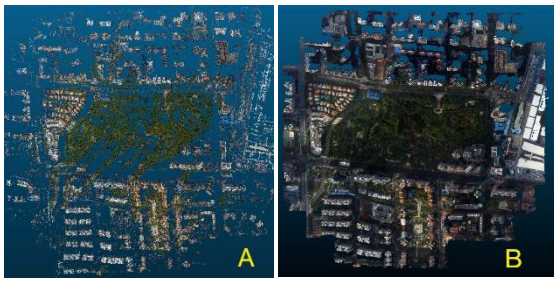


**FIGURE 6.** The sparse point clouds (A) and dense point clouds (B).

**TABLE 1.** U-Net model accuracy.

| | Precision | Recall | F-score | IoU |
|---|---|---|---|---|
| Building | 0.91 | 0.88 | 0.89 | 0.80 |
| Vegetation | 0.91 | 0.93 | 0.92 | 0.84 |
| Road | 0.77 | 0.81 | 0.79 | 0.64 |
| Water | 0.87 | 0.67 | 0.76 | 0.60 |
| Other | 0.11 | 0.16 | 0.13 | 0.06 |

DoN is first calculated for each point within its multi-scale neighbors to separate the points based on the surface normal difference. Then, the DoNs of all points are classified with a simple Euclidean distance threshold-based clustering algorithm [34]. Finally, the planar and nonplanar segments are segmented based on their distances and connectivity. Theoretically, it is challenging to extract geometrical complex building architecture with complex normals. However, the difficulty was significantly reduced by image segmentation. The calculation of the DoN operator $\Delta_{\hat{n}}$ for any point, $p$ in a point cloud $P$ is defined as:

$$\Delta_{\hat{n}}\,(\mathrm{p}, r_1, r_2) = \frac{\hat{n}\,(p, r_1) - \hat{n}\,(p, r_2)}{2} \tag{5}$$

where $r_1, r_2 \in \mathrm{R}, r_1 < r_2$, and $\hat{n}\,(p, r)$ is the surface normal estimation at point $p$, given the support radius $r$. In our building extraction, the DoN vectors were selected based on their magnitudes $\|\Delta_{\hat{n}}\,(\mathrm{p})\|$.

## IV. RESULTS

### A. U-NET SEGMENTATION RESULTS

After iterative training, the model finally converged and achieved 96% classification accuracy on the training set. The segmentation result of the U-Net model is shown in Figure 5.

It is shown that the U-Net method is effective for high-resolution image extraction. The detailed information can accurately be identified and extracted, such as vegetation located between buildings and roads. The model can also effectively distinguish the differences between the buildings and roads accurately. The relatively accurate outline boundaries and internal details of the target object could be identified, and there is no apparent confusion between the categories.

The overall classification accuracy of the U-Net model is 87% (Table 1). Among them, vegetation is the highest, with an F-score of 0.92, IoU of 0.84, and buildings with an F-score of 0.89 and IoU of 0.80.

### B. LiDAR AND PHOTOGRAMMETRIC POINT CLOUD FUSION

The whole SfM process took 34 mins, including 6 mins for feature extraction, 4 mins for matching, and 24 mins for bundle adjustment to solve the re-projection formulas. All images were calibrated in the process, and 310, 050 sparse point clouds were generated. The dense matching used to calculate the depth of each pixel took 1 hour 9 mins and generated 28, 357, 085 points for the entire area (Figure 6).

Points generated using photogrammetric techniques typically contain noise and errors [35]. Here, a statistical analysis method was used to trim the outliers [36]. Assuming that the resulting distribution is Gaussian with a mean value of 50 and a standard deviation of 1, all points with mean distances outside of an interval defined by the global mean and standard deviation distance can be considered outliers and are trimmed from the dataset. The point density decreased by about 10% after the noise-removal process. After noise removal, there were about 8 points/m2 for LiDAR and 21 points/m2 for photogrammetric point clouds. For point cloud segmentation, the segmentation polygons from U-Net predictions were firstly selected. Thus, according to the polygons, the point cloud data could be clipped into several chunks.

