remote sensing

MDPI

Article

Building Extraction and Floor Area Estimation at the Village Level in Rural China Via a Comprehensive Method Integrating UAV Photogrammetry and the Novel EDSANet

Jie Zhou [1,2], Yaohui Liu [3,4,5,*], Gaozhong Nie [1,2], Hao Cheng [6], Xinyue Yang [3], Xiaoxian Chen [3] and Lutz Gross [5]

1 Institute of Geology, China Earthquake Administration, Beijing 100029, China
2 Key Laboratory of Seismic and Volcanic Hazards, China Earthquake Administration, Beijing 100029, China
3 School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China
4 College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
5 School of Earth and Environmental Sciences, The University of Queensland, Brisbane, QLD 4072, Australia
6 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
* Correspondence: liuyaohui20@sdjzu.edu.cn

Abstract: Dynamic monitoring of building environments is essential for observing rural land changes and socio-economic development, especially in agricultural countries, such as China. Rapid and accurate building extraction and floor area estimation at the village level are vital for the overall planning of rural development and intensive land use and the "beautiful countryside" construction policy in China. Traditional in situ field surveys are an effective way to collect building information but are time-consuming and labor-intensive. Moreover, rural buildings are usually covered by vegetation and trees, leading to incomplete boundaries. This paper proposes a comprehensive method to perform village-level homestead area estimation by combining unmanned aerial vehicle (UAV) photogrammetry and deep learning technology. First, to tackle the problem of complex surface feature scenes in remote sensing images, we proposed a novel Efficient Deep-wise Spatial Attention Network (EDSANet), which uses dual attention extraction and attention feature refinement to aggregate multi-level semantics and enhance the accuracy of building extraction, especially for high-spatial-resolution imagery. Qualitative and quantitative experiments were conducted with the newly built dataset (named the rural Weinan building dataset) with different deep learning networks to examine the performance of the EDSANet model in the task of rural building extraction. Then, the number of floors of each building was estimated using the normalized digital surface model (nDSM) generated from UAV oblique photogrammetry. The floor area of the entire village was rapidly calculated by multiplying the area of each building in the village by the number of floors. The case study was conducted in Helan village, Shannxi province, China. The results show that the overall accuracy of the building extraction from UAV images with the EDSANet model was 0.939 and that the precision reached 0.949. The buildings in Helan village primarily have two stories, and their total floor area is 3.1×10^5 m^2. The field survey results verified that the accuracy of the nDSM model was 0.94; the RMSE was 0.243. The proposed workflow and experimental results highlight the potential of UAV oblique photogrammetry and deep learning for rapid and efficient village-level building extraction and floor area estimation in China, as well as worldwide.

Keywords: building extraction; floor area estimation; rural China; deep learning; UAV

Citation: Zhou, J.; Liu, Y.; Nie, G.; Cheng, H.; Yang, X.; Chen, X.; Gross, L. Building Extraction and Floor Area Estimation at the Village Level in Rural China Via a Comprehensive Method Integrating UAV Photogrammetry and the Novel EDSANet. *Remote Sens.* **2022**, *14*, 5175. https://doi.org/10.3390/rs14205175

Academic Editors: Giovanni Laneve, Chenghai Yang, Wenjiang Huang and Yingying Dong

Received: 4 September 2022
Accepted: 13 October 2022
Published: 16 October 2022

1. Introduction

Homesteads are an important part of basic rural geographic information and multi-functional complex spaces for rural residents [1–3]. With the advancement of urban–rural economic integration in China, many farmers have migrated to cities. From 2000 to 2016, the rural resident population in China decreased from 808 to 589 million (a decline of 27.1%) [4]. The migration of rural residents to cities reduces the area of rural homestead land. However,

due to the free acquisition and use of the homestead system, local governments launch new rural construction without proper scientific planning, which has increased the area of idle rural homesteads by 20.6% [5], from 0.99 to 1.21 million km^2 [4]. Compared to developed cities, rural areas are dominated by low-rise buildings, and the excessive occupation of land resources by farmers affects land-use efficiency [6]. To promote rural development, the Chinese government has proposed "beautiful countryside" construction. In-depth investigations should be conducted on the living conditions of farmers, and land-use areas in rural areas should be rationally planned. Field surveys can provide accurate information about the residents of farms but require time and labor. Moreover, land use for rural homesteads in developing countries is usually scattered, resulting in barriers to the acquisition of rural building information. Therefore, additional methods should be proposed to quickly and accurately extract building information and estimate floor area in rural environments.

