remote sensing

MDPI

Article

Optimization of OpenStreetMap Building Footprints Based on Semantic Information of Oblique UAV Images

Xiangyu Zhuo [1,2,*], Friedrich Fraundorfer [1,3], Franz Kurz [1] and Peter Reinartz [1]

[1] Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany; fraundorfer@icg.tugraz.at (F.F.); franz.kurz@dlr.de (F.K.); Peter.Reinartz@dlr.de (P.R.)
[2] Remote Sensing Technology, Technische Universität München, 80333 Munich, Germany
[3] Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria
* Correspondence: xiangyu.zhuo@dlr.de; Tel.: +49-8153-28-4235

Received: 12 March 2018 ; Accepted: 16 April 2018; Published: 18 April 2018

Abstract: Building footprint information is vital for 3D building modeling. Traditionally, in remote sensing, building footprints are extracted and delineated from aerial imagery and/or LiDAR point cloud. Taking a different approach, this paper is dedicated to the optimization of OpenStreetMap (OSM) building footprints exploiting the contour information, which is derived from deep learning-based semantic segmentation of oblique images acquired by the Unmanned Aerial Vehicle (UAV). First, a simplified 3D building model of Level of Detail 1 (LoD 1) is initialized using the footprint information from OSM and the elevation information from Digital Surface Model (DSM). In parallel, a deep neural network for pixel-wise semantic image segmentation is trained in order to extract the building boundaries as contour evidence. Subsequently, an optimization integrating the contour evidence from multi-view images as a constraint results in a refined 3D building model with optimized footprints and height. Our method is leveraged to optimize OSM building footprints for four datasets with different building types, demonstrating robust performance for both individual buildings and multiple buildings regardless of image resolution. Finally, we compare our result with reference data from German Authority Topographic-Cartographic Information System (ATKIS). Quantitative and qualitative evaluations reveal that the original OSM building footprints have large offset, but can be significantly improved from meter level to decimeter level after optimization.

Keywords: building footprint; oblique UAV images; semantic segmentation; deep neural network

1. Introduction

OpenStreetMap (OSM) is a collaborative project for creating a free editable map of the world based on volunteered geographic information. It is able to provide free and updated geographic information despite restrictions on usage or availability of georeferenced data across most of the world. In recent years, OSM has widely expanded its coverage and gained increasing popularity in many applications. One example is the generation of 3D building models from OSM building footprints [1]. Therefore, the quality of building reconstruction and modeling strongly relies on the quality of building footprints. A detailed analysis for OSM building footprints [2] assessed a high completeness accuracy and a position accuracy of about 4 m on average for these data. Therefore, OSM building footprints can be safely regarded as a rough approximation of the real scene.

Though numerous approaches for building footprint generation have been developed, most of them exploit information from airborne imagery [3] or point cloud data [4], where the building footprints are represented by roofs and therefore usually mixed with overhangs. In contrast, building façades naturally contain critical information about footprints. In this sense, data that presents façade

information, such as oblique airborne imagery and terrestrial point clouds, can facilitate accurate building footprints' generation. Among different data sources, oblique UAV imagery stands out as it bridges the gap between aerial and terrestrial mapping, thus enabling data acquisition of both building roofs and façades simultaneously.

Given an initial hypothesis of the building boundary, the refinement with well-designed constraints plays a vital role in improving the accuracy of building footprints. The constraints usually come from the 3D features embedded in DSMs or point clouds as well as from the 2D features in images. For the case of oblique UAV images, 3D features such as lines [5] and planes usually have low geometric accuracy due to the change of the viewing directions in oblique images [6]. Apart from that, image features can also be employed as effective constraints. Traditional methods employed color features [7] in early stages, which are vulnerable to shadows and illumination. Some methods extracted building boundaries by detecting 2D lines [8] or corners [9]. However, the detected edges or corners have uncertain semantic meanings and therefore can only be used as weak evidence. In contrast, pixel-wise semantic image segmentation provides an effective solution to this problem. Various handcrafted features have been proposed in traditional machine learning based classification tasks. For instance, 2D image features, 2.5D topographic features and 3D geometric features are integrated in [10] for supervised classification using an SVM classifier. With the rapid development of deep neural networks, deep-learning based segmentation methods have demonstrated their conspicuous advantages in yielding reliable and robust semantic segmentation compared to traditional machine-learning segmentation methods. Deconvolution networks are firstly applied in [11] for building extraction from remote sensing images and demonstrate promising segmentation accuracy.

In this paper, we aim to refine the building footprint in OSM by deploying textural features from multi-view images as constraints. Based on the above discussion on previous research, we are motivated to use oblique UAV images as data sources for footprint generation. Constraints for solving the optimization problem are defined by building boundaries, i.e., the projection of the 3D building model on images, are expected to lie on the boundary between building façades and the ground. The contour evidence is extracted from pixel-wise semantic segmentation via deep convolution neural networks. The proposed method is composed of these steps: first, we geo-register the UAV images by matching them with high-accuracy aerial images; meanwhile, we perform semantic segmentation of UAV images using a Fully Convolutional Network (FCN) and extract the boundaries between building façades, roof and ground as contour evidence; then, we initialize a 3D building model of LoD 1 from the OSM footprints, followed by an optimization that integrates the contour evidence of multi-view images as a constraint. In the end, not only the footprints, but also the building heights get optimized. The proposed method is tested on different datasets. The accuracy of the optimized OSM footprints is evaluated by comparison with the ATKIS data, whose position accuracy is around 0.5 m [12].

The main innovations of this paper lie in the following aspects:

- The footprints addressed in previous research are the roof areas with overhangs. In contrast, our method is able to detect the real building footprints excluding roof overhangs, i.e., the edges where the building façades meet the ground.
- Instead of directly detecting buildings in 3D space, we introduce an optimization scheme using the image evidence from pixel-wise segmentation as a constraint, i.e., the image projection of the building model is encouraged to be identical to the building areas detected via pixel-wise image segmentation.
- Our method is able to refine simultaneously the building footprint and its height.

The paper is organized as follows: Section 2 gives a brief literature review on building footprint generation and points out the drawbacks of state-of-the-art methods regarding this topic. Section 3 describes our approach for OSM footprint optimization in detail. In Section 4, experiments on various datasets are carried out to validate the feasibility and robustness of the proposed method. Furthermore, the accuracy of the optimized building footprints is evaluated both qualitatively and quantitatively by

comparing with the ATKIS data. Finally, Section 5 discusses potentials and limitations of the proposed method and describes further applications.

2. Related Work

Obtaining accurate building footprints is of paramount importance in many applications such as urban planning or the real estate industry. Many attempts have been made to automatic ally extract building footprints in the last few decades. Traditionally, satellite imagery, aerial imagery and LiDAR data are among the most widely used data sources in this context. Some of the studies exploit solely information from images via pixel-based or object-based segmentation. Various segmentation descriptors [13] have been developed and other image features such as shadows [14] are also used for this purpose. Nevertheless, it is usually difficult to extract accurate building outlines from only images due to occlusion, shades and low illumination. Therefore, some studies explore the geometric features embedded in 3D data, e.g., Digital Surface Model (DSM) [15], point cloud reconstructed from images [16] and LiDAR data [17], or integrate the information from imagery and 3D data together [18,19]. However, these approaches are also prone to occlusions and have difficulties in detecting precise building boundaries. Moreover, as the building façades are inherently hardly visible in nadir-view remote sensing data, the aforementioned approaches actually extract building roofs rather than the real footprint without overhangs.