Before fine registration, there were about 2 meters of positional drift in the photogrammetric geo-localization compared to LiDAR. In our study, most of the main streets selected had widths greater than 5 meters so that the buildings could be entirely selected in the LiDAR and photogrammetric point clouds. This provided good initial correspondence between the two point clouds for their further accurate alignment. Due to the overlaps between the LiDAR and photogrammetric point clouds in 3D space, the ICP algorithm was applied to register them in each segmentation chunk. In this process, the LiDAR point cloud data was used as a reference for its high geometric accuracy. For the urban building area, the transformation matrix between the LiDAR and photogrammetric point clouds was calculated. After accurate alignment by applying the transformation matrix, the pho-


**FIGURE 7.** Photogrammetric point clouds (A) and fused point clouds (B) of Plots B.

**TABLE 2.** Densities and spaces for the LIDAR and photogrammetric point clouds (pts/m2).

| Plot | LiDAR | Photo | Fused |
|---|---|---|---|
| A | 6.13(0.40) | 25.49 (0.20) | 30.64 (0.18) |
| B | 10.45(0.31) | 21.06 (0.22) | 28.96 (0.19) |
| C | 7.98(0.35) | 17.57 (0.24) | 23.11 (0.21) |

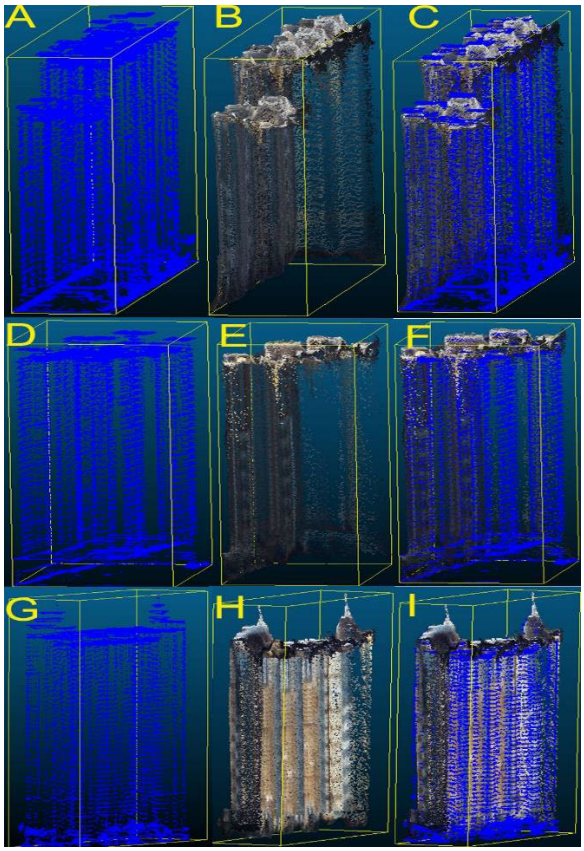


**FIGURE 8.** The LiDAR (A, D, G), photogrammetry (B, E, H), and fused point clouds (C, F, I) for individual buildings.

togrammetric point clouds and LiDAR data were aligned well (Figure 7). Due to the relatively good density of points in the photogrammetric point clouds, ICP is considered to have achieved a fairly good result in our experiments.

Although there was a high point density (Table 2) in the photogrammetric point clouds, it shows that most of the points existed on the building surface, and the point clouds were clustered. It can be seen from the photogrammetric point clouds that there are some holes in the area (Figure 7A). In this kind of point distribution, it is difficult to distinguish the ground from the photogrammetric point clouds compared to LiDAR data which contains a more even distribution. However, with fused data, the holes are filled with LiDAR points (Figure 7B), so the ground points can be easily classified. In our study, the ground points were classified after being distinguished by the CSF algorithm. In Plot A, there are 2,046,970 points, comprised of 398,249 LiDAR points and 1,648,721 photogrammetric point cloud points. There are 201,507 ground points, including 10% of the total points. In Plot B, there are 4,398,137 points, comprised of 1,505,114 LiDAR points and 2,893,023 photogrammetric point cloud points. There are 371,879 ground points, comprising 9% of the total points. Plot C has a total of 4,326,956 points, comprised of 1,450,378 LiDAR points and 2,876,578 photogrammetric point cloud points. There are 566,744 ground points, comprising 14% of the total points.

For comparison, several individual buildings were selected in each plot (Figure 8). For LiDAR data, the points were evenly distributed in the horizontal and vertical directions. For photogrammetric data, most of the point cloud lies on the building surface, for which there was a high point density. And some building facades were missing in the vertical direction. LiDAR provided enough reliable ground points for the fused data, essential for calculating the DTM and subsequent point heights. The missing facades were also supplemented in the fused data. So, the structure and shape of the buildings could be delineated from the fused point clouds with a higher point density. Furthermore, color information is another advantage, which improves the visualization and aids the identification of building types and other properties.

