Automatic Building Extraction via Adaptive Iterative Segmentation With LiDAR Data and High Spatial Resolution Imagery Fusion

Shanxiong Chen, Wenzhong Shi, Mingting Zhou, Min Zhang, and Pengfei Chen

***Abstract*—Extracting buildings from remotely sensed data is a fundamental task in many geospatial applications. However, this task is resistant to automation due to variability in building shapes and the environmental complexity surrounding buildings. To solve this problem, this article introduces a novel automatic building extraction method that integrates LiDAR data and high spatial resolution imagery using adaptive iterative segmentation and hierarchical overlay analysis based on data fusion. An adaptive iterative segmentation method overcomes over- and undersegmentation based on the globalized probability of boundary contour detection algorithm. A data-fusion-based hierarchical overlay analysis extracts building candidate regions based on segmentation results. A morphological operation optimizes a building candidate region to obtain final building results. Experiments were conducted on the international society for photogrammetry and remote sensing (ISPRS) Vaihingen benchmark dataset. The extracted building footprints were compared with those extracted using the state-of-the-art methods. Evaluation results show that the proposed method achieved the highest area-based quality compared to results from the other tested methods on the ISPRS website. A detailed comparison with four state-of-the-art methods shows that the proposed method requiring no samples achieves competitive extraction results. Furthermore, the proposed method achieved a completeness of 94.1%, a correctness of 90.3%, and a quality of 85.5% over the whole Vaihingen dataset, indicating that the method is robust, with great potential in practical applications.**

***Index Terms*—Adaptive segmentation, building extraction, data fusion, high spatial resolution imagery (HSRI), LiDAR.**

Manuscript received January 12, 2020; revised April 14, 2020; accepted April 27, 2020. Date of publication May 6, 2020; date of current version May 22, 2020. This work was supported by the Ministry of Science and Technology of the People's Republic of China under Project 2017YFB0503604, The Hong Kong Polytechnic University (1-ZVN6; Smart Cities Research Institute, The Hong Kong Polytechnic University), and in part by the State Bureau of Surveying and Mapping of the People's Republic of China (1-ZVE8). *(Corresponding author: Mingting Zhou.)*

Shanxiong Chen and Min Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: shanxiongchen@whu.edu.cn; 007zhangmin@whu.edu.cn).

Wenzhong Shi is with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: john.wz.shi@polyu.edu.hk).

Mingting Zhou is with the State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: mintyzhou@whu.edu.cn).

Pengfei Chen is with the School of Geospatial Engineering and Science, Sun Yat-Sen University, Guangzhou 510275, China and also with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: pfchen@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.2992298

I. Introduction

BUILDING extraction from remote sensing data is for the starting point in many real-world applications. These include cartographic mapping, urban planning, three-dimensional (3D) city modeling, and disaster emergency response [1]–[4]. Manual interpretation of building areas from massive remote sensing data is laborious and inefficient [5]. To satisfy the increasing need for accurate building outline of the urban regions, and their continuous update demands, more automated methods, with higher accuracy geo-information, are required. Developing automatic and robust algorithms for building extraction is a research frontier in the field of remote sensing [6].

Automated building extraction methods include image-based, LiDAR-based, and data fusion-based methods based on the input data [7]. Image-based methods rely on spectral properties derived from high spatial resolution imagery (HSRI). The major obstacle stems from spectral ambiguities and shadow occlusions that lead to significant errors [8], as well as low-level automation [9]. LiDAR-based methods employ the intensity, echo, and geometric attributes of the LiDAR point cloud to extract buildings. Although LiDAR has improved the level of automation in the building detection process [9], the use of raw or interpolated data alone suffers from poor horizontal accuracy of building boundaries [10]. Given the pros and cons of LiDAR and HSRI, it has been suggested that these data be fused to improve the degree of automation and the robustness of automatic building extraction [11], [12]. Data fusion-based methods, using both HSRI and LiDAR data have attracted more attention, but questions remain. The methods optimally combining HSRI and LiDAR data so that their disadvantages are effectively compensated is an active area of current research.

Building detection techniques integrating LiDAR data and imagery can be divided into two groups [13]. One group uses LiDAR data as the primary cue for building detection, and images to only remove vegetation [9], [14]–[17]. Thus, they have poor horizontal accuracy for detected buildings. The other group is an integration method [13], [18]–[27], that uses LiDAR data and images as the primary cues to delineate building outlines, as well as images to remove vegetation. Consequently, the horizontal accuracy for the detected buildings is improved. Our proposed building detection approach falls into latter category. Recently developed deep learning based method provides a new optimal way to combine these two data sources.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Building detection approaches based on the convolutional neural networks (CNNs) have shown superior performance since CNNs can learn high-level and discriminative features automatically [28], [29]. Long *et al.* [30] extended the original CNN structure to enable dense prediction by a pixels-to-pixels fully convolutional network (FCN). FCN yields full resolution classification maps, has now become a common framework for recent building extraction methods [6], [31]–[33]. Despite the high performance of learning-based methods, these data fusion-based methods heavily rely on a considerable body of well-annotated training samples.

Existing data-fusion-based methods have other limitations. Many methods require specific threshold settings or certain empirical rules when, using low- or mid-level features to extract building footprints. Methods taking image segmentation as a prerequisite step, are highly dependent on segmentation parameter settings and are easily affected by such as solar radiation, shadows, and random noise found in HSRIs [6]. Deep learning methods using high-level semantic features require a large number of well-annotated samples and their generalize ability is limited by the training domain.

An automatic building extraction method integrating LiDAR data and HSRI could address these problems. In our proposed method, an adaptive iterative segmentation method is designed to overcome over- and undersegmentation problems based on the globalized probability of boundary (gPb) algorithm. In the adaptive iterative segmentation process, both LiDAR data and the HRSI are segmented to obtain hierarchical segmentation results, and a data-fusion-based hierarchical overlay analysis based on the segmentation results overcomes shadow occlusion. In the proposed method, during data-fusion-based hierarchical overlay analysis process, other nonbuilding backgrounds such as vegetation under shadowed areas, are hierarchically eliminated from initial candidate regions. The main contributions of this study are as follows.

1) A new data-fusion-based method is proposed for automatic building extraction in complex urban scenes.
2) An adaptive iterative segmentation method is designed to overcome over- and undersegmentation based on gPb contour detection algorithm.
3) A data-fusion-based hierarchical overlay analysis is designed to overcome shadow occlusion.

The remainder of this article is organized as follows. Section II presents a brief review of the data-fusion-based method and contour detection. Section III describes the proposed method, Section IV presents and analyses the experimental results, and the main conclusions are presented in Section V.

II. Related Work

A. Data-Fusion-Based Method

Advances in sensor technologies make the acquisition of data from different sources much easier. The combination of different data sources is becoming widely used for object segmentation and recognition in both the computer vision and remote sensing communities. Many algorithms combining red, green, blue (RGB) and depth information for object segmentation/recognition [34]–[38] have been proposed in the computer vision community. A comprehensive review is beyond the scope of this article, instead focusing on data-fusion-based building extraction methods in remote sensing.

The latest data-fusion-based methods are reviewed in this section as research on building extraction has increasingly focused on using imagery and LiDAR data. Rottensteiner *et al.* [15] proposed a supervised classification-based building extraction framework. However, traditional supervised classification-based methods are affected by inappropriate feature selection, underestimation of urban classes, and insufficient training samples [39]. Therefore, researchers have developed hierarchical approaches, which aim to exclude nonbuilding areas/pixels in a step-wise fashion. Moussa and El-Sheimy [16] proposed an object-based two-stage classification method to integrate LiDAR data and HSRI. In the first stage, the entire digital surface model (DSM) data is segmented into objects based on height variation. Then, the objects are first classified into buildings, trees, and ground according to a minimum area threshold. The second stage of classification is conducted to tune the preliminary classification of the first stage according to rules. This iterative classification scheme was further expanded to include more features based on the previous successive classification phases. Chen *et al.* [14] generated initial building segments by truncating both normalized DSM (nDSM) and normalized difference vegetation index (NDVI) sequentially. The final building masks are determined by a set of rules related to the region size and the spatial relation between trees and buildings. To incorporate both height and spectral information in the segmentation, Gerke and Xiao [17] presented a method that exploits accurate, homogeneous, and complete 3-D geometry from the point cloud, and spectral information from images to detect urban buildings, trees, natural, and sealed ground objects. These methods use LiDAR data as the primary cue extracting primitives of buildings in LiDAR data, and fusing different types of source data for classification. However, these methods were less reliable at building edges. Therefore, researchers have developed data fusion methods, which use both the LiDAR data and the imagery as the primary cues to delineate building outlines.

Many data-fusion-based methods use the two data sources as the primary cues to use low- or mid-level features to extract buildings. Sohn and Dowman [19] employed a data-driven approach on the IKONOS imagery and a model-driven approach on the LiDAR with low point density to extract rectilinear lines around buildings. Extracted lines were regularized by analyzing the dominant line angles. The results showed that this system could successfully delineate most buildings in a complex scene; but tends to overlook some buildings because only the building edges with parallel and orthogonal structures are considered. Cheng *et al.* [21] proposed a similar technique with a precise geometric position. To deal with the problems encountered when detecting building edges solely from point clouds or images, Li and Wu [22] proposed a new adaptive method for building edge detection by fusing the two data sources. Nonbuilding objects are removed by mathematical morphology and region growing techniques. Edge buffer areas are created in the image space using edge points of the individual roof patches. The pixels with a local maximal gradient in a buffer area are judged as candidate edges. The ultimate boundaries are determined by fusing the

edges in the image and the roof patch using a morphological operation. The experimental results show that the method is adaptive for various building shapes. Zhang *et al.*[24] used many cues to remove irrelevant candidates, such as height, to remove low height objects (e.g., bushes), and width to exclude trees with small horizontal coverages. Image entropy and color information were jointly applied to remove easily distinguishable trees. A rule-based procedure using the edge orientation histogram from the imagery eliminates false-positive candidates. However, this proposed algorithm is moderately slow as compared to the original detector in [13]. In some unusual cases (e.g., buildings with green roofs, vegetation with shadows, and self-occlusions.), the improved algorithm will fail altogether.

To establish a relationship between the low-level image primitives (e.g., line segments) and the higher level geospatial objects (e.g., intersections and closed boundaries), the hypothesis/verification paradigm is adopted in some approaches. Based on the assumption that building roofs are planar, Lee *et al.* [23] proposed a new approach to extract the boundaries of complex buildings from LiDAR and photogrammetric imagery. To do this, they used several methods to group low-level features, such as height distributions, segments, and edges, into higher-level features by using directional histograms, entropy, region segmentation, and merging, line segments matching, and perceptual grouping. Gilani *et al.* [25] proposed a fully data-driven building extraction and regularization method using detected candidate building regions and line segments in an image. The buildings are extracted, including partly occluded and shadowed after the vegetation removal, employing multisource data and grid index structure. Building footprints are generated using the image lines and the extracted building boundaries. Although these methods provide accurate extraction results, they often require specific threshold settings or certain empirical rules when using low-level or mid-level features.