Considering that building façades convey vital information on building footprints, the data containing façade information, such as terrestrial data or oblique airborne imagery, can facilitate building footprint generation. A building detection method based on oblique aerial images is proposed in [20], while the façade information in terrestrial LiDAR point cloud is exploited [21,22]. As a bridge between terrestrial and airborne photogrammetry, the UAV stands out for its ability to achieve high spatial and temporal resolutions compared to traditional remote sensing platforms. Additionally, it has a great advantage in the application of building footprint generation for its ability in delivering information on both building roofs and façades. The accuracy of 3D building modeling based on both nadir and oblique UAV images is studied in [23], demonstrating that the integration of oblique UAV images can substantially increase the achievable accuracy comparing to traditional modeling using the terrestrial point cloud.

3. Methodology

In this section, we give a detailed account of the proposed approach for optimization of the OSM building footprint. The proposed workflow is comprised of the following steps: (1) geo-registration of oblique-view UAV images; (2) semantic segmentation of UAV images using a Fully Convolutional Network (FCN); and (3) optimization of the building model initialized from OSM footprints. We also point out the conditions and restrictions for the proposed method. Figure 1 depicts the workflow of the proposed approach. The external input data includes the building footprint extracted from OSM and the DSM reconstructed from aerial images, from which we can initialize a simple building sketch. In parallel, we create a ground truth dataset of UAV images and fine-tune the FCN-8s model. The trained network is then applied to segment UAV images. Finally, we optimize the building sketch by minimizing the chamfer distance between the building outline from projection and the contour evidence from image segmentation. Details of each step are explained in the following paragraphs.

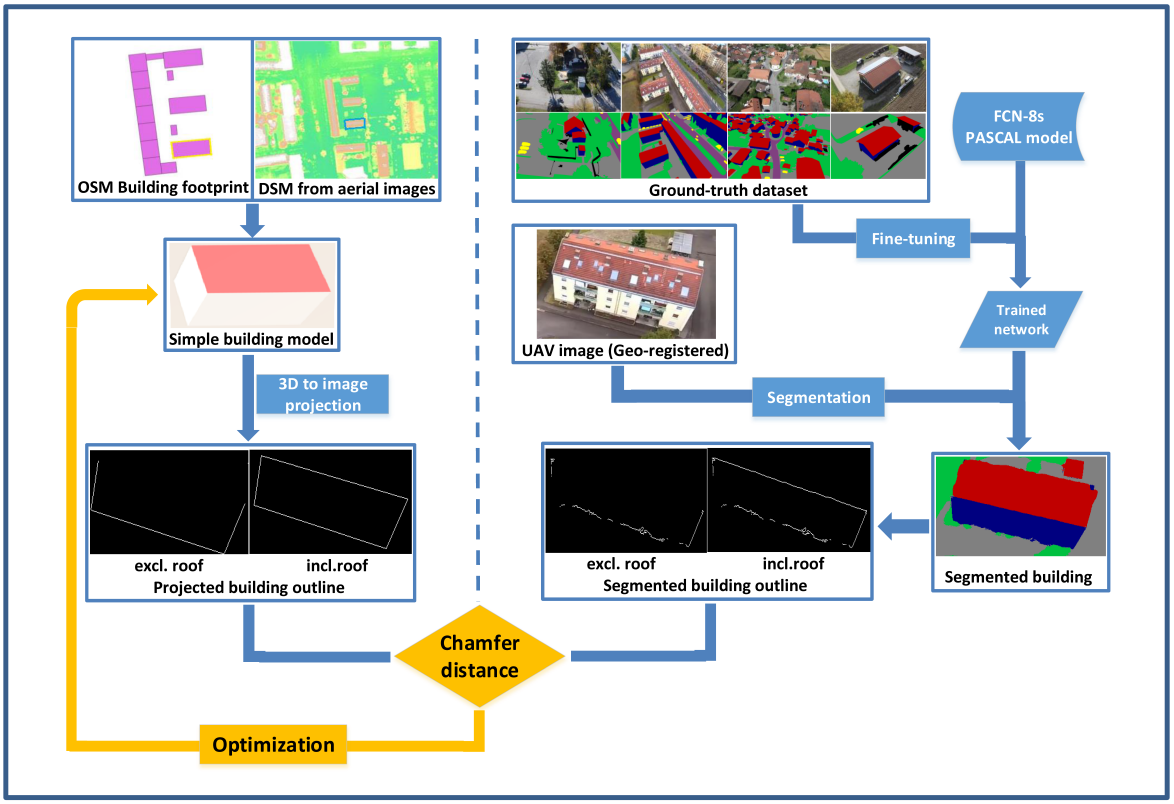


Figure 1. Workflow of the proposed method. External input data includes the building footprint extracted from OSM and DSM reconstructed from aerial images, from which a building sketch is initialized. Meanwhile, we create a ground truth dataset and fine-tune the FCN-8s model for image segmentation. We optimize the building sketch by minimizing the chamfer distance between the building outline from projection and the contour evidence from image segmentation.

3.1. Geo-Registration of UAV Images

The geo-registration of UAV images has already been discussed in numerous research works. One of the biggest challenges lies in the automated orientation of oblique UAV images. Towards this goal, various solutions have been proposed. Commercial softwares (e.g., Pix4D, Agisoft) and open source softwares (e.g., Bundler, PMVS, VisualSFM, MicMac) are widely used for matching and structure from motion (SFM) of oblique-view UAV images [24,25]. An image pyramid-based stratified matching method for matching nadir and oblique images from four-combined cameras in the step of structure from motion is proposed in [26], while an AKAZE interest operator-based matching strategy is presented in [27] for automatic registration of oblique UAV imagery to oblique aerial imagery.

Another challenge is the accurate geo-registration of UAV image blocks. Although UAV images are usually coupled with on-board GNSS/INS information, their absolute accuracy, solely 3 to 5 m in cases without a correction signal, is no higher than the one of OSM building footprints. In contrast, the accuracy of aerial photogrammetry can achieve a 10^{-2} m level, which notably exceeds OSM and is sufficient for our application. We adopt the approach proposed in [28] for co-registration between low-accuracy UAV images and high-accuracy aerial images. In short, the method assumes that the aerial images are geo-referenced and have common overlap with UAV images. First, the camera poses of sequential UAV images are solved via Structure From Motion (SFM), and then the nadir UAV images with the aerial images are matched using the proposed matching scheme and generate thousands of reliable image correspondences. Given accurate camera poses of the aerial images, the 3D coordinates of those common image correspondences can be calculated via image-to-ground projection. These 3D points are then adopted to estimate the camera poses of the corresponding nadir-view UAV images. In the end,

those UAV images with known camera poses are involved in a global optimization for camera poses of all UAV images. In this way, all UAV images are co-registered to the aerial images. The absolute accuracy of the geo-registered UAV images, according to the paper, can be as good as a 10^{-1} m level.

The aforementioned approach requires georeferenced aerial images of the surveyed area. In the absence of such reference images, the UAV images can be geo-registered with manually established Ground Control Points (GCPs), which may come from RTK GPS surveys or measurements from geo-spatial products (e.g., Basemap) with higher accuracy. For example, we can create some GCPs by measuring their planar coordinates (x, y) on an orthophoto and their elevation values z on a DSM. The GCPs with coordinates (x, y, z) can then be deployed to geo-register the UAV images dataset.