### C. BUILDING EXTRACTION BASED ON POINT CLOUDS

In the DoN implementation for building extraction, the small radius ($r_1$) and large radius ($r_2$) were set to 1 m and 10 m, respectively. Such DoN parameters settings have been found to provide sound isolation of points in urban building areas [9]. The roofs and facades were clustered with the scene based on the Euclidean cluster. For segmentation, a threshold value of 0.1 was applied for building roofs and facades and 0.4 for trees. It shows that most buildings could be successfully extracted from fused point clouds (Figure 9).

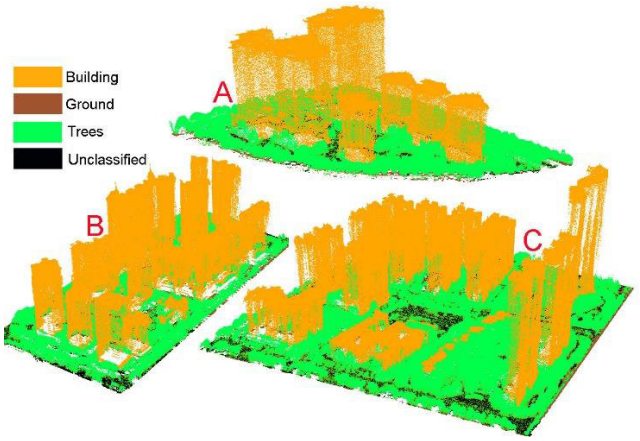


**FIGURE 9.** Building extraction based on fused point clouds for sections A, B, and C.

**TABLE 3.** Classification results and performance metrics.

| Data | Section | TP | FP | FN | Precision | Recall | *F*-score |
|---|---|---|---|---|---|---|---|
| LiDAR | A | 60182 | 720 | 4734 | 0.99 | 0.93 | 0.96 |
| | B | 336721 | 13035 | 44137 | 0.96 | 0.88 | 0.92 |
| | C | 358365 | 10035 | 81036 | 0.97 | 0.82 | 0.89 |
| Photogramm etry | A | 369684 | 30667 | 5692 | 0.92 | 0.98 | 0.95 |
| | B | 1044366 | 28126 | 97624 | 0.97 | 0.91 | 0.94 |
| | C | 627395 | 123814 | 69678 | 0.84 | 0.90 | 0.87 |
| Fused | A | 453712 | 34146 | 10938 | 0.93 | 0.98 | 0.95 |
| | B | 1432031 | 89983 | 40273 | 0.94 | 0.97 | 0.96 |
| | C | 1238631 | 72206 | 64705 | 0.94 | 0.95 | 0.95 |

To evaluate the results of the DoN segmentation with the three types of data, the building truth was labeled manually using the GIS processing software QGIS (v3.4). The Precision, Recall, and *F*-score performance metrics were calculated for each of the three selected sections.

Table 3 illustrates the results of our evaluation in the form of a *Precision/Recall/F-score* over ground truth objects. For each cluster, the point classification was compared with each of the ground truth labels. It was found that the majority of the results had a good performance for the three types of data. The LiDAR extraction results have some advantages over the photogrammetric and fused ones in terms of *Precision*. However, the *Recall* and *F*-score results appear best for the fused point clouds. It shows that the fused data combines the advantages of both LiDAR and photogrammetric point clouds and could improve the building extraction.

## V. DISCUSSION

This study proposed a practical way to fuse LiDAR data with photogrammetric point clouds to improve building extraction. A laser can penetrate the vegetation canopy to provide accurate ground geometric measurements [37]. The LiDAR data also contain a lot of laser signal parameters, such as pulse number and scan angle, which are very important for the extraction of specific parameters. In contrast, photogrammetric point clouds are indirectly obtained from photographs and primarily lie on the object's surface, making it difficult to judge height in the absence of sufficient ground point data. However, its high point density and color attributes are essential for shape delineation and object recognition. So, theoretically, the fusion of the two types of datasets can potentially facilitate building extraction.

In our study, photogrammetric point clouds are generated using SfM and a dense matching method. First, an initial geo-localization process is performed according to the GPS metadata in the camera images to reduce significant differences in rotation, scale, and translation. It was found that there were about 2 meters of positional drift in photogrammetric geo-localization compared to the LiDAR data.