To ameliorate adverse social problems, building density regulations (such as those for building heights or floor area ratios) are common practices in urban planning and management worldwide [6]. Various remote sensing products and classification methods have been used to extract building coverage areas [7,8] and building heights [9,10]; the nDSM [11] (the difference between a DSM and a digital terrain models (DTM)) is widely used in height estimation [12]. Ji and Tang [13] proposed three methods for gross floor area estimation from monocular optical imagery using the NoS R-CNN model. Given the densely populated villages and scattered land-use layout in China, UAVs have become the latest trend in rural homestead detection because of their flexibility, low cost, real-time results, and high resolution [14]. Nyaruhuma et al. [15] used oblique photogrammetry to reconstruct 3D buildings on an urban scale. High-resolution UAV images can obtain sufficiently detailed information and provide new challenges to existing methods of building extraction [16,17]. Previous studies have primarily focused on the extraction of architectural features based on machine learning, including maximum likelihood classification [18], support vector machines [19], and object-based classification methods [20]. However, machine learning algorithms based on feature extraction rely heavily on manual parameter setting and expert knowledge, which usually leads to poor generalization with different environmental backgrounds [21,22]. In rural areas with more complex surface compositions, the use of traditional algorithms for ground object classification can be improved further [23].

Owing to the complexity of image backgrounds and the semantic texture of buildings, automatic and high-precision building extraction from UAV images presents uncertainty [24,25]. Recently, scholars have employed deep learning technology to identify building contour information [26–29]. Long et al. proposed the FCN model for pixel-level semantic segmentation, which is the first end-to-end fully convolutional network that accepts any size input for image segmentation, and it has successfully led to a new wave of semantic segmentation tasks [30]. Subsequently, many variant FCN-based models have improved the feature expression capabilities to obtain better experimental results (such as SegNet [31], U-Net [32], and ERFNet [33]). Liu et al. [34] proposed a novel convolutional neural network combined encoder–decoder and spatial pyramid pooling module named USPP for building extraction from high-resolution remote sensing images. Konstantinidis et al. [35] proposed a modular CNN to improve the performance of building detectors by employing a histogram of oriented gradients and local binary patterns in a remote sensing dataset. Zhang [36] developed a method for estimating homestead areas based on UAV images and the U-Net algorithm. The results demonstrate that, in rural areas with complex surface compositions, the deep learning method can achieve fast, stable, and high-precision results. Liao et al. [37] proposed a boundary-preserved model that works by jointly learning the contours and structures of buildings. Experiments on the WHU, Aerial, and Massachusetts Building Datasets showed that the proposed model outperformed other state-of-the-art methods. Xiao et al. [38] proposed a shifted-window transformer-based encoding booster to capture the semantic information of large buildings in high-resolution remote sensing images. Li et al. [39] proposed a novel end-to-end network integrating lightweight spatial and channel attention modules to refine features adaptively for building

extraction tasks. Wei et al. [40] proposed a multi-branch network for the extraction of rural homesteads based on aerial images. Jing et al. [41] proposed an efficient memory module to enhance the learning ability of deep learning models in building extraction. Li et al. [42] proposed a global style and local matching contrastive learning model for image-level and pixel-level representation. However, most existing deep learning models focus on stacking complex architectures and parameter settings to improve accuracy, which also has disadvantages, such as requiring extensive calculations and slow iteration speed [43]. Moreover, in the extraction of comprehensive building information, high-resolution remote sensing images cannot directly identify the numbers of floors in homesteads. The use of remote sensing data with high spatiotemporal resolution to estimate the area of village-level homesteads at the pixel level still remains challenging. Comprehensive methods and models should be combined with building extraction and floor area estimation at the village level.