3.2. Semantic Segmentation of UAV Images

Extracting building outlines in images is essentially an issue of object recognition. Meanwhile, building outlines can also be viewed as class boundaries, which can be obtained from pixel-wise semantic segmentation. In this sense, the semantic segmentation of UAV images plays a crucial role in our pipeline as our optimization relies on the building outlines extracted from the segmentation as the constraint. Considering the fact that deep learning based methods significantly outperform traditional machine learning methods using handcrafted features, we attempt to train a deep neural network for the task of semantic segmentation.

Typically, a Convolutional Neural Network (CNN) is composed of an input layer, an output layer and multiple hidden layers in between. The hidden layers generally include convolutional layers, pooling layers, fully connected layers and normalization layers. In particular, convolutional layers apply nonlinear operations on the input with a set of adjustable parameters, which can be learned during the training process. The results are then passed to the next layer. Pooling layers take the outputs of neuron blocks of one layer and subsample them into a single neuron. A CNN may contain several convolutional layers and pooling layers. All the neurons in previous layers are then connected by a fully connected layer to each individual neuron in another layer. In order to adapt the classifier for a dense prediction, a solution has to consider these fully connected layers as convolutions with kernels covering their entire input regions [29]. Compared to the evaluation of the original classification network on overlapping input patches, the adapted classifier, namely FCN, is more efficient since computational burdens are shared by overlapping regions of patches. The network used in this paper is a modification of [29] by changing the number of outputs according to our demand. In our experiment, there are seven classes in total, i.e., building, roof, ground, road, vegetation, vehicle and clutter. The architecture of the neural network is depicted in Figure 2.

The aforementioned FCN also has shortcomings. First, its receptive field is as large as 32 pixels, resulting in segments with non-sharp boundaries and blob-like shapes. However, sharp boundaries between buildings and surroundings are preferred for us. Second, the prediction of FCN does not take the smoothness and the consistency of label assignments into consideration. To solve this problem, we plug in a Conditional Random Field (CRF) represented as Recurrent Neural Network (CRFasRNN) [30,31] at the end of the FCN, which combines the strengths of both the CNN and CRF based graphical model in one unified framework. In this model, unary energies are obtained from the FCN, which predict labels of pixel without considering the smoothness and the consistency of the label assignments; meanwhile, the pairwise energies provide a data-dependent smoothing term that encourages semantic label coherence for pixels with similar properties. Such combination enhances the consistency of the changes in labeling and in image intensity, resulting in sharp boundaries between adjacent segments.

For the compensation of limited training data, we implement data augmentation by cropping, rotating and scaling. Instead of training the network from scratch, we initialize the network with pre-trained weights from the FCN-8s model and then fine-tune it with our own ground-truth data. Details of implementation and parameter settings are described in Section 4.

In the end, we deploy the trained network to generate pixel-wise semantic segmentations of the images. Since the building footprint is defined as the boundary of a building where it meets the ground, the edge model of building footprints, denoted by L_0, can be therefore extracted as the boundary between the class "building" and the class "ground" from the semantically labeled images. It needs to be pointed out that semantic segmentation itself does not have a concept of an object; however, we are interested in class boundaries, which can be seen as objects that should be detected.

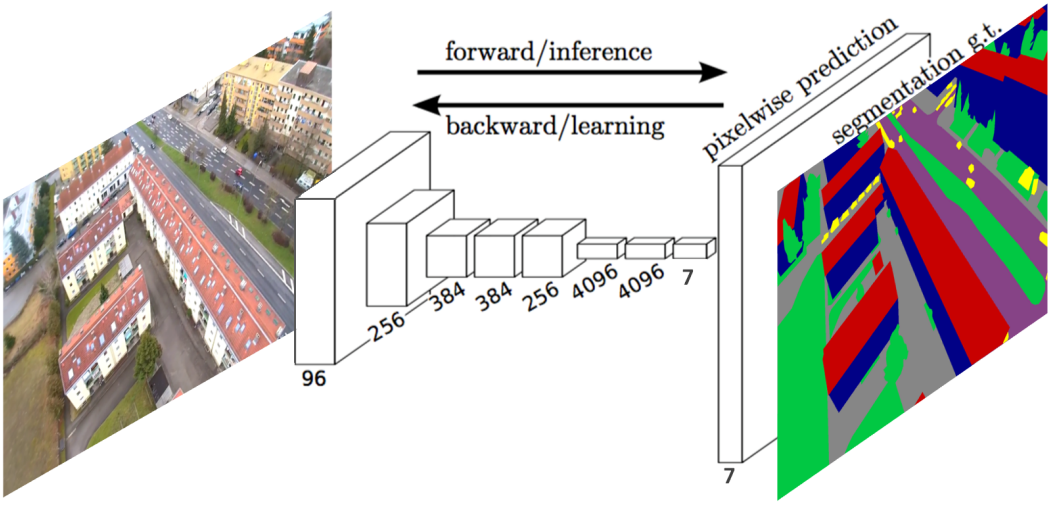


Figure 2. Architecture of the FCN network used in this paper.

3.3. Optimization of Building Footprints

We consider the building model as a polyhedron in 3D space featured by the building footprint $\mathcal{P}$ and height H. In particular, $\mathcal{P}$ is comprised of a set of vertices $\{P_1, \ldots, P_n \mid P_i \subset \mathbb{R}^3\}$, where P_i denotes the 3D coordinates $\begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix}$ of each corner of the building footprint, which can be directly extracted from OSM. For the cases without OSM height information, we can obtain the elevation of the ground at the foot of the building as well as the elevation of the roof, denoted by Z_{ground} and Z_{roof}, respectively, in a DSM from aerial imagery using a Top-Hat transform algorithm [32,33]. Then, the elevation Z_{ground} is assigned to Z_i and the difference $Z_{roof} - Z_{ground}$ is assigned to the building height H. At this point, a simple building model formulated by footprint and a uniform height has been established.

With geo-registered UAV images, the corresponding image point projection of $\mathcal{P}$ can be simply computed by means of ground-to-image projection, resulting in a 2D polygon denoted by $\mathcal{S} \subset \mathbb{R}^2$. Theoretically, if $\mathcal{P}$ is absolutely accurate, its projection $\mathcal{S}$ should be exactly identical with the building area in the UAV image. Under the assumption that the image segmentation result is reliable, the edge model of the polygon $\mathcal{S}$, denoted by L_1, should be close to the edge model L_0 extracted from the segmented image. Hereby, we adopt the Chamfer Distance [34] as a measurement for the difference between L_0 and L_1. We first cut off the area of L_0 from the image with a buffer zone of 100 pixels as the region of interest (ROI), and then generate the distance image of L_0. Afterwards, we superimpose L_1 on the distance image and the Chamfer Distance between L_0 and L_1 is defined as:

$$D(L_0, L_1) = \sqrt{\frac{1}{N} \sum_{l \in L_1} d_{I(l)}^2}, \tag{1}$$

where $d_{I(l)}$ stands for the distance values where the edge model L_1 hits the distance image of L_0, while N is the number of points in L_1.