Then, the study area was segmented into several polygons based on U-Net deep learning model. The overall classification accuracy of the U-Net model is 87%. This model has a high classification accuracy for buildings with an F-score of 0.89 and IoU of 0.80. In fact, the object-oriented segmentation method, FCN, and Resnet deep-learning models were also tested in our study. The overall classification accuracy of object-oriented segmentation is 75%, and FCN is 82%. Although Resnet model could reach 90% accuracy in the commercial cluster, it is not a cost-performance way to be implemented in our server. So the U-Net model was selected for image segmentation in our study, especially it can completely and accurately extract house information. Through quantitative analysis, it is verified that the deep learning U-Net method has advantages in the extraction of high-resolution remote sensing images with complex backgrounds, and the classification results are accurate and reliable. Therefore, the urban building areas could be chosen entirely without loss using either LiDAR or photogrammetric point clouds.

Moreover, the segmentation provided sufficient initial correspondences in each chunk. So, the ICP algorithm was performed to register the two sets of point clouds in each chunk. In this process, correspondences were iteratively assigned based on a closest-distance criterion, and the matrix was solved through a least-squares approach until a local minimum was reached. Due to its high geometric accuracy, the LiDAR point cloud data was used as a reference during the assignments. For the building area, a transformation matrix between LiDAR and photogrammetric point clouds was calculated. Although the density of the photogrammetric point clouds was high, they always included many outliers and noisy points. Many factors, such as feature selection, correspondence matching, and patch-matching, affect the quality of the clouds. Most of the points represented the building surface for the photogrammetric point clouds, and the point clouds were clustered according to the building distribution. For this kind of point distribution, it isn't easy to distinguish ground points from point clouds. Due to its penetration ability, LiDAR data exists a more even distribution with enough ground points. So, ground and lower-canopy points supplemented each other to create a more even distribution for fused data, making it easier to classify ground points to calculate the DTM and subsequent point heights.

In this study, the DoN segmentation strategy was carried out to extract buildings. Selecting the parameters $r_1$ and $r_2$ for DoN may cause a significant difference and affect the normal calculation, even the segmentation results. In the extraction of building roofs and facades, the responses of the normals across two different radii $r_1 < r_2$ were detailed compared. Our experiments with different buildings found that using $r_1 = 1$ m and $r_2 = 10$ m gives satisfactory results in this area [33]. These small radii with enough neighboring points can provide a reasonable estimation of the surface normal. The metrics FP and FN could be better balanced in this situation. Such DoN parameter settings provided sound isolation of points in urban LiDAR scenes [5]. After DoN calculation, the Euclidean cluster extraction method was performed to classify the scenes. For each point clouds cluster, a threshold value of 0.1 m was applied for building roof and facades planar fitting. A low measure threshold (0.1) yielded horizontal and planar surfaces that were mainly classified as buildings. On the other hand, a high value (0.4) produced rough surfaces, which indicates that the points represent trees in our study area.

The segmentation quality was quantitatively evaluated on the LiDAR, photogrammetric and fused point cloud datasets. Building roofs and facades were automatically segmented from these datasets. It was shown that the majority of the results had a precision $> 0.9$, and the Recall and F-score results appear favorable. Overall, the LiDAR extraction results have some advantages over the photogrammetric and fused ones for the *precision*. However, for the comprehensive F-score metric, fused point cloud appears best extraction result. The comprehensive analysis showed that the fused point clouds maintain reliable geometric accuracy and provide detailed shape information. So, theoretically, it combines the advantages of both LiDAR and photogrammetric point clouds and could improve the individual building extraction in urban areas. It is better to obtain high spatial resolution/point density data to carry out the image segmentation and 3D building extraction. However, it is always a tradeoff between high-performance data acquisition and cost. The accuracy of UNET could translate further into the accuracy of 3D building extraction because the point cloud fusion was conducted in each polygon obtained from UNET. If the polygon accuracy is bad or there is no polygon available, the two point clouds dataset couldn't be fused very well because the initial geo-localization of the dense point cloud from SfM couldn't match the lidar point cloud with high position accuracy. However, it is difficult to evaluate the translation accuracy between the polygon extraction accuracy using UNET and 3D extraction from point clouds. So only the separate accuracy was evaluated in each step.