Another stream of data-fusion-based methods uses the two data sources as the primary cues to use data classification or segmentation to extract buildings. Qin and Fang [18] obtained an initial building mask hierarchically by considering the shadow and off-terrain objects. A graph cut optimization algorithm based on spectral and height similarities refines the mask by exploiting the connectivity between the building and nonbuilding pixels. This method can handle shadows and small buildings to an extent. However, the building patches on steep slopes, roof parts under shadows, and the roofs with vegetation cannot be extracted correctly. Moreover, more automated parameter tuning for truncating thresholds is needed for more challenging datasets. Zarea and Mohammadzadeh [26] utilized support vector machines (SVMs) to separate buildings from trees based on features found in both LiDAR data and aerial images. In the SVM, an automatic procedure was used for selecting the training data. Awad [27] proposed an innovative fusion method for segmentation, which reduces oversegmentation through increasing the success rate of feature extraction. Based on the improved fusion method, the confusion between different urban classes and over-segmentation is reduced. A disadvantage of these methods is that they are dependent on segmentation parameter settings, but deep learning offers a potential solution using artificial neural networks.

It is now possible to learn image features automatically instead of extracting them by classical methods given the tremendous jump in development in the field of artificial neural networks. Maltezos *et al.* [40] introduced an efficient deep learning framework based on CNNs that extract buildings from orthoimages and dense image matching point clouds. Experimental results indicate that a combination of raw image data with height information provides potentials in robust and efficient building detection. They further employ a CNN classifier for building extraction from the LiDAR data [41]. The proposed deep learning classifier outperforms the compared linear and nonlinear classification methods. Bittner *et al.* [31] presented a novel method to segment buildings in complex urban areas using multiple types of remote sensing data based on FCNs. Their end-to-end Fused-FCN4s framework integrates relevant contextual features from spectral and height information within a single architecture for pixelwise classification, producing a unique binary building mask. Huang *et al.* [6] developed an end-to-end trainable gated residual refinement network (GRRNet) that fuses high-resolution aerial images and LiDAR point clouds for building extraction. A modified residual learning network is applied as the encoder in GRRNet to learn multilevel features from the fused data. A gated feature labeling unit reduces unnecessary feature transmissions and refines classification results. The proposed model-GRRNet was tested on a publicly available dataset with urban and suburban scenes, illustrating that GRRNet delivers competitive building extraction performance in comparison with other approaches. Despite the high performance of these learning-based methods, their performances rely heavily on a large amount of well-annotated training samples. The proposed method based on contour detection addresses these problems. To understand the method further, below is a brief introduction to contour detection.

B. Contour Detection

Contour detection refers to finding closed boundaries between objects or segments. Arbelaez *et al.* [42] proposed a gPb contour detection algorithm. gPb contour detection provides accurate contour results compared to other approaches on image segmentation (e.g., mean shift, multiscale normalized cuts and region merging) and edge detection (e.g., Prewitt, Sobel, Roberts operator, and Canny detector) [42], often referred to as a state-of-the-art method for contour detection. The algorithm is fast, with no parameters to tune. This is achieved by combining edge detection and hierarchical image segmentation while integrating texture, color, and brightness image information at both the local and a global scale. The algorithm creates an ultrametric contour map (UCM) [43], with values reflecting the contrast between neighboring regions.

The processing pipeline of gPb contour detection is as follows. In the first step, oriented gradient operators for brightness, color, and texture are calculated on two halves of differently scaled discs to obtain local image information. The cues are merged based on a logistic regression classifier resulting in a posterior probability of a boundary, i.e., edge strength per pixel. The global image information is obtained through spectral clustering detecting the most salient edges only. This is done by

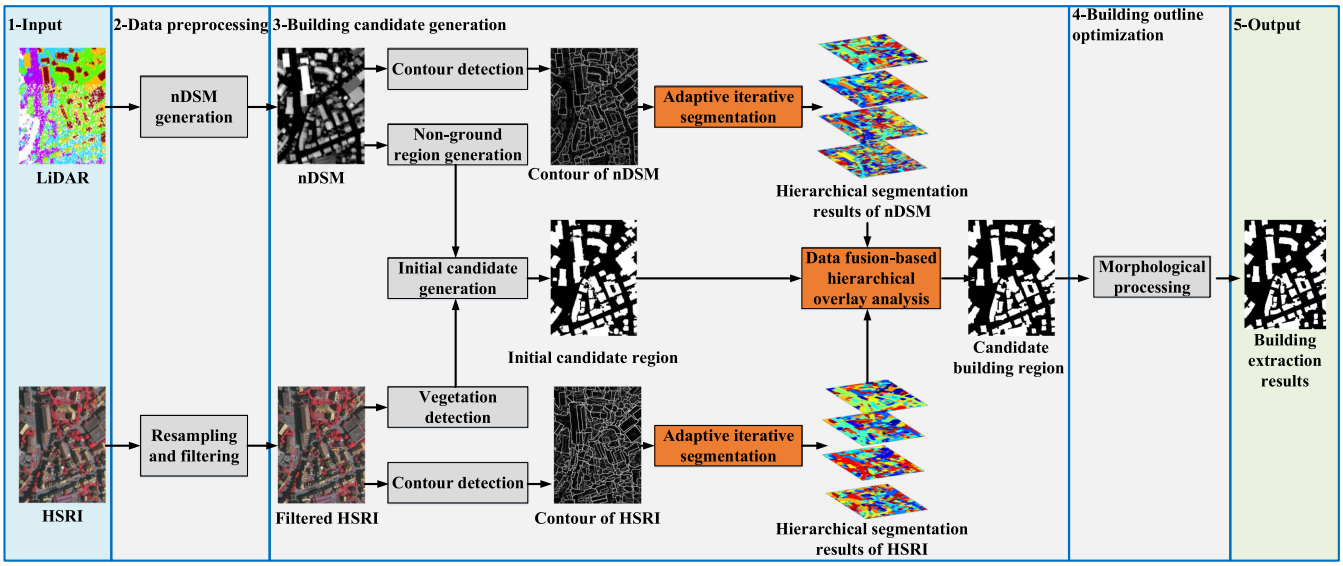


Fig. 1. Workflow of the proposed method.

examining a radius of pixels around a target pixel in terms of oriented gradient operators as for the local image information. The local and global information are combined through learning techniques and trained on natural images from the "Berkeley Segmentation Dataset and Benchmark" [44]. By considering image information on different scales, relevant boundaries are verified, while irrelevant ones, e.g., in textured regions, are eliminated. In the second step, initial regions are formed from the oriented contour signal provided by a contour detector through oriented watershed transformation. Subsequently, hierarchical segmentation is performed by weighting each boundary and their agglomerative clusters to create a UCM that defines the hierarchical segmentation. Thresholding the resulting UCM with some global threshold k provides by definition a set of closed curves, the boundaries of the segmentation at scale k. The lower the levels of k, the fewer contours are transferred from the contour map to the binary boundary map.

An adaptive iterative segmentation with LiDAR data and HSRI fusion method based on gPb contour detection algorithm is proposed. Crommelinck *et al.* [45] applied gPb contour detection to aerial imagery for automated cadastral mapping. However, contours solely from the images are affected by shadows and occlusion. The LiDAR point clouds provide initially closed edges around the roof patches, no matter the shadow and occlusion. Therefore, the two contours from the two sources can be fused to complement each other. To our knowledge, it is the first time gPb contour detection algorithms have been used for nDSM and combined with HSRI contour detection. On the other hand, previous methods have used a specific threshold to produce segmentation results from UCM, thus creating over- or under-segmentation problems. By proposing an adaptive iterative threshold method, hierarchical segmentation results were generated, thus, conquer the over- and undersegmentation problem. Furthermore, the UCM has no semantic meaning except for contrast between regions, by fusing the height information provided by LiDAR data, the contour is given better semantic information. Therefore, by overlay analysis of the initial candidate area and hierarchical segmentation results, the building will be detected automatically.

III. Proposed Method

The workflow of the proposed method consists of three stages as in Fig. 1. In the data preprocessing stage, an nDSM and filtered HSRI are generated for use at later stages. Data preprocessing arranges the LiDAR point cloud and HSRIs into a grid format at the same resolution and removes noise from both datasets. In the stage of building candidate generation, candidate building regions are extracted through adaptive iterative segmentation and data-fusion-based hierarchical overlay analysis. Subsequently, candidate building regions are optimized with morphological processing, to produce final output buildings.

Data preprocessing, building candidate generation, and building outline optimization stages depicted in Fig. 1 are discussed separately in Section III-A through Section III-C. The Vaihingen Area1 dataset was selected as a use case to illustrate the processes applied in the proposed approach.

A. Data Preprocessing

The proposed method takes LiDAR, ortho imagery as an input. The preprocessing module removes noise from the LiDAR and HSRI datasets, generate an nDSM with LiDAR, and unifies the resolution of the two data sources. The workflow of data preprocessing module is shown in Fig. 2.

As shown in Fig. 2(a), outlier removal and ground points filtering generate an nDSM from LiDAR data. LiDAR point clouds include noise points, which have anomalously high or low elevations in comparison to the elevations expected for ground, vegetation, and structures in the survey area. LasTools (downloaded via http://www.cs.unc.edu/ isenburg/lastools/) is a commonly used point cloud preprocessing tool. In this research, we used two LasTools subtools *lasnoise* and *lasground* to filter the noise points as well as separate ground points from

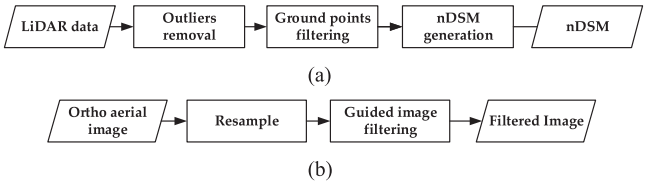


Fig. 2. Workflow of data preprocess module. (a) LiDAR preprocess. (b) HSRI preprocess.

nonground points. The *lasnoise* tool removes all points that have only four or fewer other points in their surrounding 3*3*3 grid (with the respective point in the center cell) where each cell is 1*1*1 meters in size. For *lasground* tool, the terrain type is set to metropolis and the other parameters are default. It classifies all LAS files with the default settings but uses even wider spacing to allow for very large buildings. The separated ground points and the whole LiDAR point cloud are interpolated into DSM and digital terrain model (DTM) in raster format to make them consistent with aerial images. The DSM and DTM cell size is set as close as possible to the reciprocal of the average point density. The nDSM is obtained by subtracting the DTM from DSM.