One building may be present in multiple images from different viewing directions, and these images may have different segmentation accuracy in building areas. Despite the high interior accuracy

of the images' block, the image derived building contours in different images have a certain degree of variance, resulting in different Chamfer Distance. Therefore, it makes sense to take all cases into consideration. The adjustment of OSM footprints can be formulated as an energy minimization problem whose energy term is defined as Equation (2). In particular, $\mathbb{J}$ denotes the set of images that show the building, vector P_i denotes each vertex in the OSM footprint to be optimized, while H stands for the height of the building:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & E = \sum_{i \in \mathbb{J}} D(L_0^{(x,i)}, D(L_1^{(x,i)}), \\ & x = \{\{P_1, \ldots, P_n, H \mid P_i \subset \mathbb{R}^3, H \in \mathbb{R}\}\}. \end{aligned} \tag{2}$$

We solve the minimization problem using a modification of Powell's method [35,36], which performs sequential one-dimensional minimizations along each vector of the directions set. After optimization, the accuracy of the building footprint and height improves.

3.4. Application Conditions

It should be pointed out that building footprints can only be partially optimized in the presence of occlusions. Figure 3a shows the projection of an original OSM footprint in an aerial image, while Figure 3b highlights the edges that can be optimized with the proposed method. Figure 3c is an aerial image of the surveyed area with a slightly oblique view. Given that no additional images are available, only the visible building borders (marked with red lines) can be optimized. Towards the goal to achieve a complete optimization of the building footprints, it is thereby advised to acquire UAV images of the buildings of interest from different viewing directions so that all the façades are visible in the images.

(**a**)

(**b**)

(**c**)

Figure 3. Overview of optimizable building vertices in the presence of occlusions. Red lines are the projection of original OSM building footprints before optimization, highlighting the building edges that can be optimized.

4. Experiments

In order to validate the generalization ability of the proposed method, we test the methodology on four datasets with different building types. Furthermore, we compare the optimization results with the reference data. Qualitative and quantitative analyses are carried out to verify the accuracy of the optimized building footprint and height.

4.1. Data Description

- **Image data**

We collect images from four scenarios containing different types of buildings. The main characteristics of the datasets are listed in Table 1. In particular, Scenario A is targeted at optimization of footprints of individual buildings. To this end, 375 oblique images of an isolated cabin were captured by a UAV flight. The survey site lies on a bare agricultural land in Finning, Germany. Here, the building of interest is free from occlusions. The images were acquired by a Canon EOS-1DX camera mounted on a rotary wing platform at altitudes ranging from 20 m to 50 m above ground and with pitch angle of 40 $\sim$ 50°. The average ground sampling distance is 0.96 cm.

Scenario B presents a small kindergarten surrounded by trees and bushes in Oberpfaffenhofen, Germany. The dataset also serves for optimization of a footprint of an individual building, yet the

scene is more complex with the presence of occlusions and shades. 142 images were acquired by a Canon EOS-1DX camera, including 32 nadir-view images and 110 oblique-view images with a pitch angle of about 45°. The flight height ranges from 20 m to 45 m above ground, resulting in an average Ground Sampling Distance (GSD) of 1.09 cm.

In addition, we also exploit the possibility to optimize footprints of multiple buildings. With the increasing popularity of UAVs, more and more people are able to take photos or videos using their own drones and spontaneously share the data on the Internet with free access. In this context, Scenario C was established by extracting a series of frames from a YouTube video captured by a drone. The survey site is located in an urban residential area in Munich, Germany, containing many modern buildings, and most of them are partially occluded. The flight height is estimated to 40 m above ground with pitch angle around 40°. In total, 169 images with a GSD at image centers of 14.33 cm were extracted for the experiment. Each building can be visible or partly visible in 24–86 images, depending on its location in the survey area. It has to be noted that the low image resolution and the presence of occlusions cause difficulties for the subsequent optimization step.

Scenario D is an open dataset provided by the company senseFly [37]. The 37 oblique images were collected in a small village in Switzerland, featuring many traditional-style buildings surrounded or occluded by vegetation. Each building can be visible or partly visible in 31–37 images. The images were taken by a Canon PowerShot camera from about 100 m above the ground with a pitch angle of $-50°$, and the average image GSD is 5.46 cm.

- **OSM data**

The OSM data used in experiments have been downloaded on 21 January 2018. The footprints contain only the planar coordinates of building footprints but no height information. According to the detailed quality assessment for OSM building footprints data [2], the OSM footprints in our survey area (Munich) have a high completeness accuracy and a position accuracy of about 4 m in average. Therefore, they can be safely adopted to initialize a building model.

- **Reference data**

For evaluation of the experimental results, we take the German ATKIS (Amtliches Topographisch-Kartographisches Informationssystem) data as reference. ATKIS has been developed as a common project of the Working Committees of the Survey Administrations of the States of the Federal Republic of Germany (AdV), containing information of objects of the 'Real World' like roads, rivers or woodland [38]. The position accuracy of building footprint in ATKIS is ± 0.5 m [12]. It needs to be pointed out that ATKIS data is not available to the public in Germany, therefore we can only request the ATKIS footprint data for small areas as ground truth. Specifically, the ATKIS data used in the experiments were published on 27 January 2016 and the building data is formatted as LoD1 CityGML model, i.e., the value of building height describes the difference in meters between the highest point of the roof and the ground.

Table 1. Characteristics of the datasets used in the experiment. AA: automatically co-registered to aerial data; MA: manually co-registered to aerial data; -: pre-georeferenced.

Dataset	UAV Image						
	Date	Resolution (pix)	Height (m)	Pitch Angle	GSD (cm)	Number of Images per Building	Registration
A	10/2016	5184 × 3456	20 - 50	40–50°	0.96	375	MA
B	10/2016	5184 × 3456	20 - 45	45°, 90°	1.09	142	AA
C	01/2016	1296 × 728	40	50°	14.33	24–86	MA
D	06/2014	4000 × 3000	100	50°	5.46	31–37	MA

4.2. Geo-Registration of UAV Images

Due to payload limitations, UAVs are usually equipped with low-quality GNSS/IMUs and can therefore achieve a direct geo-referencing accuracy of merely 3–5 m. To improve the position accuracy of OSM building footprints, however, UAV images are expected to have higher accuracy than OSM. To this end, we made various attempts at improving the geo-registration accuracy of UAV images. In particular, Scenarios A, C and D are manually geo-registered using measurements from geo-spatial products that have higher accuracy than the OSM data, whereas Scenario B is geo-registered to aerial images in a fully automated way.

For Scenario A, we manually established some GCPs, whose planar coordinates (x,y) were measured on Bavaria DOP80 (digital orthophoto of 80 cm resolution, provided by Bavarian State Office for Survey and Geoinformation in Germany) and the elevation values z were extracted from DTK25 (Digital Topographic Model [DTM] of 25cm resolution, provided by Bavarian State Office for Survey and Geoinformation in Germany). These GCPs were then used to geo-register the UAV images dataset. Similarly, Scenario B was geo-registered using the GCPs measured from the orthophoto and DTM [39] reconstructed from the aerial images acquired by the DLR 3K camera system with cm-level accuracy [40]. In Scenario D, GCPs were measured on SWISSIMAGE 25 (digital orthophoto of 25 cm resolution) and swissNAMES3D (Topographical Landscape Model [TLM] of 0.2∼1.5 m accuracy).