## VI. CONCLUSION

This study provides a practical procedure for aligning and fusing LiDAR and photogrammetric data to create a single point cloud for extracting urban buildings. Since the photogrammetric point cloud uses a local coordinate system and LiDAR data uses a georeferenced coordinate system, there are large translation, rotation, and scale differences. Therefore, geo-localization is performed to approximately transform the photogrammetric point cloud data into georeferenced coordinates, which reduces these differences and allows alignment of the point clouds. Then, an urban building map from U-Net segmentation was utilized for point cloud segmentation. The two types of point cloud data and polygons all use the same coordinate system. The U-Net convolutional neural network model was applied for image segmentation. The model extracts image context information through continuous dual convolutional and pooling layers, and uses deconvolution and jump connections to achieve precise positioning. To prevent the model from overfitting and enhance the robustness and generalization of the model, a BN layer and data enhancement operations are added to the network. The experimental results prove that the deep learning algorithm based on U-Net has a higher classification accuracy for high-resolution remote sensing images, especially for buildings and vegetation. The overall accuracy is 87%. In addition, U-Net can automatically acquire deep semantic features through the multi-dimensional feature learning of convolutional neural networks, effectively reducing the classification process's noise. So, in recent years, U-Net model was successfully used for image segmentation [38], [39]. The LiDAR and photogrammetric point clouds were also segmented based on the polygons from U-Net model segmentation results. Each chunk of the LiDAR and photogrammetric point clouds were finely registered and merged based on the ICP algorithm. Hence, the two types of data were accurately co-registered, and the photogrammetric point cloud was incorporated into the LiDAR dataset for a given street polygon.

Applying the DoN methods to LiDAR, photogrammetric, and fused data of dense urban areas qualitatively demonstrated the effectiveness of fused point clouds in classifying buildings. It shows that photogrammetric point clouds provide lower geometric accuracy than LiDAR ones. However, the point density of photogrammetric point clouds is much higher and may include much more redundant data. In our study, the fused point clouds combine the advantages together. So, it reached higher accuracy when using DoN for detecting these nonplanar points with the appropriate size of the neighborhood. At the same time, color information could be used in the future to improve the accuracy of extracting the related metrics. Future work should exploit the DoN scale operator with color information for building extraction and integrate it with cluster-recognition methods. The comprehensive analysis shows that the fused point clouds maintain reliable geometric accuracy and provide detailed shape and color information. Therefore, the proposed procedure provides a practical way to conduct individual building extraction in urban areas.

## REFERENCES

[1] B. Benciolini, V. Ruggiero, A. Vitti, and M. Zanetti, "Roof planes detection via a second-order variational model," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 101–120, Apr. 2018.

[2] S. Du, Y. Zhang, Z. Zou, S. Xu, X. He, and S. Chen, "Automatic building extraction from LiDAR data fusion of point and grid-based features," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 294–307, Aug. 2017.

[3] X. Lai, J. Yang, Y. Li, and M. Wang, "A building extraction approach based on the fusion of LiDAR point cloud and elevation map texture features," *Remote Sens.*, vol. 11, no. 14, p. 1636, Jul. 2019.

[4] R. W. Kulawardhana, S. C. Popescu, and R. A. Feagin, "Fusion of LiDAR and multispectral data to quantify salt Marsh carbon stocks," *Remote Sens. Environ.*, vol. 154, pp. 345–357, Nov. 2014.

[5] W. Yang, Y. Liu, H. He, H. Lin, G. Qiu, and L. Guo, "Airborne LiDAR and photogrammetric point cloud fusion for extraction of urban tree metrics according to street network segmentation," *IEEE Access*, vol. 9, pp. 97834–97842, 2021.

[6] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 1, pp. 43–63, 2007.

[7] F. Rottensteiner and C. Briese, "A new method for building extraction in urban areas from high-resolution LiDAR data," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 34, pp. 295–301, Sep. 2001.

[8] S. L. Ullo, C. Zarro, K. Wojtowicz, G. Meoli, and M. Focareta, "LiDAR-based system and optical VHR data for building detection and mapping," *Sensors*, vol. 20, no. 5, p. 1285, Feb. 2020.

[9] S. Raschka, *Python Machine Learning*. Birmingham, U.K.: Packt, 2015.

[10] M. Brédif, O. Tournaire, B. Vallet, and N. Champion, "Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework," *ISPRS J. Photogramm. Remote Sens.*, vol. 77, no. 1, pp. 57–65, 2013.

[11] M. Baatz and A. Schape, "Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation," *Angew. Geographische Inf.-Verarbeitung*, vol. 12, pp. 12–23, Jan. 2000.

[12] Y. Liu, J. Zhou, W. Qi, X. Li, L. Gross, Q. Shao, Z. Zhao, L. Ni, X. Fan, and Z. Li, "ARC-Net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020.