As shown in Fig. 2(b), a filtered image is generated using ortho aerial image by resampling and guided image filtering [46]. The ortho image is resampled into the nDSM spatial resolution and smoothed through guided filter. Guided filter is an edge-preserving smoothing technique, which is widely utilized in computer vision and remote sensing community [32]. The basic idea of guided filtering is to establish a local linear model between the guided image and the filtered result. With the input image as the guidance image, the filtering result is more structured and less smoothed. For more details, the reader can refer to [46]. The ortho image is processed through guided filtering with itself as guiders to reduce the negative effects of noises while preserving edge smoothness. A guided filter with default settings was applied in a MATLAB implementation. The preprocessed data was fed to the building candidate generation module.

B. Building Candidate Generation

Candidate building regions are generated by removing vegetation and other nonbuilding backgrounds. Building candidate generation includes two steps, initial candidate area generation and removal of other nonbuilding objects. An initial candidate area is generated by removing vegetation over nonground regions. The other nonbuilding backgrounds are eliminated by adaptive iterative segmentation and data-fusion-based hierarchical overlay analysis. Adaptive iterative segmentation was adopted based on contour detection to generate hierarchical segmentation results, and data-fusion-based hierarchical overlay analysis in further filtering of nonbuilding backgrounds in the nonground area to obtain more accurate candidate building regions.

1) Initial Candidate Area Generation: The main interference in the nonground regions obtained during data preprocessing is vegetation, which needs to be removed in order to obtain initial candidate area. Vegetation cannot be removed by using elevation information alone, so spectral information is introduced

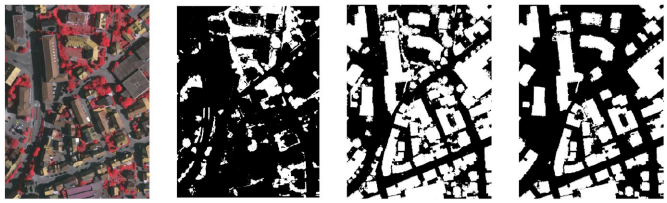


Fig. 3. Initial candidate generation on the Vaihingen Area1 image: (a) image; (b) binary vegetation mask; (c) nonground binary mask; (d) initial candidate area. Note that the generated areas are shown in white color.

to improve vegetation removal results. Because of its particular biological structure, vegetation appears relatively dark in the red band and relatively bright in the near-infrared (NIR) and green bands [47]. Based on this property, many vegetation indices have been developed. They can be divided into two categories: NIR band-based and green band-based [48]. For NIR band-based vegetation indices, the NDVI [47] is currently the most popular index. There are also some green band-based vegetation indices, among them, the normalized difference green band-based index (NDGI) [49] combines the green and the red bands in a normalized way to offset the influence of the light, to some extent. The green band-based indices are inferior to the NIR band-based indices. Nevertheless, for many aerial images and close-digital images without the NIR band, the green band-based indices are the only option.

The two indices are defined as follows:

$$\text{NDVI} = \frac{\text{NIR-Red}}{\text{NIR} + \text{Red}} \tag{1}$$

$$\text{NDGI} = \frac{\text{Green} - \text{Red}}{\text{Green} + \text{Red}} \tag{2}$$

where Red, Green, and NIR stand for the spectral reflectance measurements acquired in the red (visible), green, and near-infrared regions, respectively.

A problem occurs in shadow areas, because vegetation has a low value in the two indices. To solve this issue, researches [26], [48], [50] improved the accuracy of vegetation detection in the shadow areas by detecting the shadow. However, these methods require shadow extraction and, therefore, affected by shadow detection algorithms. Moreover, some methods also require manually set thresholds, which reduce the degree of automation.

The NDVI is binarized by Otsu's method [51] to detect vegetation areas automatically. The vegetation detection result on the Vaihingen example image is shown in Fig. 3(b). It can be seen that the NDVI binarized by Otsu's method suffered from underdetection. Nearly, all the detected vegetation was located in the sunlit areas, while the vegetation in the shadowed areas was missed for the most part. But that is enough for our proposed method to detect buildings.

Non-ground binary masks (NGBM) are identified from an nDSM using Tsai's moment preserving automatic threshold method [52], as shown in Fig. 3(c). Moment preserving thresholding is a parametric method that segments an image based on the condition that a thresholded image has the same moments

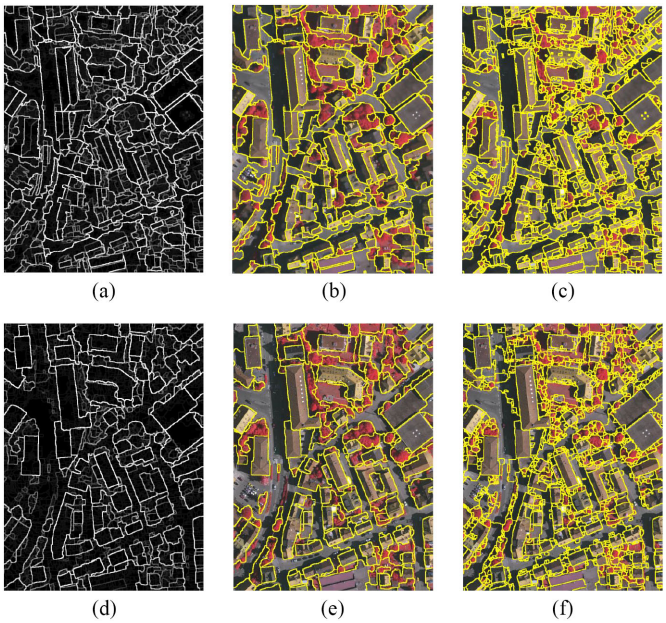


Fig. 4. Hierarchical segmentation results on the Vaihingen Area1 image and nDSM: (a) UCM of aerial image; (b) contour of *iseg4*; (c) contour of *iseg1*; (d) UCM of nDSM; (e) contour of *nseg4*; and (f) contour of *nseg1*. Note that the contour is shown in yellow color. All the image is dilated for better vision.

as the original image. It can be seen that the NGBM preserves the nonground object and effectively eliminated ground objects. Fig. 3 shows the process for initial candidate area generation. Given a preprocessed image, automatic vegetation detection produces a binary vegetation mask (BVM) to identify vegetation pixels. Nonground region generation produces a NGBM to identify nonground pixels. The intersection of NGBM with BVM, which identified vegetation over nonground areas, is eliminated from NGBM to obtain the initial candidate area, as shown in Fig. 3(d). Consequently, the nonground plant is partially eliminated.

2) Adaptive Iterative Segmentation Based on Contour Detection: After the initial candidate area is obtained, nonbuilding backgrounds areas still need to be removed. We obtain candidate building regions by object-oriented analysis. Object-oriented analysis takes image segmentation as a prerequisite step. The segmentation results are highly dependent on the segmentation parameter settings. To solve this problem, an adaptive iterative segmentation algorithm is proposed based on contour detection. UCM of image and nDSM are generated using the gPb contour detection algorithm. The UCM is a weighted contour image that, by construction, has the property of producing a set of closed curves for any threshold. The value range of UCM is 0 to 1, and large values indicate high contrast. The lower the threshold value, the smaller the segmentation object. Experiments showed that the strong contrast between edges with UCM values above 0.4 is consistent in the effect of distinguishing buildings. Thus, the UCM is iteratively binarized with decreasing thresholds from 0.4 to 0.1, at 0.1 intervals. The hierarchical segmentation results are provided based on the connected component analysis [53] of the binary UCM map. The same procedure is adopted for both UCM of the image and UCM of nDSM. Fig. 4 shows the contour of segmentation results on the Vaihingen Area1 image for UCM of the image and UCM of nDSM at different thresholds. In the following, we refer to the eight segmentation results as *iseg4*, *iseg3*, *iseg2*, *iseg1*, *nseg4*, *nseg3*, *nseg2*, and *nseg1*. The *iseg4* image is the segmentation result from the UCM of the image at threshold 0.4, *nseg4* indicates the segmentation result from the UCM of the nDSM at threshold 0.4, and so on.

Fig. 4(a)–(c) shows processing results of the Vaihingen Area1 aerial image. Fig. 4(a) is the UCM generated by gPb contour detection. Fig. 4(b) shows a undersegmentation result. Fig. 4(c) shows an oversegmentation result. The edge of segmentation results is precise. In Fig. 4(b), a segmentation object contain several targets simultaneously. In Fig. 4(c), a target consist of several segmentation objects. At the same time, there is a hierarchical relationship between Fig. 4(c) and (b). Because contours in Fig. 4(b) all exist in Fig. 4(c). The generation of Fig. 4(c) is based on Fig. 4(b), so they have a hierarchical relationship. Contour detection solely from the aerial image is affected by shadow and occlusion. Targets under shadow and occlusion are not segmented.

Fig. 4(d)–(f) shows processing results of the Vaihingen Area1 nDSM image. Fig. 4(d) is the UCM generated by gPb contour detection. Fig. 4(e) shows a undersegmentation result. Fig. 4(f) shows an oversegmentation result. The edge of segmentation results is not precise as in aerial image. The hierarchical relationship is exist between Fig. 4(e) and (f) too. Furthermore, contour detection solely from the nDSM is not affected by shadow and occlusion. Targets under shadow and occlusion are segmented ideally.

3) Data-Fusion-Based Hierarchical Overlay Analysis: Other nonbuilding backgrounds are eliminated by hierarchical overlay analysis of the eight segmentation results and initial candidate area, discussed in the following section. Fig. 5 shows hierarchical overlay analysis for a simulated data. As shown in Fig. 5, the hierarchical overlay analysis module obtains candidate building regions by overlay analysis of segmentation results and initial candidate area. The hierarchical overlay analysis process zooms in at the bottom of Fig. 5. The initial candidate region is overlaid on segmented image Seg4 (undersegmentation). If the overlapped area exceeds area proportional threshold (APT), it is considered as a building object. The area proportion is the ratio of the area of the initial candidate region to the area of the segmentation object. APT is a value that determines whether the degree of overlap between the initial candidate region and each segmented object meets the conditions for building candidate. This is discussed in more detail in Section IV-B. This object in initial candidate regions would be erased at the latter stage. The remaining initial candidate regions is overlaid on the segmented image Seg3, and the operation is repeated until the segmented image Seg1 (oversegmentation) is processed. All the reserved objects are merged to obtain the candidate building region.