Here, we propose a comprehensive method for building extraction and floor area estimation of village-level homesteads by combining UAV oblique photogrammetry and deep learning technology. First, the footprint of buildings is identified using the novel EDSANet model proposed, which employs dual attention extraction and attention feature refinement to enhance the accuracy of building extraction. Then, the number of floors of each building is estimated using the nDSM generated from UAV remote sensing. The total floor area of the homestead is rapidly calculated by multiplying the floor area of each building by the number of floors. A case study was conducted in Helan village, Shaanxi province, China. The experiments demonstrate that the proposed method can achieve rapid and low-cost results in building extraction and floor area estimation in rural villages. To summarize, the main contributions of this paper are as follows:

(1) We propose a comprehensive method combining UAV oblique photogrammetry and deep learning technology for building extraction and floor area estimation of village-level homesteads. A novel EDSANet model is proposed to tackle the problem of complex surface feature scenes in remote sensing images and improve performance in building extraction;
(2) We designed a semantic encoding module by applying three down-sample stages (with atrous convolution) to enlarge the receptive field and a spatial information encoding module with only six layers and three stages using one eighth of the original input to enrich spatial details and improve the accuracy in building extraction;
(3) A dual attention module is proposed to extract useful information from the kernel and channel, respectively. To adjust the excessive convergence of building feature information after attention extraction, we propose an attention feature refinement module to further improve the extraction effect of the model for useful features by redefining the attention features, thereby improving the accuracy.

The remainder of this paper is organized as follows: Section 2 describes the study area and data. Section 3 presents deep learning methods for building extraction and the UAV oblique photogrammetry method for floor area estimation. Section 4 introduces the results of the building extraction and floor area estimation. The discussion and conclusions are presented in Sections 5 and 6, respectively.

2. Study Area and Data

2.1. Study Area

Weinan City is located in Shaanxi province, China, from 34°13′E to 35°52′E and 108°50′N to 110°38′N. According to the 2017 census, the total population of the city is approximately 5.38 million, and it has an area of 13,134 km^2. Since 2018, Shaanxi province has vigorously promoted rural innovation and reform and accelerated the implementation of the rural revitalization strategy. The pilot reform project in Weinan city achieved remarkable success. Based on a field survey, this research selected Helan village, Fuping county, Weinan city, Shaanxi province as the research area. The village is located between the Guanzhong Plain and the northern Shaanxi Plateau. The village has an area of

3.88 × 10^4 m^2 with 205 households (of which 151 are residents) and a registered population of 321. The buildings are densely distributed in the research area, the village roads are planted with regular arbor forests, and parts of the homesteads are shaded by tall trees or shrubs. An overview of the study area is presented in Figure 1.

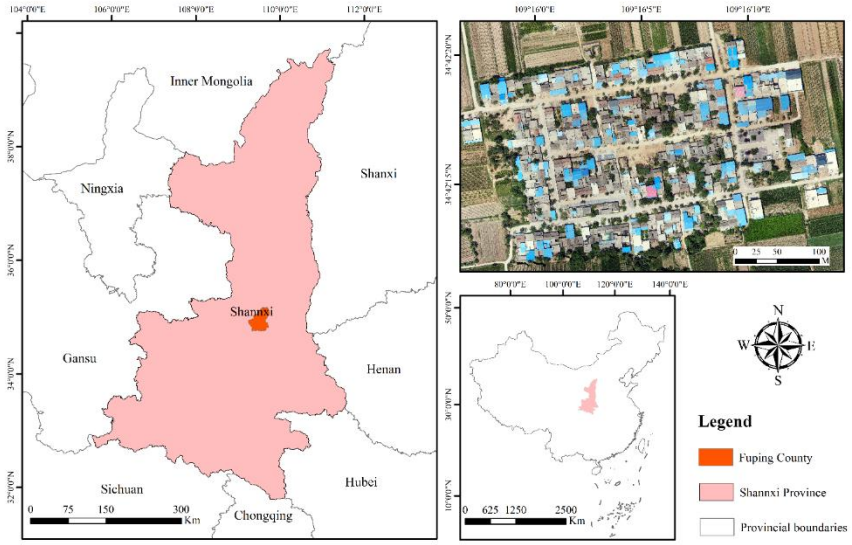


Figure 1. The geographical location of the study area.