Scenario B contains both nadir-view and oblique-view UAV images. Here, the corresponding nadir-view aerial images with cm-level global accuracy are also available, thus we co-registered the low-accuracy UAV images to the high-accuracy aerial images following the approach proposed in [28]. More specifically, we first solve the camera poses of the UAV images via Structure From Motion (SFM), and then match the nadir UAV images with nadir aerial images using the proposed matching scheme, resulting in thousands of reliable image correspondences. Since the aerial images are pre-georeferenced, 3D coordinates of those common image correspondences can be derived via image-to-ground projection of the aerial images, and these 3D points are then adopted to estimate the camera poses of the corresponding nadir-view UAV images. In the end, those UAV images with known camera poses are involved in a global optimization for camera poses of all UAV images. In this way, all UAV images get geo-registered.

We used the software Pix4Dmapper Pro (version 4.0.25) for the process and orientation of the UAV data. The mean reprojection errors of the four datasets are in the range of 0.15–0.2 pixels.

4.3. Semantic Image Segmentation Using CRFasRNN

For a robust and generalized training of the neural network, we collect training images evenly distributed from the four datasets, to ensure that different types of buildings are all included in the training dataset. The training images were manually labeled with seven categories: building, roof, ground, road, vegetation, vehicle and clutter. Among them, categories like building, roof and ground are of most interest for our application. In order to compensate for the shortage of training data, we implemented data augmentation by cropping, rotating and scaling the training data. Around 10,000 annotated images with a size of 300 × 300 pixels were generated for training.

The deep learning procedure was implemented under the framework Caffe [41]. Instead of training the network from scratch, we fine-tuned the FCN-8s PASCAL model from the Berkeley Vision and Learning Center (BVLC) on our own dataset. As the boundary between different classes is of interest for our application, we plugged in the CRF-RNN layer in order to achieve sharp edges at class borders. The training process started to converge at iteration 6000 and was stopped at iteration 74,000 before over-fitting. Figure 4 depicts the segmentation results of the trained network on the test data. Figure 4a,b are test images from Scenario A, while Figure 4e,f are the segmentation results. It can be seen that the roofs, façades, building and the surrounding clutter are basically correctly segmented; Figure 4c,d,g,h are respectively the original and segmented images from Scenario D, the segmentation in building areas is noisy due to shading and poor illumination. Figure 4i–p in the last two rows display segmentation results from Scenarios C and D with multiple buildings.

To conclude, despite a few incorrect segmentations in areas with complex textures or structures, the overall segmentation achieved a remarkable performance and yielded reliable image labels.

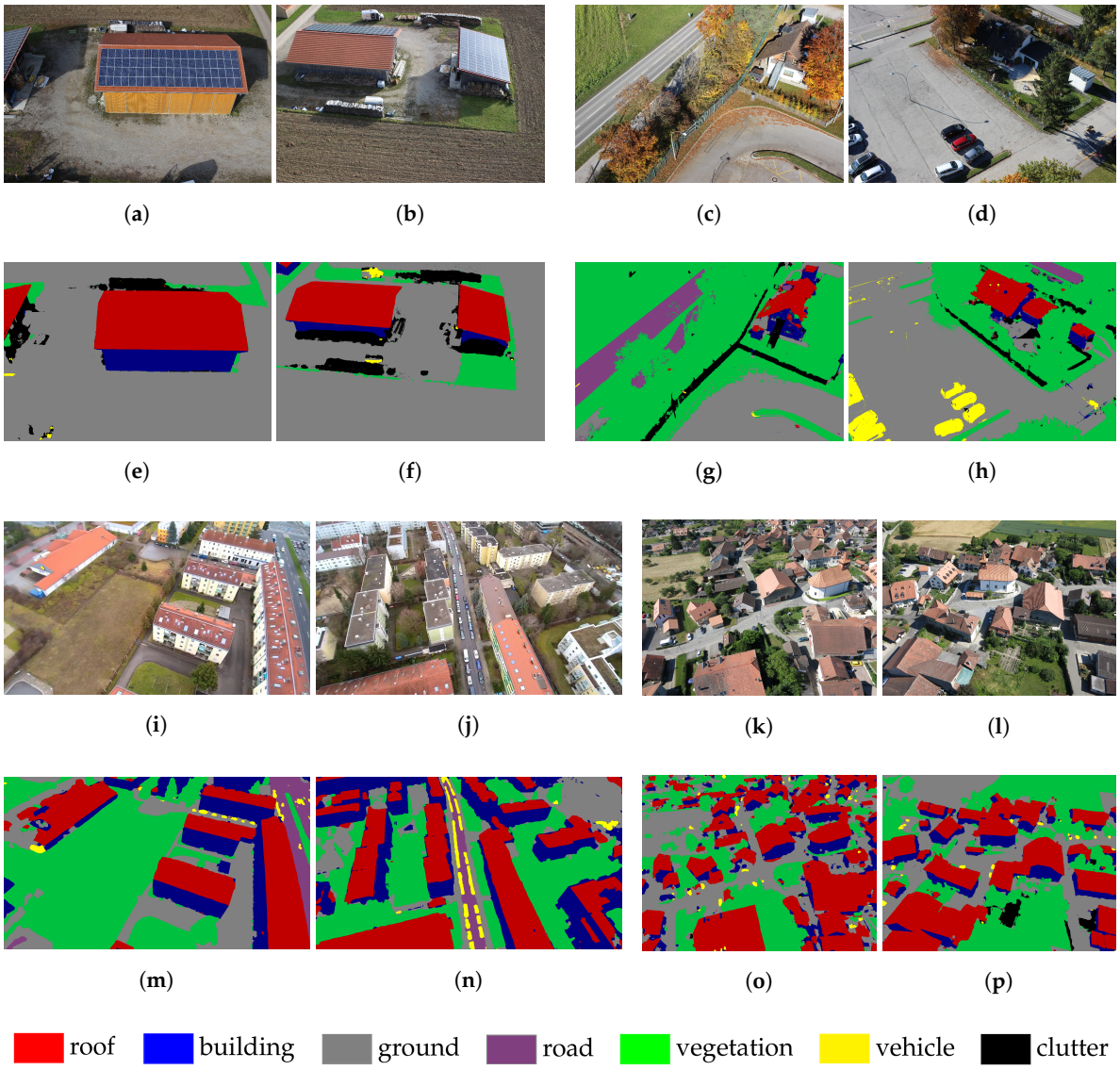


Figure 4. Segmentation results of four scenarios. (**a**) and (**b**) are test images of Scenario A while (**e**) and (**f**) are corresponding segmentation results; (**c**) and (**d**) are test images of Scenario B while (**g**) and (**h**) are corresponding segmentation results; (**i**) and (**j**) are test images of Scenario C while (**m**) and (**n**) are corresponding segmentation results; (**k**) and (**l**) are test images of Scenario D while (**o**) and (**p**) are corresponding segmentation results.