The initial candidate area belonging to the building will account for a larger proportion of the segmentation object, while the remaining nonbuilding backgrounds are the opposite. In HSRIs, buildings usually have low spectral variation corresponding to the building body and a high spectral variation

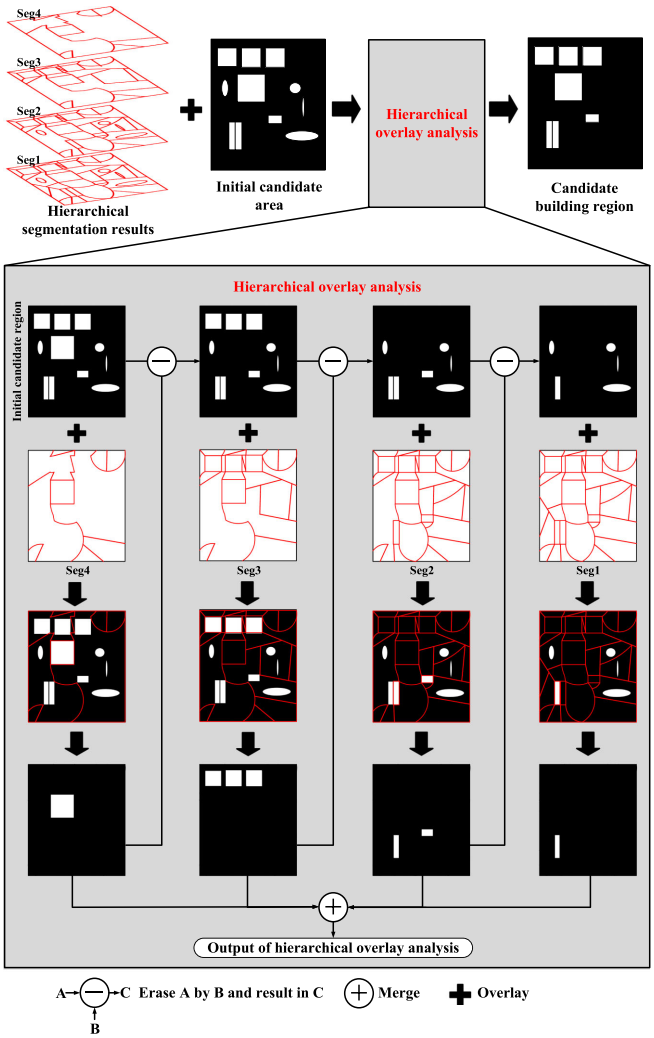


Fig. 5. Workflow of hierarchical overlay analysis module for a simulated data.

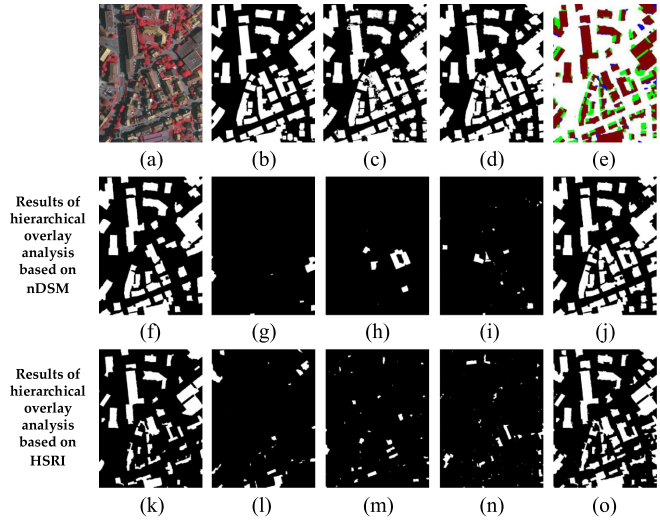


Fig. 6. Results of data-fusion based hierarchical overlay analysis at Vaihingen Area1: (a) aerial image; (b) ground truth; (c) initial candidate area; (d) building candidate region; (e) color version of building candidate region; (f) remain object at *nseg4* overlay analysis; (g) remain object at *nseg3* overlay analysis; (h) remain object at *nseg2* overlay analysis; (i) remain object at *nseg1* overlay analysis; (j) merge result of nDSM's segmentation result; (k) remain object at *iseg4* overlay analysis; (l) remain object at *iseg3* overlay analysis; (m) remain object at *iseg2* overlay analysis; (n) remain object at *iseg1* overlay analysis; (o) merge result of image's segmentation result.

corresponding to the building periphery, so they often have strong edges and high contrast with the surrounding environment [54]. In LiDAR data, there are height differences between ground points and nonground points, especially for buildings, where the height difference is generally vast. Due to these characteristics, buildings tend to have higher UCM values in both an nDSM and HSRI. In addition, buildings are above-ground objects. Thus, the segmented objects generated by UCM belong to buildings tend to have a larger proportion of initial candidate regions in it. The nonbuilding backgrounds consist of two parts. One is the vegetation left after removing the bright vegetation. It is generally shaded vegetation or sparse vegetation. Such vegetation is usually lower than buildings in height, smaller in area than buildings, and the internal elevation of vegetation is irregular. Thus, their flatness and contrast of the spectrum and elevation are lower than those of buildings. The other nonground backgrounds are usually smaller than buildings and have no similar characteristics to buildings. In conclusion, the nonground backgrounds tend to have smaller UCM values in both an nDSM and HSRI, and occupy a lower proportion in the overlay analysis of segmentation results and initial candidate area.

Binarizing a UCM at a larger value results in a larger object; therefore, the path from *nseg4* to *nseg1* is a hierarchical process from undersegmentation to oversegmentation. If the initial candidate area is considered a building in an object of *nseg4* (undersegmentation), it must also be a building in an object of *nseg1* (oversegmentation). In order to reduce redundant operations, the overlay analysis is performed in the order from undersegmentation results to oversegmentation results. Each segmentation result is marked as many connected objects. For each object, counting the ratio of the number of the initial candidate region pixels to the number of the object pixels and setting an APT, the candidate building region is generated. Specifically, *iseg4* is first used to superimpose the initial candidate area, and the object whose area ratio exceeds APT is retained, and the segment is removed from the initial candidate region, then the *iseg3*, *iseg2*, and *iseg1* are repeated the abovementioned procedure. The four results are combined, and the result from the UCM of the nDSM and UCM of the image are merged. The whole process, for the test data, is shown in Fig. 6.

As shown in Fig. 6, through the data-fusion hierarchical overlay analysis in this section, high-quality building candidates are obtained. Fig. 6(c) is the initial candidate area result after removing vegetation from nonground areas. The outline of Fig. 6(c) is not very accurate and contains many nonbuilding backgrounds. Fig. 6(f) is the result of overlay analysis on the *nseg4*, the segmentation result from UCM of nDSM at threshold 0.4, and the initial candidate area. The optimized result is effectively filtering the fragmented nonbuilding backgrounds. However, due to the undersegmentation of *nseg4*, some buildings were also filtered out. Fig. 6(g)–(i) are the results of using *nseg3*, *nseg2*, and *nseg1* to supplement buildings missed by *nseg4* undersegmentation.

Fig. 6(j) is the union of Fig. 6(f)–(i). Compared to Fig. 6(c), the building extraction results in Fig. 6(j) remove most of the nonbuilding background. However, because of the nDSM-based method, the edges are inaccurate, and the vegetation adjacent to the building cannot be removed. Therefore, UCM of HRSI is utilized to hierarchical overlay analysis to take full advantage of nDSM and HRSI to improve the initial building outline further. The third row in Fig. 6 is the result of hierarchical overlay analysis based on HRSI. Fig. 6(k) is the result from the overlay analysis of on the *iseg4*, the segmentation result from the UCM of HRSI at threshold 0.4. The optimized result is effectively extracted the precise boundary of the building. However, due to the undersegmentation of *iseg4*, some buildings were also filtered out. Fig. 6(l)–(n) are the results of using *iseg3*, *iseg2*, and *iseg1* to supplement buildings missed by undersegmentation. Fig. 6(o) is the union of Fig. 6(k)–(n). Compared to Fig. 6(c), the building extraction results in Fig. 6(o) remove most of the nonbuilding backgrounds, and get more accurate edges, but omit the buildings obscured by shadows. The shaded buildings were retained in the nDSM results. Finally, by fusing the two results, we get Fig. 6(d). It can be seen that Fig. 6(d) removes most of the nonbuilding background in Fig. 6(c). Fig. 6(e) is candidate building region shown in color for higher clarity. The results of nDSM only, the results of HSRI only, and the results of both nDSM and HSRI are colored in green, blue, and red. In contrast to Fig. 6(o), Fig. 6(e) retains buildings obscured by shadows. Object boundaries in Fig. 6(e) are mostly green except for the shaded areas, indicating Fig. 6(e) mainly takes the boundary obtained from the HSRI as the final boundary. Fig. 6(e) has more accurate boundaries than in Fig. 6(j), and retains some missed detections. However, Fig. 6(e) is a simple merge of results from the UCM of the nDSM and UCM of the image. The false detections from nDSM-based procedure were also passed to the candidate building region, for further processing.

C. Building Outline Optimization

The building outline must be refined and small noises removed. The boundary of the candidate building region generated as the previous section is still not very accurate. Since we are merging the two segmentation results, the boundary of the results may have some glitches as well as some false detection of objects above the ground (such as vegetation) adjacent to the building. At the same time, some small areas also are retained in the candidate area. The candidate building region is filtered with the initial candidate region to eliminated confused vegetation region adjacent to the building in complex scene. To suppress small noises, a threshold of 10 m^2 for the shape attribute area is applied to the extraction results. Morphological filtering refines the building boundary.

IV. Experiment and Analysis

Experiments and discussion are presented in this section, which is divided into three following sections. Section IV-A describes the datasets and evaluation metrics. Section IV-B presents the analysis of parameter settings. The experimental results and qualitative and quantitative analysis of the tested methods are presented in Section IV-C.