2.2. UAV Data

The experimental data utilized in this research were tokens from a small four-rotor unmanned aerial vehicle (UAV). The drone model was an INSPIRE 2 (Shenzhen DJI Innovation Technology Co., Ltd., Shenzhen, China) equipped with a Zenmuse X5s HD camera, an effective pixel count of 16 million for the four thirds CMOS, and a built-in optical imaging lens camera composed of nine glass sheets in seven groups. The UAV was equipped with GPS and GLONASS dual satellite navigation systems, which can be used to autonomously plan the flight path in a study area. Table 1 presents detailed information on the UAV equipment.

Table 1. Detailed information on the UAV equipment.

Parameters	Value
Takeoff Weight	1280 g
Image Size	4608 × 3456
Flight Duration	27 min
Focal Length	15 mm
Ground Sample Distance	0.23 cm
Spectral Range	0.38–0.76 μm
Working Temperature	0–40°
Maximum Flight Altitude	6000 m
Maximum Horizontal Flight Speed	18 m/s
GPS Module	GPS/GLONASS dual mode
Image Coordinate System	WGS 84/UTM Zone 49N
UAV Flight Permission	Needed

A warm, clear, and windless day (2 August 2018) was chosen to ensure stability for the UAV photography. The flight track ranged from 108°50′E to 110°38′E and 34°13′N to 35°52′N (Figure 2). The flight route was from the southeast corner to the northwest corner of the study area, and pictures were taken along the S route. To construct photogrammetric stereo pairs, the two adjacent images were set with an 85% heading overlap and 75% inside overlap. The spatial resolution of the UAV data reached 2.3 cm.

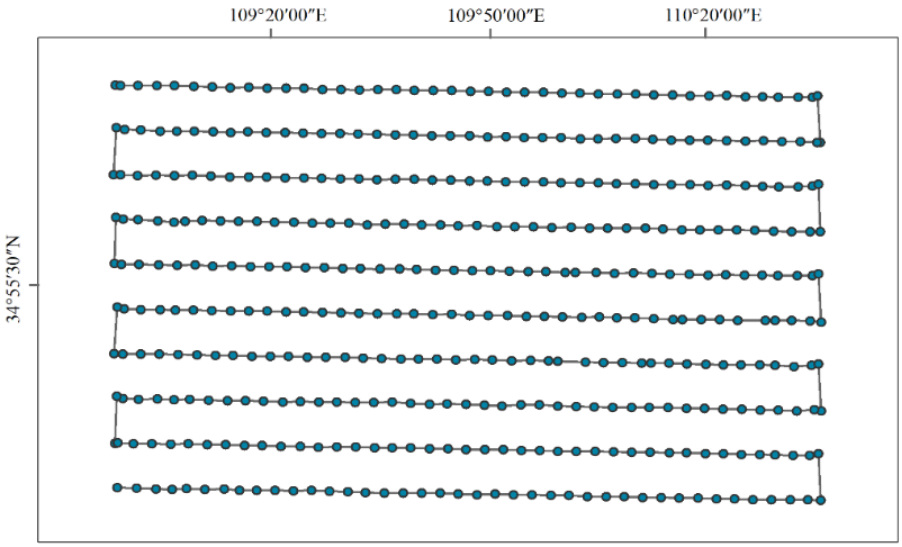


Figure 2. Flight route map in the research area.

3. Methodology

Figure 3 illustrates the detailed workflow, which includes seven principal steps. The first and second steps consisted of obtaining the orthophoto of the research area from the aerial UAV images. The orthophoto of the research area and the building sample dataset were produced through data preprocessing and augmentation. The proposed EDSANet model was then used to extract the building footprint of the study area, the accuracy was evaluated using five metrics, and the segmented images were merged into an entire image. Based on the UAV point cloud data, the tilt photogrammetry method was applied to generate the DSM, DTM, and nDSM to determine the building height. Lastly, the floor area of the homesteads in the study area was calculated based on the building footprints and the numbers of floors.