4.4. OSM Building Footprint Optimization

As we regard the building boundaries extracted from the segmented images as a constraint for optimization, it is crucial that the image segmentation, at least for the building and its surroundings, should yield accurate and reliable results. In practice, however, there are inevitably some poorly segmented images in a dataset, thus we need to select those images that satisfy those requirements:

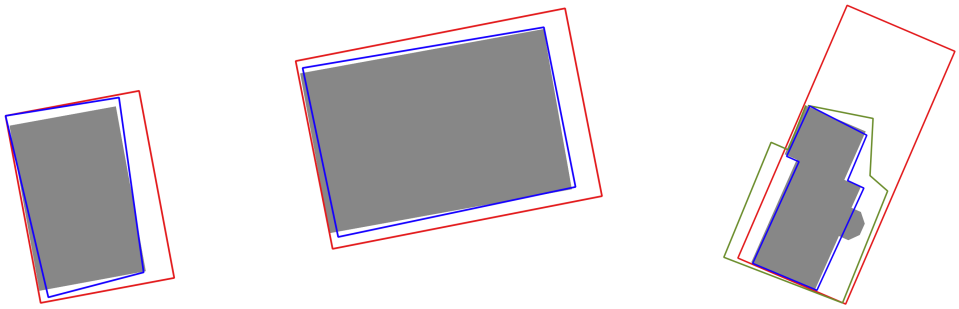
1. The segmented building areas have accurate boundaries;
2. Buildings are not occluded by vegetation or obstacles;
3. The selected images are expected to be taken from different viewpoints so that all vertices of the building footprint can be optimized.

Figure 5 demonstrates the results of footprint optimization process. Figure 5a–c are projections of original OSM footprints with the height extracted from DSM, while Figure 5d–f are projections of the optimized building sketch of Scenario A. Figure 5g,h are projections of original OSM footprints with the height extracted from DSM, while Figure 5j,k are projections of the optimized building sketch of Scenario B. Figure 5i,l illustrate a combined footprint of two adjacent buildings with different heights before and after optimization, therefore only the footprint get optimized. The original projections of the footprints extracted from OSM with the height from DSM are highlighted by the red lines, which have large position shift with respect to the building. The projections of the footprints after optimization using the proposed method are marked by blue lines, which fit precisely the building borders. It is evident that the image projection accuracy of the optimized footprints have improved conspicuously compared to the original OSM footprints.

Figure 5. Image projections of building sketch before and after optimization. Red lines are projections of original OSM footprints with height measured from DSM, blue lines show projections of optimized building footprints and heights, and the green line shows a combined footprint for two buildings with different height.

For the visualization of the absolute position accuracy of the optimized footprints, we overlap the footprints before and after optimization together with ATKIS data. As shown in Figure 6, the gray areas are reference footprints from ATKIS data, red lines indicate original footprints extracted from OSM, and blue lines show the footprints after optimization using the proposed method. All footprints are overlapped together in the same coordinate reference system. To be specific, Figure 6a shows the footprints of the two cabins from Scenario A. All corners of the larger building on the right get optimized with a significant improvement in position accuracy. It should be pointed out that the smaller building on the left is not our main target and therefore only appears in a few images, still all its three visible corners got optimized

with satisfying accuracy. Figure 6b shows the footprints of a house from Scenario B, which is composed of two small adjacent houses. The building footprint in OSM, however, is simplified to a rectangular shape with a large position shift. As a correct hypothesis of the building shape is the prerequisite for reasonable optimization, we extracted eight corners of the roof based on the corresponding orthophoto and DSM. The new building footprint, colored in green in Figure 6b, was used as the initial value for optimization. The optimized footprint, highlighted in blue, matches the reference data well. The experimental results demonstrate that, given the correct hypothesis of the building shape, our method is able to efficiently optimize the footprint of individual buildings even if the initial values are far from accurate.



(**a**) (**b**)

Figure 6. Optimization for multiple buildings. (**a**) and (**b**) show the result of Scenario A Scenario B respectively; gray areas represent the reference footprints from ATKIS data, red lines indicate original footprints extracted from OSM, blue lines show the footprints after optimization using the proposed method, and green represents the initial lines for optimization.

In contrast, Scenario C and Scenario D feature multiple buildings that are partly occluded. As a consequence, only the visible building corners may get optimized. In addition, the performance of optimization is also affected by the quality of image segmentation.

Figure 7a,b illustrate the optimization results of Scenario C. Blue lines indicate an overall projection of all optimized buildings, while the optimization for the rest of the buildings failed as a result of severe occlusions or poor segmentation. It can be seen that more than half of the visible building corners were successfully optimized despite the low image resolution, and the estimated building height aligns well with the border between the roof and the building.

Within Scenario D, most buildings are surrounded by thick vegetation, and we extract only the boundary between the ground and the building. For that reason, there are few effective contours available for optimization. Figures 7c–f show some of the optimized building footprints, where the red lines correspond to the original OSM building footprints and blue lines refer to the optimized building footprints. It should be noted that the OSM footprints data for rural areas exhibit much larger errors than in the urban area. Nevertheless, our method still achieves accurate optimization results for visible building edges.

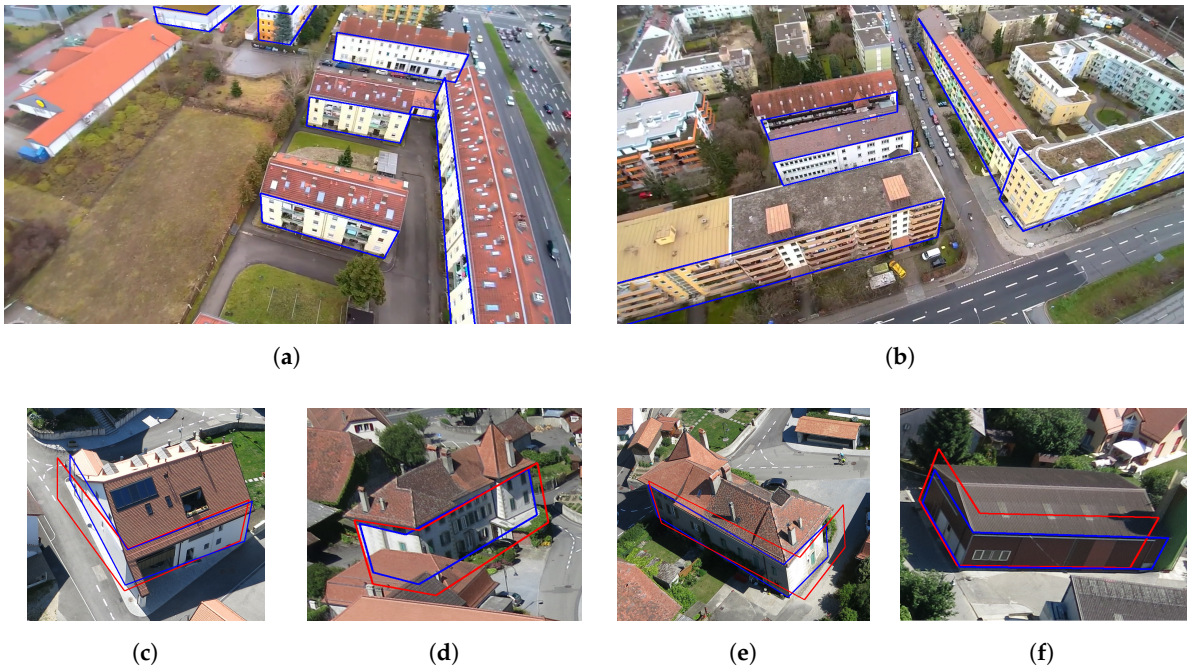


Figure 7. Optimization for multiple buildings. Red lines are the projections of the original OSM building footprints while blue lines correspond to the optimized building footprints; (**a**) and (**b**) give an overall view of all the optimized buildings in Scenario C; (**c**) to (**f**) enumerate some of the optimized buildings of Scenario D.