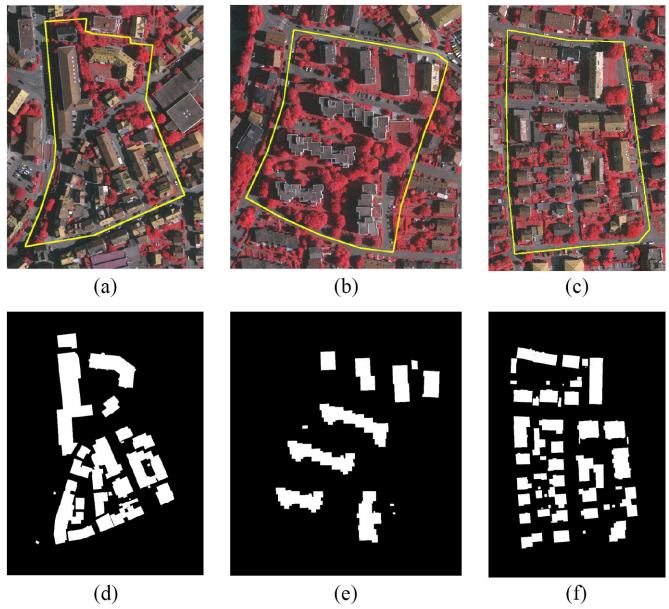


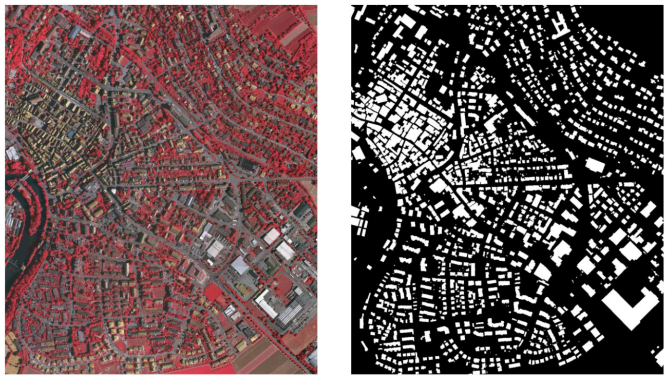
Fig. 7. (a)–(c) Aerial images of areas 1–3. Region of interest is specified with yellow lines in aerial images. (d)–(f) Reference data for buildings of areas 1–3.

A. Dataset and Error Metric

To verify the effectiveness of our method, extensive experiments were conducted on the Vaihingen test dataset, which was captured over Vaihingen in Germany [55] with an aerial camera. We first evaluate the proposed algorithm with the three test sites by comparing the extracted buildings with the reference data, and then compare the whole dataset with the manually sketched building masks on the aerial image.

1) Dataset: To assess the performance of our approach, we conducted experimental evaluations on the international society for photogrammetry and remote sensing (ISPRS) Vaihingen 2-D semantic labeling challenge. This is an open benchmark dataset provided at http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html. The data includes aerial images and LiDAR data. In the Vaihingen area, ISPRS-WGIII/4 has determined three study areas: area 1, area 2, and area 3. The aerial images have three bands of infrared (IR), red (R), and green (G). Each area has a point density of 3.5, 3.9, and 3.5 points/m^2, respectively. Area 1 is characterized by dense development consisting of historic buildings with rather complex shapes along with roads and trees. Area 2 is characterized by a few high-rising residential buildings surrounded by trees. Area 3 is a purely residential area with detached houses and many surrounding trees. In these test areas, reference data were generated by manual stereo plotting. The reference for building detection consists of roof outline polygons at a planimetric accuracy of about 10 cm. The aerial image of three study areas and reference data is shown in Fig. 7.

To validate the effectiveness of our proposed method further, we tested the proposed algorithm on the entire dataset. The same



(a) (b)

Fig. 8. (a) Aerial images of the whole Vaihingen dataset. (b) Reference data for buildings of the whole Vaihingen dataset.

range of the aerial images and the LiDAR data was clipped. The ground truth of the whole area was obtained by manual editing based on the ground truth published by ISPRS. The image and ground truth of the whole Vaihingen dataset are shown in Fig. 8.

2) Error Metric: To evaluate the building extraction results quantitatively, the index system adopted by ISPRS [56] was applied. The following indices were used to measure the quality of the results.

$\mathrm{Comp}_{ar}, \mathrm{Corr}_{ar}, Q_{ar}$: Completeness, correctness and quality determined on a per-area level. These indices are related to the area that was correctly classified.

$\mathrm{Comp}_{\mathrm{obj}}, \mathrm{Corr}_{\mathrm{obj}}, Q_{\mathrm{obj}}$: Completeness, correctness, and quality determined on a per-object level. These indices count the number of objects that are correctly detected. A minimum overlap of 50% for an extracted object with the reference is required for the object to be counted as a true positive.

$\mathrm{Comp}_{50}, \mathrm{Corr}_{50}, Q_{50}$: Completeness, correctness, and quality determined on a per-object level, but only considering objects larger than 50 m^2. These indices are useful to analyze the dependency of per-object quality metrics on the object size. The threshold was chosen to select the most representative buildings per plot and the largest trees in the scene [57].

The correctness, completeness, and quality equations are calculated by

$$\text{Correctness} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$\text{Completeness} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

$$\text{Quality} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

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

B. Parameter Settings

The proposed method mainly involves the following parameters: the APT and the minimum area threshold. Objects with an area smaller than the minimum area threshold are removed in the building detection procedure. A detailed description of the selection of these parameters is as follows.

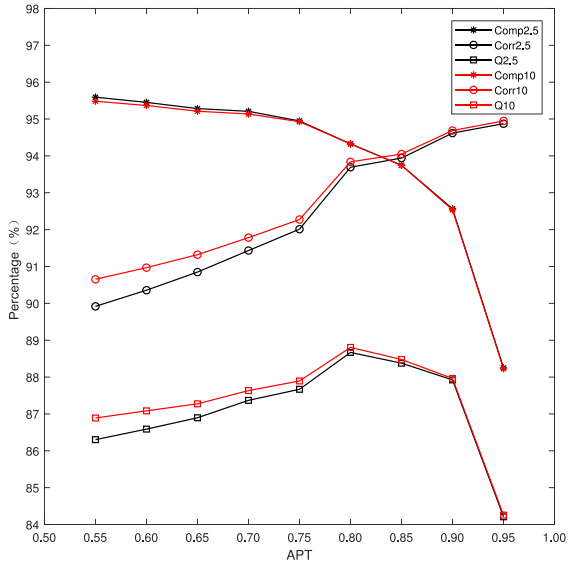


Fig. 9. APT and minimum area threshold parameter analysis.

According to [58] and [25], we compare the difference of 2.5 m^2 and 10 m^2 with the minimum area threshold parameter. The APT value ranges from 0.55 to 0.95, at an interval of 0.05. Fig. 9 shows the results, based on completeness, correctness, and quality determined on a per-area level. Fig. 9 has six polylines. The red polylines show the changes in completeness, correctness and quality with the APT when the minimum area threshold was 10 m^2. The black polylines show the changes in completeness, correctness, and quality with the APT when the minimum area threshold was 2.5 m^2. The symbol $*$, $\circ$, and $\square$ in the figure indicate completeness, correctness, and quality, respectively.

As shown in Fig. 9, comparing lines with the same symbol and different colors at every specific APT value (e.g., 0.70), the three evaluation indices show little change when the area threshold was set to 2.5 m^2 or 10 m^2. Especially when the APT increased, the difference between the two parameters was almost negligible. The quality of the extraction results is relatively high when the minimum area threshold was set to 10 m^2. This indicates that our method is prone to miss buildings under 10 m^2 in area.

When analyzing the three red polylines in which the minimum area threshold was set to 10 m^2, the smaller the APT, the higher the completeness, and the correctness increased as the APT increased, as expected. As APT increases, the value of quality first increased, then decreased, and achieved a maximum value at 0.8. The red polyline with a square shows that when the APT value was 0.8, the decrease in completeness and the increase in correctness achieved an optimal balance. Therefore, the optimal parameters for the experiment were set as follows: the APT value was set to 0.8 and the minimum area threshold was set to 10 m^2. For the other parameters, the same values for all test data were used. These parameters included the window size and structuring element of the morphological closing as well as opening operations. They were set to a 3 pixel radius disk shape.

TABLE I
ISPRS-WGIII/4 AVERAGE EVALUATION RESULTS (%) FOR DETECTED BUILDINGS OF THE PROPOSED METHOD COMPARED TO OTHER METHODS

Abbrev.	$Comp_{ar}$	$Corr_{ar}$	Q_{ar}	$Comp_{obj}$	$Corr_{obj}$	Q_{obj}	$Comp_{50}$	$Corr_{50}$	Q_{50}
CAL1	89.8	95.1	85.8	76.2	**100.0**	76.2	96.5	**100.0**	96.5
LJU1	93.6	94.2	88.5	84.2	93.9	79.8	**100.0**	**100.0**	**100.0**
TUM	83.4	94.8	79.8	74.4	99.0	73.8	93.9	**100.0**	93.9
WHUZ	83.4	91.5	77.4	70.2	61.9	49.0	88.9	95.7	85.5
HAND	88.6	91.0	81.5	75.6	83.9	66.0	93.9	97.2	91.4
TEH	83.5	93.3	78.8	74.6	94.0	71.2	94.4	**100.0**	94.4
KNTU	88.8	95.4	85.2	83.3	96.2	80.6	**100.0**	**100.0**	**100.0**
MEL	89.2	87.5	79.1	78.2	89.2	71.4	97.4	98.0	95.5
ITCM	92.9	87.5	82.0	84.8	61.4	55.3	99.1	**100.0**	99.1
ITCR	90.3	93.2	84.7	80.0	89.9	73.4	98.2	**100.0**	98.2
CAL2	88.9	**97.6**	87.0	78.2	97.2	76.5	**100.0**	**100.0**	**100.0**
RMA	92.7	93.0	86.7	82.7	96.9	80.6	**100.0**	**100.0**	**100.0**
DLR	92.3	95.8	88.7	82.7	86.7	73.4	**100.0**	**100.0**	**100.0**
SZU	92.0	93.2	86.2	**91.1**	89.7	82.5	**100.0**	**100.0**	**100.0**
IIST	82.8	82.1	70.1	75.0	80.5	63.5	89.6	96.7	86.9
KNTU_mod	89.2	92.6	83.3	85.1	96.6	82.6	**100.0**	**100.0**	**100.0**
FED_1	89.0	86.4	78.1	83.3	**100.0**	**83.3**	99.1	**100.0**	99.1
FED_2	89.6	90.0	81.5	83.9	98.6	82.9	**100.0**	**100.0**	**100.0**
IIST2	85.6	85.8	75.0	79.1	90.7	73.2	96.4	95.7	92.4
WHU_ZZ	90.8	89.3	81.9	82.1	84.4	71.3	98.2	94.1	92.5
MON3	92.3	88.4	82.3	82.4	97.5	80.7	99.1	**100.0**	99.1
Proposed	**94.3**	93.8	**88.8**	75.0	97.0	73.2	**100.0**	**100.0**	**100.0**

The best values per column are highlighted by bold font.

C. Results and Analysis

According to this analysis, the minimum area threshold parameter was set to 10 m^2 and the APT parameter set to 0.8. Then, the building detection and accuracy evaluation were performed on the three areas.

1) ISPRS Benchmark Results: The quantitative evaluation of the Vaihingen data set (Area 1, Area 2, and Area 3) is available on the ISPRS' website.[1] Since the proposed method integrates LiDAR and aerial imagery automatically, we compared our method to methods that (1) use both LiDAR and images, (2) automatic, and (3) unsupervised. The average results of areas 1–3 are compared with the results on the ISPRS website. The methods are shown in Table I. Table I has twenty-three rows and ten columns. The first column shows the abbreviation for the 22 methods. The first row shows the nine indices. Each row is the value of the indices of the relevant method. The best values per column are highlighted by bold font.