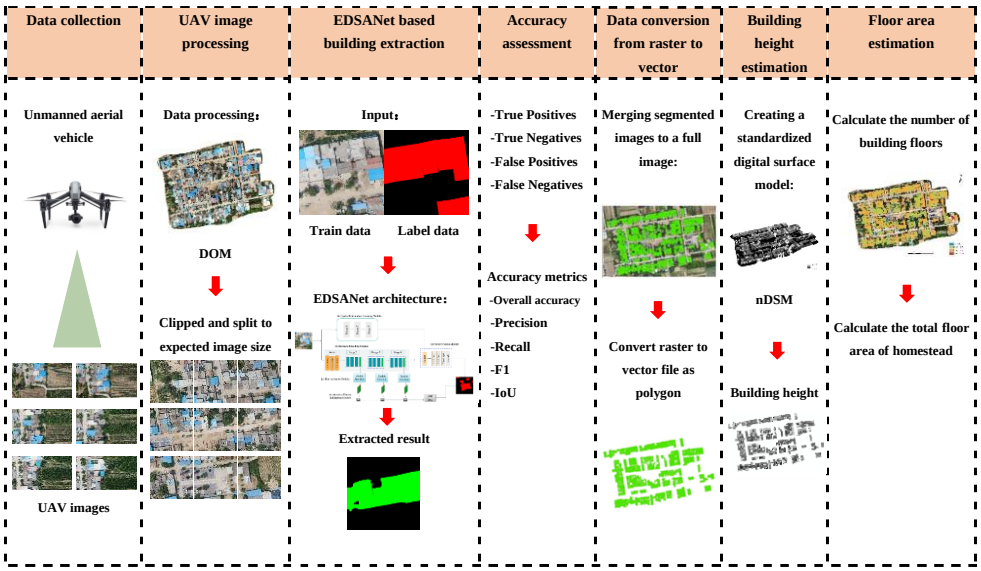


Figure 3. Flowchart of building extraction and floor area estimation in this research.

3.1. Methodology

3.1.1. EDSANet Architecture

We propose a novel fully connected network named the Efficient Deep-wise Spatial Attention Network (EDSANet) to tackle the problem of complex surface feature scenes in

remote sensing images and improve the efficiency and accuracy of building extraction tasks. Figure 4 shows an overview of the EDSANet architecture, including two branch networks composed of four units. (1) We first designed a semantic encoding module (SEM, Figure 4b), which employs channel splitting and shuffling to reduce computation and maintain higher segmentation accuracy. (2) A dual attention module (DAM, Figure 4d), consisting of spatial attention and channel attention, and an attention feature refinement module (AFRM, Figure 4e) were designed to make full use of the multi-level feature maps simultaneously, which helps predict the pixel-wise labels in each stage. (3) A spatial information encoding module (SIEM, Figure 4a) was used to enhance spatial semantic information and preserve spatial details. (4) We developed a simple feature fusion module (FFM, Figure 4c) to better aggregate the context information and spatial information [44].

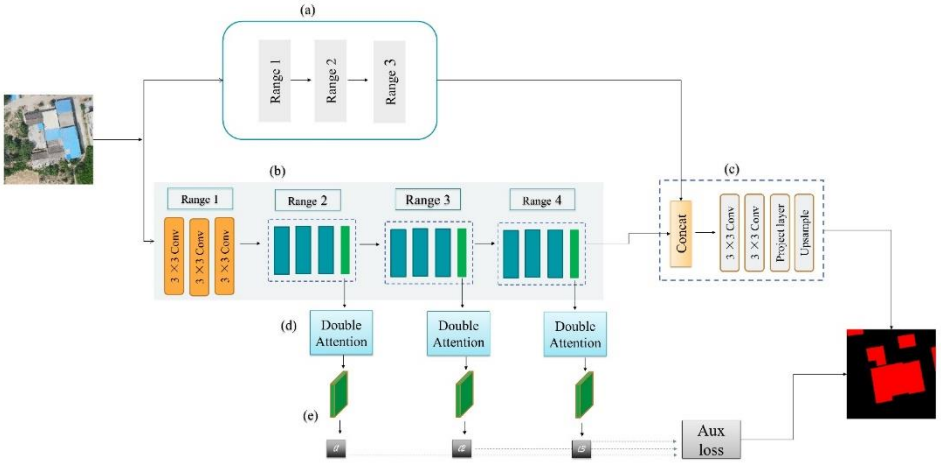


Figure 4. The architecture of the EDSANet model consists of two parts: the semantic encoding branch and the spatial information encoding branch. (**a**) Spatial information encoding module, (**b**) semantic encoding module, (**c**) feature fusion module, (**d**) dual attention module, and (**e**) attention feature refinement module.