[13] A. Abdollahi, B. Pradhan, S. Gite, and A. Alamri, "Building footprint extraction from high resolution aerial images using generative adversarial network (GAN) architecture," *IEEE Access*, vol. 8, pp. 209517–209527, 2020.

[14] A. Krizhevsky and I. G. S. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, p. 25.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[16] O. Ronneberger and P. T. F. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[18] T. L. Erdody and L. M. Moskal, "Fusion of LiDAR and imagery for estimating forest canopy fuels," *Remote Sens. Environ.*, vol. 114, no. 4, pp. 725–737, Apr. 2010.

[19] J. P. Dandois and E. C. Ellis, "High spatial resolution three-dimensional mapping of vegetation spectral dynamics using computer vision," *Remote Sens. Environ.*, vol. 136, pp. 259–276, Sep. 2013.

[20] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 93, pp. 256–271, Jul. 2014.

[21] H. Aliakbarpour, V. B. S. Prasath, K. Palaniappan, G. Seetharaman, and J. Dias, "Heterogeneous multi-view information fusion: Review of 3-D reconstruction methods and a new registration with uncertainty modeling," *IEEE Access*, vol. 4, pp. 8264–8285, 2016.

[22] A. G. Melo, M. F. Pinto, L. M. Honorio, F. M. Dias, and J. E. N. Masson, "3D correspondence and point projection method for structures deformation analysis," *IEEE Access*, vol. 8, pp. 177823–177836, 2020.

[23] F. Remondino, M. G. Spera, E. Nocerino, F. Menna, and F. Nex, "State of the art in high density image matching," *Photogramm. Rec.*, vol. 29, pp. 144–166, Jun. 2014.

[24] S. Shen, "Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1901–1914, May 2013.

[25] D. Dey, L. Mummert, and R. Sukthankar, "Classification of plant structures from uncalibrated image sequences," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2012, pp. 329–336.

[26] W. Tao, Y. Lei, and P. Mooney, "Dense point cloud extraction from UAV captured images in forest area," in *Proc. IEEE Int. Conf. Spatial Data Mining Geographical Knowl. Services*, Jun. 2011, pp. 389–392.

[27] S. Shen, "Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1901–1914, May 2013.

[28] E. P. Baltsavias, "A comparison between photogrammetry and laser scanning," *ISPRS J. Photogramm. Remote Sens.*, vol. 54, nos. 2–3, pp. 83–94, Jul. 1999.

[29] L. Guo, X. Deng, Y. Liu, H. He, H. Lin, G. Qiu, and W. Yang, "Extraction of dense urban buildings from photogrammetric and LiDAR point clouds," *IEEE Access*, vol. 9, pp. 111823–111832, 2021.

[30] P. Moulon, P. Monasse, and R. Marlet, "Adaptive structure from motion with a contrario model estimation," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 257–270.

[31] P. Moulon, P. Monasse, and R. Marlet, "Global fusion of relative motions for robust, accurate and scalable structure from motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3248–3255.

[32] W. Zhang, J. Qi, P. Wan, H. Wang, D. Xie, X. Wang, and G. Yan, "An easy-to-use airborne LiDAR data filtering method based on cloth simulation," *Remote Sens.*, vol. 8, no. 6, p. 501, Jun. 2016.

[33] Y. Ioannou, B. Taati, R. Harrap, and M. Greenspan, "Difference of normals as a multi-scale operator in unorganized point clouds," in *Proc. 2nd Int. Conf. 3D Imag., Model., Process., Visualizat. Transmiss.*, Oct. 2012, pp. 501–508.

[34] R. R. Bogdan, "Semantic 3D object maps for everyday manipulation in human living environments," *KI-Künstliche Intelligenz*, vol. 24, no. 4, pp. 345–348, Aug. 2010.

[35] K. Zhang, S.-C. Chen, D. Whitman, M.-L. Shyu, J. Yan, and C. Zhang, "A progressive morphological filter for removing nonground measurements from airborne LIDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 4, pp. 872–882, Apr. 2003.

[36] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 2155–2162.

[37] S. Gao, Z. Niu, G. Sun, D. Zhao, K. Jia, and Y. Qin, "Height extraction of maize using airborne full-waveform LIDAR data and a deconvolution algorithm," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1978–1982, Sep. 2015.

[38] O. J. Afolabi, G. P. Mabuza-Hocquet, F. V. Nelwamondo, and B. S. Paul, "The use of U-Net lite and extreme gradient boost (XGB) for glaucoma detection," *IEEE Access*, vol. 9, pp. 47411–47424, 2021.

[39] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, Sep. 2021.