As can be seen from Table I, our method has achieved the highest Comp_{ar} and Q_{ar} compared to the other methods. The corresponding completeness and correctness values were 94.3% and 93.8%, indicating that 94.3% of all building pixels in the reference map were correctly detected, whereas 93.8% of the detected building pixels are also building pixels in the reference map. The quality index for our method in the current study area was 88.8%. CAL2 method gets the highest Corr_{ar}, but its Comp_{ar} was much lower than our method. The Corr_{ar} of our method was slightly lower than the CAL2 method; but the Q_{ar} of our method was higher than CAL2. Moreover, the CAL2 algorithm applies many empirical thresholds in building detection that are not required by our proposed method. In object-based evaluation, if all objects in the reference data are considered, and objects less than 10 m^2 are not removed, then our method only achieves a Comp_{obj} value of 75%. There are two reasons for this. First, the total number of buildings in the test area is less, but there are many small buildings of less than 10 m^2. Second, our method is not suitable for detecting buildings of less than 10 m^2. Our method, however, is still relatively accurate, having a Corr_{obj} of 97%. The FED_1 achieved the highest performance in quality determined on a per-object level. However, the index was evaluated only for buildings larger than 10 m^2. Further comparison of our proposed method and FED method presented in Section IV-C2, but in the metrics for objects larger than 50 m^2, our method also achieved the highest performance, reaching 100% quality.

2) Comparative Analysis: To further verify our proposed method, we made a detailed comparison with four state-of-the-art approaches. They were the FED_2 algorithm [25], Maltezos' method [40], U-Net [59], and DeepLabv3 [60]. The FED_2 is a fully data driven and automatic approach. Maltezos' method is an efficient CNN-based approach only needs a small collection of training samples. U-Net is a typical and widely used FCN architecture with elegant encoder-decoder structures. DeepLabv3 attains comparable performance with other state-of-the-art models on semantic image segmentation benchmark. Although training data are needed, comparisons with these state-of-the-art methods can further prove the effectiveness of the proposed method. The results of FED_2 and Maltezos' method were taken from the paper. U-Net and DeepLabv3 were trained from scratch. For a fair comparison, we manually selected the image tiles covering Area1, Area2, and Area3 as the test set. The U-Net and DeepLabv3 implementation details are as follows.

All the U-Net/DeepLabv3 experiments were conducted on a server with NVIDIA GRID M60-8Q virtual GPU accelerator, with 8 GB GPU memory. To train U-Net and DeepLabv3, the whole Vaihingen dataset was seamlessly cropped into image

[1] http://www2.isprs.org/commissions/comm3/wg4/results.html

TABLE II
AREAS 1-3 EVALUATION RESULTS (%) FOR DETECTED BUILDINGS OF THE PROPOSED METHOD COMPARED TO FOUR STATE-OF-THE-ART METHODS

Area	Method	Comp_{ar}	Corr_{ar}	Q_{ar}	Comp_{10}	Corr_{10}	Q_{10}	Comp_{50}	Corr_{50}	Q_{50}
Area1	Maltezos [40]	**96.1**	83.7	80.9	_	_	_	_	_	_
	U-Net [59]	94.8	93.0	88.5	**100.0**	94.4	**94.4**	**100.0**	**100.0**	**100.0**
	DeeplabV3 [60]	94.5	93.5	**88.7**	94.4	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**
	FED_2 [25]	85.4	86.4	75.3	83.8	**100.0**	83.8	**100.0**	**100.0**	**100.0**
	Proposed	93.1	**95.0**	**88.7**	94.4	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**
Area2	Maltezos [40]	96.0	88.0	84.9	_	_	_	_	_	_
	U-Net [59]	96.7	91.4	88.6	90.0	90.0	81.8	**100.0**	**100.0**	**100.0**
	DeeplabV3 [60]	**99.2**	88.7	88.0	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
	FED_2 [25]	88.8	84.5	76.4	85.7	**100.0**	85.7	**100.0**	**100.0**	**100.0**
	Proposed	96.4	**92.9**	**89.8**	**100.0**	90.9	90.9	**100.0**	**100.0**	**100.0**
Area3	Maltezos [40]	92.6	88.0	82.2	_	_	_	_	_	_
	U-Net [59]	93.0	93.8	87.6	87.8	**100.0**	87.8	96.9	**100.0**	96.9
	DeeplabV3 [60]	**95.6**	**94.7**	**90.7**	**97.6**	**100.0**	**97.6**	**100.0**	**100.0**	**100.0**
	FED_2 [25]	89.9	84.7	77.3	82.1	95.7	79.2	**100.0**	**100.0**	**100.0**
	Proposed	93.5	93.6	87.9	85.0	**100.0**	85.0	**100.0**	**100.0**	**100.0**
Average	Maltezos [40]	94.9	86.6	82.7	_	_	_	_	_	_
	U-Net [59]	94.8	92.7	88.2	92.6	94.8	88.0	99.0	**100.0**	99.0
	DeeplabV3 [60]	**96.4**	92.3	**89.1**	**97.3**	**100.0**	**97.3**	**100.0**	**100.0**	**100.0**
	FED_2 [25]	88.0	85.2	76.3	83.9	98.6	82.9	**100.0**	**100.0**	**100.0**
	Proposed	94.3	**93.8**	88.8	93.1	97.0	90.1	**100.0**	**100.0**	**100.0**

The best results are marked in bold and the secondary ones are underlined.

tiles with the size of 512 × 512 pixels. To increase the data volume and improve the generalization ability of the model, we augmented the data before training, including random rotation and random crop. Both nDSM and HSRI are regarded as the inputs to train the network. We trained the network until the loss converged. For U-Net, the training loss converged after 57k iterations, and the entire training process took about 22 h. For DeepLabv3, it takes about 54 h to converge, occurring after 69k iterations. The trained model was used to predict the test set to obtain the prediction result. For a fair comparison, the same postprocess as ours was applied to the prediction results. The results were mosaiced and clipped to a range entirely consistent with Area1, Area2, and Area3 for comparative analysis.

Table II shows the detailed evaluation results for the three test areas of the Vaihingen dataset, the buildings with an area size of below 10 m^2 were not included in the per-object evaluation process for the methods. Only pixel level accuracy assessment was reported in [40]. The best results are marked in bold and the secondary ones are underlined. Fig. 10 shows the per-pixel level visual evaluation of all the test areas for our proposed method, U-Net, and DeepLabv3, respectively.

Table II have 21 rows and 11 columns. The first column shows the test area. The second column is the abbreviation of the methods. From the 3rd to the 11th column shows the value of the evaluation index. As can be seen from the last row of Table II, the DeepLabv3 achieves the highest performance, and our proposed method ranks the second. Compared with DeepLabv3 in the per-area level evaluation index, our method has lower completeness and higher correctness, and the quality only differs by 0.3%. As can be seen from Fig. 10 row 3, the DeepLabv3 suffers from the omission error of small buildings and imprecise edges. Our method, however, showed stronger performance on building boundaries. In the object-level evaluation index of more than 50 m^2, all methods except U-Net achieved 100% quality. As shown in Fig. 10 row 2, U-Net has a missed detection in Area3. U-Net has false detections in each test area and has poor performance on small target detection. DeepLabv3 and U-Net require a large number of labeled samples, while Maltezos' method only needs a very small number of samples. The FED_2 and our method do not require samples. Therefore, these three methods were compared in detail. The average Comp_{ar} index of Maltezos' method [40] is 0.6% higher than our method. This is due to their completeness in Area1 outperform ours with 3%. However, higher completeness had result of lower correctness in their approach. Our method is more robust and obtains higher quality. The bold font in Table II shows that FED_2 only in the Corr_{10} index of Area2 and average of three regions delivered higher performance outcomes than our method. As shown in Fig. 10 (b), our method has a false positive (red object) in Area2, which is a plant at the edge of the shadow. This problem can be attributed to the simultaneous occurrence of three factors. First, the false positive has similar characteristics as buildings. Second, it was occluded by shadow. Third, it is at the edge of the shadow. Therefore, it is not easy to separate this plant from buildings as

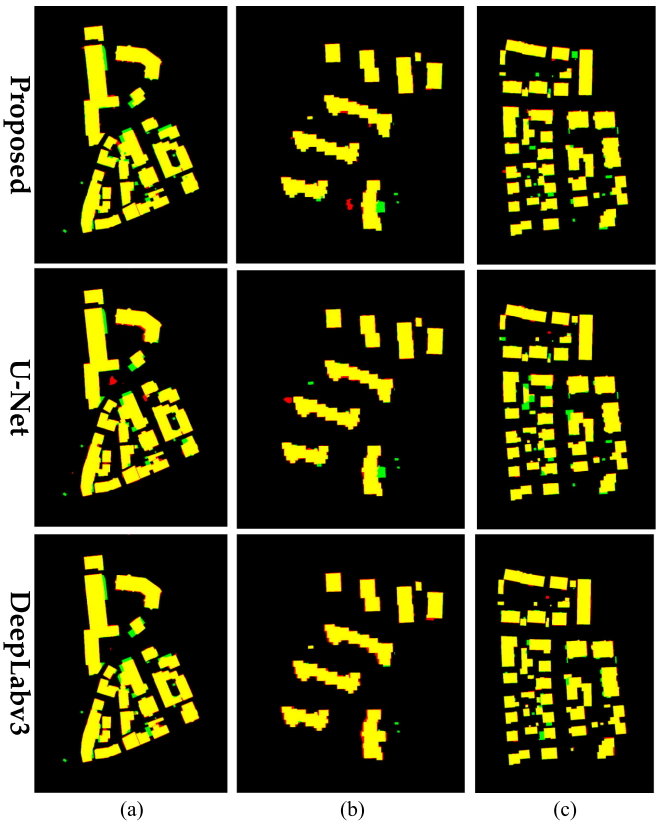


Fig. 10. Building detection result pixel-based evaluation on the Vaihingen data set: (a) Area 1; (b) Area 2; (c) Area 3. Row 1: the proposed method. Row 2: U-Net. Row 3: DeepLabv3. The TP, FP, and FN are colored in yellow, red, and blue, respectively.