First, input images are fed into the SEM to generate four feature maps ($F_{h,1}$, $F_{h,2}$, $F_{h,3}$, $F_{h,4}$) with decreasing spatial resolution. The feature maps $F_{h,3}$ and $F_{h,4}$ have the same numbers of channels, with different dilation rates, to enlarge the receptive field convolutional filters. Then, inspired by the efficiency of dilated convolution [45], we adopted a one-eighth down-sample strategy. As Equation (1) shows, the final segmentation $FFM_{h,s}$ is obtained by combining the high-resolution feature map F_h with the spatial feature map F_s from SIEM.

$$FFM_{h,s} = F_{up}(conv([F_h, F_s])), \quad (1)$$

3.1.2. Semantic Encoding Module (SEM)

This building block was designed with inspiration from lightweight image classification model strategies, such as in Ma et al. [46], Zhang et al. [47], and Sandler et al. [48]. The models mentioned above set the ratio of the input image resolution by applying five down-samplings and the size of the final output is only 1/32 of the input image size, which can lead to a significant loss in the spatial details. As Table 2 shows, our proposed SEM is based on this building block and applies three down-samplings (the output resolution is only one eighth of the original image resolution with 32, 64, and 128 channels). In stages three and four, atrous convolution is introduced to increase the receptive field.

3.1.3. Spatial Information Encoding Module (SIEM)

To improve the performance of semantic segmentation, the model aimed to effectively combine high-level semantics and low-level details. As the SEM was not designed for spatial details or low-level information, in the shallow SIEM, which has only six layers and three stages, each layer consists of a convolution operation (Conv), batch normalization

(BN), and a parametric rectified linear unit (PReLU) [49]. The first and second layers of each stage have the same number of filters (stride of 2) and output feature map size. Therefore, one eighth of the original input is extracted by the SIEM, which enriches the spatial details due to the high channel capacity.

Table 2. SEM is used to extract high-level semantic information.

Stage	Type	Filters
Input		
Stage 1	3 × 3 Conv	32
Stage 2	Down-sample	64
Stage 3	Down-sample	128
Stage 4	Building block	128

3.1.4. Dual Attention Module (DAM)

For the spatial dimension, we designed an attention mechanism based on kernel attention named the kernel attention module (KAM). For the channel dimension, the number of input channels C is normally far less than the number of pixels contained in the feature maps (i.e., $C \ll N$). Therefore, the complexity of the Softmax function for channels is not high. Thus, we utilized a channel attention mechanism based on the dot-product [50] named the channel attention module (CAM). As Figure 5 shows, using the KAM, which models the long-range dependencies of positions, and CAM, which models the long-range dependencies of channels, we designed the dual attention module (DAM) to enhance the discriminative ability of the feature maps extracted by each layer.