4.5. Accuracy Evaluation of Building Position and Height

Apart from the visual comparison of the results, we performed a quantitative analysis of the results. Following the evaluation approach in [2], we investigate the position accuracy of building footprints by calculating the average distance between the corresponding vertices pair from the optimized footprints and the reference data. In this sense, only the vertices appearing in both datasets can be compared.

In order to evaluate the optimization accuracy quantitatively, we compare the footprints before and after optimization with reference to ATKIS data. The results of Scenario A, B, and C are listed in Table 2. The second column lists the optimized buildings in each scenario, and for each building footprint, we manually measure the coordinates of each vertex and calculate the distance to the corresponding vertex in ATKIS data. The column Initial lists the errors in the x- and y-directions as well as the Euclidean distance of each vertex from the original footprint, whereas the column Optimized reports the errors of the optimized building footprint. Scenario C contains a number of optimizable buildings, from which six buildings with their optimized vertices are randomly selected as representatives. The value average shows the average distance of all building vertices of each Scenario.

Given that the building footprints in ATKIS have an average accuracy of ± 0.5 m, we can draw the conclusion that the accuracy of the building footprints has substantially increased after optimization using our method.

Table 2. Position errors of building footprints before and after optimization. The column Initial lists the errors in the x- and y-directions as well as the Euclidean distance of each vertex of the original footprints; the column Optimized reports the errors of the optimized building footprints.

Scenario	Building	Initial			Optimized		
		ΔX (m)	ΔY (m)	Distance (m)	ΔX (m)	ΔY (m)	Distance (m)
A	1	0.207	−1.213	1.230	0.623	−0.328	0.704
		−0.374	0.942	1.014	0.145	0.427	0.451
		1.601	1.545	2.225	0.020	0.137	0.139
		2.247	−0.521	2.307	0.276	0.165	0.322
	2	1.733	1.157	2.084	0.233	0.657	0.697
		2.114	−0.511	2.175	−0.164	−0.097	0.190
		0.080	−0.899	0.902	0.655	−0.481	0.813
	average			1.705			0.474
B	1	0.275	0.043	0.278	0.395	−0.037	0.397
		0.758	1.337	1.537	0.118	−0.383	0.401
		2.635	0.436	2.671	0.345	−0.074	0.353
		2.756	−0.400	2.785	0.326	−0.100	0.341
		2.848	−1.306	3.133	0.233	−0.008	0.233
		−2.807	0.522	2.856	0.142	−0.102	0.175
		−2.632	1.763	3.168	0.218	−0.227	0.314
		0.342	0.461	0.574	0.192	−0.209	0.284
	average			2.125			0.312
C	1	0.039	−2.073	2.073	−0.029	−0.682	0.683
		−0.078	−1.927	1.928	−0.174	−0.540	0.567
		0.695	−1.509	1.661	0.027	0.201	0.203
	2	−0.303	0.271	0.406	−0.252	−0.250	0.355
		0.397	0.406	0.568	0.155	0.116	0.194
	3	0.492	−1.415	1.498	−0.144	0.748	0.761
		−0.053	−1.412	1.413	0.530	0.437	0.687
	4	0.708	−1.944	2.069	0.471	−0.451	0.651
		0.303	−1.917	1.941	−0.387	−0.668	0.772
	5	0.543	−1.638	1.726	0.365	−0.417	0.554
		0.150	−1.368	1.376	0.239	−0.382	0.451
	6	0.144	0.579	0.596	0.105	0.320	0.337
		0.423	0.501	0.656	0.282	0.360	0.457
	average			1.378			0.513

Additionally, during the optimization of the planar coordinates of building footprints, our method is also able to estimate the height of the wall, i.e., the height from the top of the building façade to the ground, which cannot be directly measured from LiDAR data or DSM from aerial imagery. As aforementioned, the height value in ATKIS data describes the distance from the top of the roof to the ground, hence it only makes sense to evaluate buildings with flat roofs. Applied to our dataset, there remain only five optimized buildings with flat roofs. Table 3 compares the height values of these optimized buildings with the height measurements from ATKIS data. It can be demonstrated that the building heights are accurately estimated with an absolute error $\leq 10\%$.

Table 3. Accuracy evaluation of optimized building height.

Building	Optimized H (m)	ATKIS H (m)	Error H (m)
1	3.20	3.5	−0.30
2	10.96	11.53	−0.57
3	17.5	17.83	−0.33
4	19.7	21.00	−1.30
5	5.24	5.79	−0.55

5. Discussion

In this paper, we present a novel framework for optimizing OSM building footprints based on the contour information derived from deep learning-based semantic segmentation of UAV images. Through our methodology, the position accuracy of optimized building footprints has been improved from meter-level to decimeter-level, which is comparable with the accuracy of ATKIS data.

The applicability of the proposed method depends on the following prerequisites:

- Towards the goal of improving the absolute position accuracy of OSM building footprints, the UAV images are supposed to be accurately geo-referenced. However, it is also practical to simply align the OSM building footprint data to the users' local reference system.
- Targeted at optimization of the complete building footprint, it is advised to design the UAV flight path to surround the buildings of interest; otherwise, only the visible building edges can be optimized.
- Since we use UAVs to acquire image data, our approach is suitable for regional improvement for buildings of interest. In most large-scale applications such as navigation, web-based visualization and city planning, the accuracy of OSM footprints is already sufficient. Accurate footprints (with sub-meter level accuracy) are usually needed for specific buildings of interest, and our approach can play its role in such cases.

The merits of the proposed method mainly lie in four aspects:

- In many other regions of the world, there is no such high-quality footprint data like ATKIS; even in Germany, the ATKIS data is not freely accessible to the public. Our approach opens up the possibility to generate high-accuracy building footprints from OSM with comparable accuracy as ATKIS data.
- The realistic building footprints excluding roof overhangs can be detected, i.e., the edges where the building façades meet the ground, whereas the footprints addressed in previous research are essentially the building roof including overhangs.
- The height information of buildings can be simultaneously refined with the building footprints.
- The proposed method has good generalization ability, as it can optimize not only a single building, but also multiple buildings with high tolerance for the spatial resolution of images.

Based on the optimized building footprint and building height, we can establish a building sketch of LoD 1, which can be further applied in building information modeling (BIM).

6. Conclusions

In summary, we exploit the façades' information in oblique UAV images to optimize OSM building footprints. The framework consists of three main aspects: first, a simplified 3D building model of Level of Detail 1 (LoD 1) is initialized using the footprint information from OSM and the elevation information from the Digital Surface Model (DSM). Subsequently, a deep neural network is trained for pixel-wise semantic image segmentation and the building boundaries are extracted as contour evidence. Finally, the initial building model is optimized by integrating the contour evidence from multi-view images as a constraint, resulting in a refined 3D building model with optimized footprints and height. The result reveals the great potential of oblique UAV images in building reconstruction and modeling.

Acknowledgments: This research was funded by the German Academic Exchange Service (DAAD:DLR/DAAD Research Fellowship Nr. 50019750) for Xiangyu Zhuo.