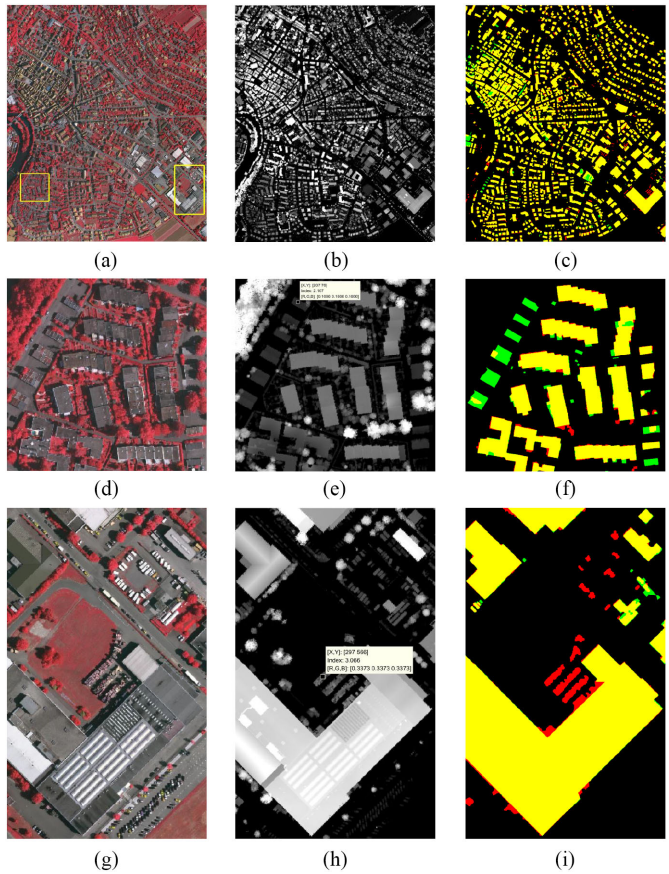


Fig. 11. Building extraction result of the whole Vaihingen dataset. (a) image; (b) nDSM; (c) pixel-based evaluation; (d)–(f) enlarged image, nDSM and pixel-based evaluation of selected bottom left yellow rectangle, respectively; (g)–(i) enlarged image, nDSM and pixel-based evaluation of selected bottom right yellow rectangle, respectively. The TP, FP, and FN are colored in yellow, red, and blue, respectively.

it has similar characteristics to buildings as well as high contrast due to occlusion from a shadow.

Columns 3rd to 5th of the last row of Table II shows that the overall average pixel-based completeness, correctness, and quality for our method were 94.3%, 93.8%, and 88.8%, respectively. These results show a significant improvement over the FED_2 method. The per-object level completeness, correctness, and quality only considered objects larger than 10 m^2 are 93.1%, 97.0%, and 90.1%. Although our method was lower in $Corr_{10}$ than FED_2, it achieves a higher Q_{10}, by nearly eight percentage points; 100% object-based accuracy was achieved on buildings larger than 50 m^2 for both methods.

While the proposed method achieves accurate building detection results, the drawback is that our method is prone to miss buildings smaller than 10 m^2, as indicated by the green colored areas in Fig. 10 row 1. This problem can be attributed to two causes. First, small buildings may have relatively small contrast in both the spectrum and elevation. Second, it is possible to identify the nonground point of the small target as the ground point at the stage of distinguishing the ground point from the nonground point.

An evaluation of the test dataset validates algorithms, but an appraisal of the robustness and practicability of a method relies on its performance on large datasets. To validate the effectiveness of our proposed method further, the proposed algorithm was tested on the whole dataset, which covers approximately 1.52 km^2 with 4527 $\times$ 5358 pixels in the orthophoto, containing over 1000 buildings. The computation was made by dividing the whole dataset into small tiles (1000 $\times$ 1000), with the same set of parameters adopted for each tile, and the accuracy evaluation results are shown in Fig. 11.

Fig. 11 shows the building extraction result of the whole Vaihingen dataset. Fig. 11(a) is the resampled aerial image. Fig. 11(b) is the nDSM image. Bright pixels indicate higher height, vice versa. Fig. 11(c) shows the pixel-based evaluation result, where TP, FP, and FN are colored in yellow, red, and blue, respectively. It can be seen from Fig. 11(c) qualitatively that most of the pixels are with yellow color, which represents the true positive, indicating that the proposed method also robust on large-scale data. Although the proposed building detection algorithm showed robustness, it still produced some building extraction errors. As shown in Fig. 11(c), there are some false-positive pixels. Meanwhile, there are also some false-negative pixels. These errors need to be further reduced to achieve more accurate building detection results. Two regions, as shown in Fig. 11(a) with a yellow rectangle, are selected for detailed analysis. The place at the bottom left of Fig. 11(a) has many omission errors (from now on referred to as $oa1$). A zoom

TABLE III
WHOLE VAIHINGEN DATASET EVALUATION RESULTS (%) FOR DETECTED BUILDINGS OF THE PROPOSED METHOD

$Comp_{ar}$	$Corr_{ar}$	Q_{ar}	$Comp$	$Corr$	Q
94.1	90.3	85.5	86.9	88.1	77.8
$Comp_{10}$	$Corr_{10}$	Q_{10}	$Comp_{50}$	$Corr_{50}$	Q_{50}
93.4	88.0	82.8	96.4	96.9	93.5

in view of $oa1$ is displayed in Fig. 11(d)–(f). The other place at the bottom right of Fig. 11(a) has many commission errors (from now on referred to as $ca1$). In the $oa1$ region, the elevation value is only about 2 m. In the spectrum, there is no visible contrast with the surrounding environment. Thus, the contrast between the spectral image and the elevation is relatively small, resulting in omission errors. Similarly, in the $ca1$ region, factory goods are stacked on the open ground, which has high contrast in the spectrum and elevation, so it was wrongly detected as buildings.

The accuracy verification table is shown in Table III. From the quantitative evaluation results in Table III, the overall pixel-based completeness, correctness, and quality for our method are 94.1%, 90.3%, and 85.5%, respectively, indicating that 94.1% of all building pixels in the reference map were correctly detected, whereas 90.3% of the detected building pixels were also building pixels in the reference map. The quality of our method in the large-scale study area is 85.5%, indicating that the method has specific adaptability, as well as revealing the practical potential of the proposed algorithm.

The per-object level completeness, correctness, and quality were 86.9%, 88.1%, and 77.8%. The per-object level completeness, correctness, and quality only considering objects larger than 10 m^2 are 93.4%, 88.0%, and 82.8%. The index Corr and the index $Corr_{10}$ has the same value. This confirms our previous observation that our method cannot detect buildings smaller than 10 m^2. Our proposed method tends to have higher completeness rather than correctness. The per-object level completeness, correctness, and quality only consider objects larger than 50 m^2 are 96.4%, 96.9%, and 93.5%, revealing the practical potential of the proposed algorithm.

D. Computational Complexity

The computation process of the proposed method is comprised of three major components: (a) guided image filtering, binarization and morphological processing, (b) UCM generation, and (c) adaptive iterative segmentation with data fusion. With raw LiDAR data preprocessed to nDSM, all the experiments are performed with MATLAB on a laptop with Intel Core i7-8550U CPU @ 1.80 GHz and 8.00 GB RAM. The most time-consuming part is the UCM generation, since it requires to consider spectral and height properties globally and locally. There are two ways to generate UCM, one is multiscale, which can get more accurate contours, but at the same time requires more computing time. The other is a single scale, which requires less computing resources while obtaining a sufficiently accurate contour. With an efficient implementation as described in [61], in our operating environment, it takes an average of 7.8 s to obtain a single-scale UCM of a 1001 × 762 size image. The proposed method needs to calculate the UCM twice, which are the UCM of the image and the UCM of the nDSM. For nDSM, only single-scale UCM was considered as its edges are inherently imprecise. For HRSI, two UCMs were generated, and the extracted building result were compared based on the per-area level quality. The detail of the processing time is shown in Table IV.

TABLE IV
COMPUTATIONAL COMPLEXITY ANALYSIS OF THE PROPOSED METHOD

Area	Size	Running time (seconds)						$Q_{ar}(\%)$
		UCM			Fusion	Other	Total	
		HSRI		nDSM				
Area1	925×691	MS	67.4	5.8	1.3	1.6	76.1	**88.7**
		SS	**6.5**				15.2	**88.7**
Area2	996×874	MS	90.2	8.2	1.4	0.5	100.3	**89.8**
		SS	**8.8**				18.9	89.4
Area3	1083×722	MS	79.7	7.7	1.2	0.5	89.1	**87.9**
		SS	**8.1**				17.5	**87.9**
Average	1001×762	MS	79.1	7.2	1.3	0.9	88.5	**88.8**
		SS	**7.8**				17.2	88.7

MS/SS are short for multiscale/ single scale testing, respectively.

Table IV has 11 rows and 9 columns. The first column shows the test area. The second column is the data size. From the third to eighth column shows running time for different components. The last column shows the Q_{ar} index. MS/SS are short for multiscale/single scale, respectively. The best values are highlighted in bold. It can be seen from Table IV that the accuracy of the proposed method differs a little at single-scale and multiscale, whereas the computation time differs largely. To extract buildings with 88.7% per-area quality, our method takes about 17.2 s to process a 1001 × 762 pixels area based on a single-scale UCM. The proposed method based on multiscale UCM have higher extraction quality, but it takes more time. If time complexity is taken into consideration, a single-scale approach is preferred. As described in Section IV-C2, it took many hours to train a model for U-Net/DeepLabv3. According to the description in [40], the total calculation time of the CNN model for each study area is about 15 min. It should be noted that they use higher resolution data. The average size is approximately 6.2 times larger than ours. Excluding training time, it takes 190 s to test the three Vaihingen area with the trained model, and the average is about 63.3 s. Our algorithm is relatively fast, with no training computation cost. According to the description in [18], the computation time of their methods takes around 2 min for a 5000 × 5000 pixels image. Our method is slower than them. But the two methods were tested on computers with different computing performance. These comparisons verify that our algorithm can meet the requirements of most applications.

V. CONCLUSION

In this article, a novel hierarchical automatic building extraction method with LiDAR data and HSRI fusion is proposed. The proposed method employs mid-level features and image segmentation to extract building boundaries efficiently and automatically. The idea is simple while practical. An adaptive iterative segmentation method is proposed to overcome the over- and under-segmentation problem. A data-fusion-based hierarchical overlay analysis is designed to extract buildings robustly and

efficiently. Compared with previous data-fusion-based methods, our proposed method had no parameters to set and needs no samples.

The performance of the proposed method was tested on Vaihingen dataset. The results show that it could not only extract partially occluded and shadowed buildings but also generates complex building shapes. The proposed method outperformed similar types of methods as shown on the ISPRS website, a quality of 88.8%. Detailed comparisons with four state-of-the-art methods shows that the proposed method with no sampling achieves competitive extraction results. The limitation of the proposed method was that few targets below $10\,m^2$ were detected because of low contrast at both elevation and spectral.