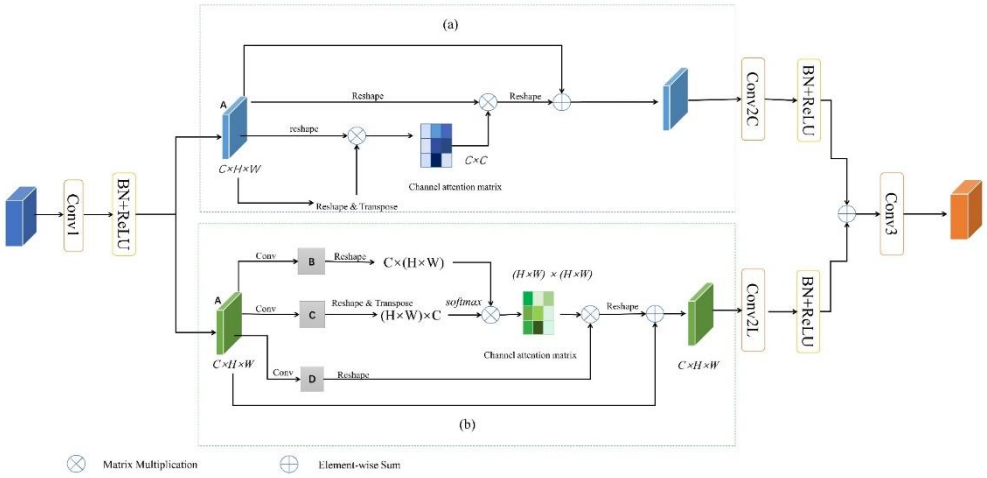


Figure 5. The architecture of the dual attention module consists of two branches: the kernel attention module and channel attention module. (**a**) Kernel attention module, and (**b**) channel attention module.

3.1.5. Deep Supervision

As providing supervision to the hidden layer reduces classification errors [21], researchers have adopted similar strategies [51] to ease the loss propagation in shallow layers. Therefore, we adopted auxiliary losses (Equation (2)) in stages two to four to supervise the predictions:

$$L_t = \alpha L_f \beta \sum_{i=1}^{n} L_i, \tag{2}$$

where α and β are the weights of the main loss function and auxiliary loss, with both weights set to 1; L_t is the total loss; L_f represents the loss for the output layer; and L_i represents the loss of the j-th stage after applying dual attention and feature refinement.

3.1.6. Loss Function

The loss function has an essential impact on the model accuracy and, usually, the most suitable loss function depends on the data properties and the class definitions [28]. Cross-entropy loss is a widely used loss function in two-dimensional semantic segmentation tasks. The aim of the learning-based remote sensing building extraction task is to train a binary classifier. The positive samples are pixels representing the buildings, whereas the negative samples are pixels containing the background. We here employed binary cross-entropy loss (Equation(3)) [52] in the training process:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)), \quad (3)$$

where y is the label (1 for green points and 0 for red points) and $p(y_i)$ is the predicted probability of the point being green for all N points.

3.2. Data Preprocessing

The data preprocessing in the deep learning technology primarily consists of image clipping and image labeling. Building segmentation is a binary classification task involving buildings and non-building elements [21]. The building samples were intended to contain various types of buildings in the study area. The building labels were manually completed in ArcGIS 10.2. The pixel values of each image were scaled to the interval [0,1] by dividing by 255. To facilitate the deep learning calculation, the original image was uniformly cropped to generate 256 × 256 pixels with an overlap of 56 pixels between two adjacent images.

Data augmentation is an effective way to enlarge a dataset and avoid overfitting [53]. As presented in Figure 6, the images were rotated by 90°, 180°, and 270°. Random horizontal and vertical flipping were performed with a probability of 0.5. After data augmentation, 4980 images with 256 × 256 pixels were generated. The spatial resolution of these images was about 2.3 to 5.3 cm. A total of 30% of the images were randomly selected as the test set, while the rest of the images were the training set. The final results of the building extraction were obtained by further applying a threshold of 0.5. No additional post-processing was performed in this study.