Author Contributions: Friedrich Fraundorfer and Xiangyu Zhuo conceived and designed the experiments; Xiangyu Zhuo performed the experiments and analyzed the data; Franz Kurz contributed to the data acquisition and pre-processing; Xiangyu Zhuo wrote the manuscript and all authors contributed to reviewing the paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

GNSS	Global Navigation Satellite System
INS	Inertial Navigation System
GPS	Global Positioning System
RTK	Real Time Kinematic

References

1. Goetz, M. Towards generating highly detailed 3D CityGML models from OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 845–865, doi:10.1080/13658816.2012.721552.
2. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719, doi:10.1080/13658816.2013.867495.
3. Müller, S.; Wilhelm Zaum, D. Robust building detection in aerial images. *Int. Arch. Photogramm. Remote Sens.* **2005**, *36*, 143–148.
4. Zhou, Q.Y.; Neumann, U. 2.5 d dual contouring: A robust approach to creating building models from aerial lidar point clouds. In *Computer Vision—ECCV 2010, Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010*; Springer: Berlin, Germany, 2010; pp. 115–128.
5. Lafarge, F.; Mallet, C. Creating large-scale city models from 3D-point clouds: A robust approach with hybrid representation. *Int. J. Comput. Vis.* **2012**, *99*, 69–85, doi:10.1007/s11263-012-0517-8.
6. Xiao, J.; Gerke, M. Building footprint extraction based on radiometric and geometric constraints in airborne oblique images. *Int. J. Image Data Fus.* **2015**, *6*, 270–287.
7. Sirmacek, B.; Unsalan, C. Building detection from aerial images using invariant color features and shadow information. In Proceedings of the 23rd International Symposium on Computer and Information Sciences, Istanbul, Turkey, 27–29 October 2008; pp. 1–5. doi:10.1109/ISCIS.2008.4717854.
8. Zhu, Q.; Jiang, W.; Zhang, J. Feature line based building detection and reconstruction from oblique airborne imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *40*, 199.
9. Wegner, J.D.; Thiele, A.; Soergel, U. Fusion of optical and InSAR features for building recognition in urban areas. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2009**, *38*, W4.
10. Gevaert, C.; Persello, C.; Sliuzas, R.; Vosselman, G. Informal settlement classification using point-cloud and image-based features from UAV data. *ISPRS J. Photogramm. Remote Sens.* **2017**, *125*, 225–236.
11. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1835–1838.
12. Müller, W.; Seyfert, E.; Brandenburg, L. Quality assurance for 2.5-D building data of the ATKIS DLM 25/2. *Int. Arch. Photogramm. Remote Sens.***1998**, *32*, 411–416.
13. Dornaika, F.; Moujahid, A.; Merabet, Y.E.; Ruichek, Y. Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors. *Expert Syst. Appl.* **2016**, *58*, 130–142.
14. Ok, A.O.; Senaras, C.; Yuksel, B. Exploiting shadow evidence and iterative graph-cuts for efficient detection of buildings in complex environments. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *XL-1/W1*, 269–274, doi:10.5194/isprsarchives-XL-1-W1-269-2013.
15. Dini, G.R.; Jacobsen, K.; Heipke, C. Delineation of building footprints from high resolution satellite stereo imagery using image matching and a GIS database. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *XL-1/W1*, 81–85, doi:10.5194/isprsarchives-XL-1-W1-81-2013.
16. Dai, Y.; Gong, J.; Li, Y.; Feng, Q. Building segmentation and outline extraction from UAV image-derived point clouds by a line growing algorithm. *Int. J. Digit. Earth* **2017**, *10*, 1077–1097, doi:10.1080/17538947.2016.1269841.
17. Awrangjeb, M.; Fraser, C.S. Automatic segmentation of raw LIDAR data for extraction of building roofs. *Remote Sens.* **2014**, *6*, 3716–3751, doi:10.3390/rs6053716.
18. Bittner, K.; Cui, S.; Reinartz, P. Building extraction from remote sensing data using fully convolutional networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 481.
19. Li, Y.; Wu, H.; An, R.; Xu, H.; He, Q.; Xu, J. An improved building boundary extraction algorithm based on fusion of optical imagery and LIDAR data. *Optik Int. J. Light Electron Opt.* **2013**, *124*, 5357–5362.
20. Nex, F.; Rupnik, E.; Remondino, F. Building footprints extraction from oblique imagery. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *II-3/W3*, 61–66.
21. Hammoudi, K.; Dornaika, F. Extracting building footprints from 3D point clouds using terrestrial laser scanning at street level. In Proceedings of the Object Extraction for 3D City Models, Road Databases and Traffic Monitoring—Concepts, Algorithms and Evaluation (CMRT09), Paris, France, 3–4 September 2009.
22. Yang, B.; Wei, Z.; Li, Q.; Li, J. Semiautomated building facade footprint extraction from mobile LiDAR point clouds. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 766–770, doi:10.1109/LGRS.2012.2222342.

23. Vacca, G.; Dessi, A.; Sacco, A. The use of nadir and oblique UAV images for building knowledge. *ISPRS Int. J. GeoInf.* **2017**, *6*, 393.
24. Lingua, A.; Noardo, F.; Spano, A.; Sanna, S.; Matrone, F. 3D model generation using oblique images acquired by UAV. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 107–115.
25. Aicardi, I.; Chiabrando, F.; Grasso, N.; Lingua, A.M.; Noardo, F.; Spanò, A. Uav photogrammetry with oblique images: First analysis on data acquisition and processing. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 835–842, doi:10.5194/isprs-archives-XLI-B1-835-2016.
26. Feifei, X.; Zongjian, L.; Dezhu, G.; Hua, L. Study on construction of 3D building based on UAV images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *39*, 469–473.
27. Onyango, F.; Nex, F.; Peter, M.; Jende, P. Accurate estimation of orientation parameters of UAV images through image registration with aerial oblique imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 599.
28. Zhuo, X.; Koch, T.; Kurz, F.; Fraundorfer, F.; Reinartz, P. Automatic UAV image geo-registration by matching UAV images to georeferenced image aata. *Remote Sens.* **2017**, *9*, 376.
29. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. doi:10.1109/TPAMI.2016.2572683.
30. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H.S. Conditional random fields as recurrent neural networks. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015.
31. Arnab, A.; Jayasumana, S.; Zheng, S.; Torr, P.H.S. Higher order conditional random fields in deep neural networks. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
32. Shih, F.Y. *Image Processing and Mathematical Morphology: Fundamentals and Applications*; CRC Press: Boca Raton, FL, USA, 2009; 439p.
33. Mongus, D.; Žalik, B. Parameter-free ground filtering of LiDAR data for automatic DTM generation. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 1–12.
34. Borgefors, G. Distance transformations in digital images. *Comput. Vis. Graph. Image Process.* **1986**, *34*, 344–371.
35. Powell, M.J. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput. J.* **1964**, *7*, 155–162.
36. Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T. *Numerical Recipes*; Cambridge University Press: Cambridge, UK, 1989.
37. senseFly. Oblique mapping of a village. URL: https://www.sensefly.com/education/datasets/.
38. Müller, W.; Seyfert, E. Intelligent imagery system: A proposed approach. *Int. Arch. Photogramm. Remote Sens.* **2000**, *33*, 710–717.
39. D'Angelo, P.; Reinartz, P. Semiglobal Matching Results on the ISPRS Stereo Matching Benchmark; High-Resolution Earth Imaging for Geospatial Information; ISPRS Hannover Workshop: Hannover, Germany, 2011.
40. Kurz, F.; Rosenbaum, D.; Meynberg, O.; Mattyus, G.; Reinartz, P. Performance of a real-time sensor and processing system on a helicopter. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2014**, *40*, 189–193, doi:10.5194/isprsarchives-XL-1-189-2014.
41. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093* **2014**, doi:10.1145/2647868.2654889.