ACKNOWLEDGMENT

The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF). The authors would like to thank the people involved in the ISPRS building reconstruction benchmarks for the evaluation of results.

REFERENCES

[1] S. Noronha and R. Nevatia, "Detection and modeling of buildings from multiple aerial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 5, pp. 501–518, May 2001.

[2] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[3] G. Zhou, C. Song, J. Simmers, and P. Cheng, "Urban 3d gis from LiDAR and digital aerial images," *Comput. Geosci.*, vol. 30, no. 4, pp. 345–353, 2004.

[4] F. Rottensteiner and C. Briese, "A new method for building extraction in urban areas from high-resolution LiDAR data," in *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34. Ottawa, ON, Canada: Natural Resources Canada, 2002, pp. 295–301.

[5] K. Khoshelham, C. Nardinocchi, E. Frontoni, A. Mancini, and P. Zingaretti, "Performance evaluation of automated approaches to building detection in multi-source aerial data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 1, pp. 123–133, 2010.

[6] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 91–105, 2019.

[7] 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. Photogrammetry Remote Sens.*, vol. 130, pp. 294–307, 2017.

[8] Q. Zhang, R. Qin, X. Huang, Y. Fang, and L. Liu, "Classification of ultra-high resolution orthophotos combined with DSM using a dual morphological top hat profile," *Remote Sens.*, vol. 7, no. 12, pp. 16422–16440, 2015.

[9] T. T. Vu, F. Yamazaki, and M. Matsuoka, "Multi-scale solution for building extraction from LiDAR and image data," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 11, no. 4, pp. 281–289, 2009.

[10] Y. Li, H. Wu, R. An, H. Xu, Q. He, and J. Xu, "An improved building boundary extraction algorithm based on fusion of optical imagery and LiDAR data," *Optik-Int. J. Light Electron Opt.*, vol. 124, no. 22, pp. 5357–5362, 2013.

[11] J. Zhang, "Multi-source remote sensing data fusion: Status and trends," *Int. J. Image Data Fusion*, vol. 1, no. 1, pp. 5–24, 2010.

[12] F. Rottensteiner, "Automatic generation of high-quality building models from LiDAR data," *IEEE Comput. Graph. Appl.*, vol. 23, no. 6, pp. 42–50, Nov./Dec. 2003.

[13] M. Awrangjeb, M. Ravanbakhsh, and C. S. Fraser, "Automatic detection of residential buildings using LiDAR data and multispectral imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 5, pp. 457–467, 2010.

[14] L. Chen, S. Zhao, W. Han, and Y. Li, "Building detection in an urban area using LiDAR data and quickbird imagery," *Int. J. Remote Sens.*, vol. 33, no. 16, pp. 5135–5148, 2012.

[15] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Using the dempster–shafer method for the fusion of LiDAR data and multi-spectral images for building detection," *Inf. Fusion*, vol. 6, no. 4, pp. 283–300, 2005.

[16] A. Moussa and N. El-Sheimy, "A new object based method for automated extraction of urban objects from airborne sensors data," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 39, pp. 309–314, 2012.

[17] M. Gerke and J. Xiao, "Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 78–92, 2014.

[18] R. Qin and W. Fang, "A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization," *Photogrammetric Eng. Remote Sens.*, vol. 80, no. 9, pp. 873–883, 2014.

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

[20] M. Jarząbek-Rychard and H.-G. Maas, "Geometric refinement of als-data derived building models using monoscopic aerial images," *Remote Sens.*, vol. 9, no. 3, pp. 282–297, 2017.

[21] L. Cheng, J. Gong, X. Chen, and P. Han, "Building boundary extraction from high resolution imagery and LiDAR data," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 37, no. Part B3, pp. 693–698, 2008.

[22] Y. Li and H. Wu, "Adaptive building edge detection by combining LiDAR data and aerial images," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 37, no. Part B1, pp. 197–202, 2008.

[23] D. H. Lee, K. M. Lee, and S. U. Lee, "Fusion of LiDAR and imagery for reliable building extraction," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, pp. 215–225, 2008.

[24] M. Awrangjeb *et al.*, "Building detection in complex scenes thorough effective separation of buildings from trees," *Photogrammetric Eng. Remote Sens.*, vol. 78, no. 7, pp. 729–745, 2012.

[25] S. Gilani, M. Awrangjeb, and G. Lu, "An automatic building extraction and regularisation technique using LiDAR point cloud data and orthoimage," *Remote Sens.*, vol. 8, no. 3, pp. 258–284, 2016.

[26] A. Zarea and A. Mohammadzadeh, "A novel building and tree detection method from LiDAR data and aerial images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1864–1875, May 2016.

[27] M. M. Awad, "Toward robust segmentation results based on fusion methods for very high resolution optical image and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 2067–2076, May 2017.

[28] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[29] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *Electron. Imag.*, vol. 2016, no. 10, pp. 1–9, 2016.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.

[31] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from vhr remote sensing images combined with normalized dsms using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[32] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, pp. 144–161, 2018.

[33] H. He, J. Zhou, M. Chen, T. Chen, D. Li, and P. Cheng, "Building extraction from uav images jointly using 6d-slic and multiscale siamese convolutional networks," *Remote Sens.*, vol. 11, no. 9, pp. 1040–1072, 2019.

[34] A. D. Doulamis, N. D. Doulamis, K. S. Ntalianis, and S. D. Kollias, "Unsupervised semantic object segmentation of stereoscopic video sequences," in *Proc. Int. Conf. Inf. Intell. Syst.*, 1999, pp. 527–533.

[35] A. Bleiweiss and M. Werman, "Fusing time-of-flight depth and color for real-time segmentation and tracking," in *Proc. Workshop Dyn. 3D Imag.*, 2009, pp. 58–69.

[36] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. Di Stefano, and M. Vincze, "Multimodal cue integration through hypotheses verification for RGB-D object recognition and 6dof pose estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 2104–2111.

[37] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.

[38] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for RGB-D action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1651–1664, Aug. 2016.

[39] S. Durrieu, T. Tormos, P. Kosuth, and C. Golden, "Influence of training sampling protocol and of feature space optimization methods on supervised classification results," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2007, pp. 2030–2033.

[40] E. Maltezos, N. Doulamis, A. Doulamis, and C. Ioannidis, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042620.

[41] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from LiDAR data applying deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 155–159, Jan. 2019.

[42] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[43] P. Arbelaez, "Boundary extraction in natural images using ultrametric contour maps," in *Proc. Conf. Comput. Vision Pattern Recognit. Workshop*, 2006, pp. 182–182.

[44] P. Arbelaez, C. Fowlkes, and D. Martin, "The berkeley segmentation dataset and benchmark," Comput. Sci. Depart., Berkeley University, Berkeley, CA, USA, 2007. [Online]. Available: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds

[45] S. Crommelinck, R. Bennett, M. Gerke, M. Yang, and G. Vosselman, "Contour detection for UAV-based cadastral mapping," *Remote Sens.*, vol. 9, no. 2, pp. 171–183, 2017.

[46] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[47] J. Rouse Jr, R. Haas, J. Schell, and D. Deering, "Monitoring vegetation systems in the great plains with erts," NASA special publication, vol. 351, pp. 309–317, 1974.

[48] Y. Meng, Z. Hu, X. Chen, and J. Yao, "Subtracted histogram: Utilizing mutual relation between features for thresholding," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7415–7435, Dec. 2018.

[49] A. Perez, F. Lopez, J. Benlloch, and S. Christensen, "Colour and shape analysis techniques for weed detection in cereal fields," *Comput. Electron. Agriculture*, vol. 25, no. 3, pp. 197–212, 2000.

[50] D. Grigillo, M. Kosmatin Fras, and D. Petrovič, "Automatic extraction and building change detection from digital surface model and multispectral orthophoto," *Geodetski Vestnik*, vol. 55, no. 1, pp. 28–45, 2011.

[51] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[52] W.-H. Tsai, "Moment-preserving thresolding: A new approach," *Comput. Vision, Graph., Image Process.*, vol. 29, no. 3, pp. 377–393, 1985.

[53] R. C. Gonzales and R. E. Woods, "Digital image processing," 2nd ed. New Jersey: Prentice Hall, 2002.

[54] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral geoeye-1 imagery," *Photogrammetric Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, 2011.

[55] M. Cramer, "The dgpf-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2010, no. 2, pp. 73–82, 2010.

[56] 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. Photogrammetry Remote Sens.*, vol. 93, pp. 256–271, 2014.

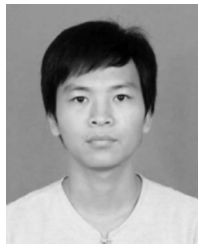
[57] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3d building reconstruction," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 1, no. 1, pp. 293–298, 2012.

[58] Y. Dong, L. Zhang, X. Cui, H. Ai, and B. Xu, "Extraction of buildings from multiple-view aerial images using a feature-level-fusion strategy," *Remote Sens.*, vol. 10, no. 12, pp. 1947–1976, 2018.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[60] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[61] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 328–335.



Shanxiong Chen received the B.S. degree in remote sensing science and technology in 2015 from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

His remote sensing research interests include building extraction, change detection, and deep learning.

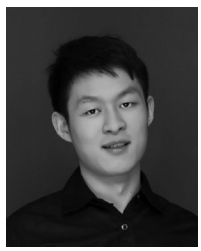
Wenzhong Shi received the Ph.D. degree in GISci and remote sensing from the University of Osnabrck, Vechta, Germany, in 1994.

He is currently a Chair Professor of Geographical Information Science and Remote Sensing with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong. He has authored and coauthored more than 400 scientific articles and 15 books. His current research interests include GISci and remote sensing, with focusing on analytics and quality control for spatial big data, object extraction and change detection from satellite images and LiDAR data, and integrated mobile mapping technology.

Prof. Shi was elected as an Academician of International Eurasian Academy of Sciences in 2019. He was a recipient of the State Natural Science Award from the State Council of China in 2007 and the Wang Zhizhuo Award from the International Society for Photogrammetry and Remote Sensing in 2012.

Mingting Zhou received the B.S. degree in photogrammetry and remote sensing in 2015 from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

Her research interests include high spatial resolution remote sensing image road detection, image processing, and deep learning.



Min Zhang received the M.S. degree in surveying engineering in 2014 from Wuhan University, Wuhan, China, where he currently working toward the Ph.D. degree with the School of Remote Sensing and Information Engineering.

His research interests include spatial data quality, change detection, and deep learning for remote sensing.

Pengfei Chen received the B.S., M.Sc., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012, 2015, and 2019, respectively. He is currently working toward the joint Ph.D. degree in geographical information system from The Hong Kong Polytechnic University, Hong Kong.

His research interests include spatial big data uncertainty analysis, human mobility modeling, and GeoAI.