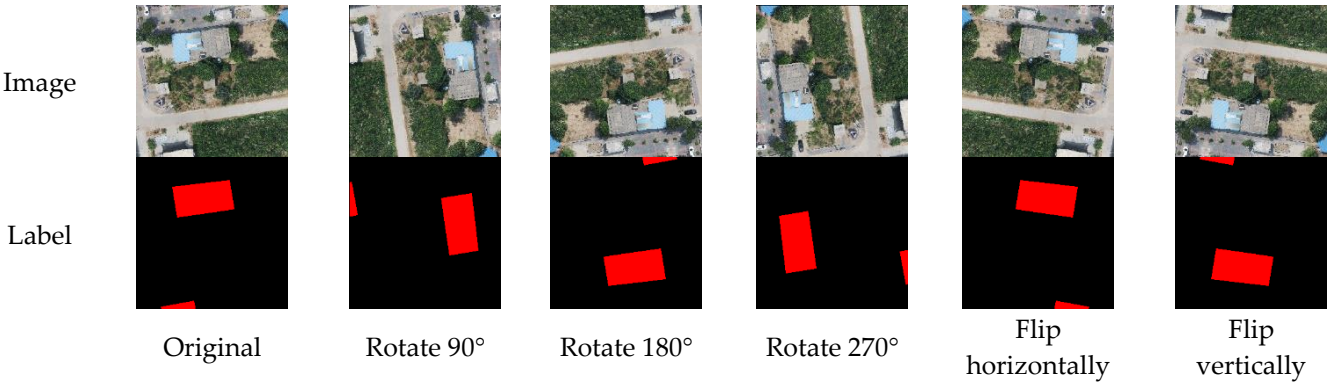


Figure 6. An example of data augmentation by rotating and flipping the rural Weinan building dataset.

3.3. Experimental Setting

The experiments were conducted using the PyTorch deep learning framework. All experiments were conducted on servers with 12th Gen Intel(R) Core™ i9-12900KF (3.20 GHz) and NVIDIA GeForce RTX 3090 (24 GB). All deep learning models were trained for 100 epochs, and 16 batches were randomly selected as the input data. The Adam optimizer was applied with an initial learning rate of 0.0001 and a weight decay of 0.0001. Figure 7 presents the dynamic changes in the accuracy and loss of the EDSANet model during the training process with the rural Weinan building dataset: the loss decreased and

the accuracy increased as the training epochs increased; after the number of epochs reached 60, the model training tended to stabilize, and the accuracy remained high.

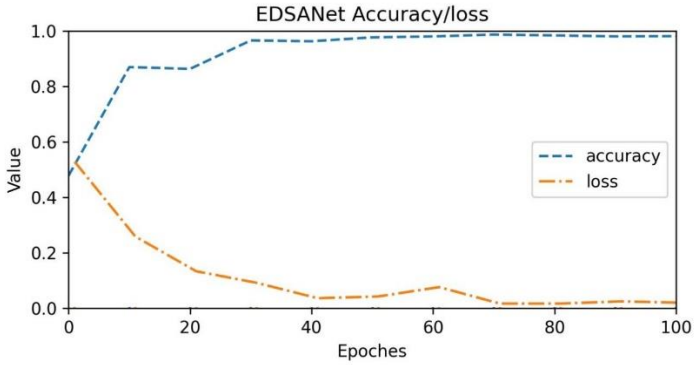


Figure 7. Changes in accuracy and loss for the EDSANet model in the training process.

3.4. Evaluation Metrics

Five common evaluation metrics were employed for quality evaluation in this research: overall accuracy (OA) (Equation (4)), precision (Equation (5)), recall (Equation (6)), F1-score (F1) (Equation (7)), and intersection-over-union (IoU) (Equation (8)). The five metrics are calculated as follows:

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{7}$$

$$IoU = \frac{TP}{TP + FP + FN}, \tag{8}$$

where P is the number of positive samples, N is the number of negative samples, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

3.5. Building Height and Floor Area Estimation

Figure 8 shows the workflow for the building height and floor area estimation. The UAV images obtained from field surveys were first fed into the Pix4Dmapper software (version 4.5.6) [54]. This software contains three image processing steps: initial processing, point cloud and mesh, and DSM orthomosaic [55]. The DSM can be extracted from overlapping aerial images obtained with photogrammetry technology using location information stored in the header file of each aerial image from the UAV flight [56]. Based on the DSM, the point-cloud filtering algorithm with mathematical morphology was employed to identify whether the filter window was the ground point. Subsequently, the ground objects on the surface (including buildings, trees, and other non-ground points) were eliminated. The DTM data, which represent the terrain elevation information, were then formed [57–59]. The difference between the DSM and the DTM is referred to as the nDSM, which is widely used in height estimation. The height of the rural elements above the terrain was then generated [60]. Based on field surveys of the usual heights of local buildings, a threshold was set to estimate the number of floors in rural buildings. Using the footprint area of the buildings identified by the EDSANet model, the numbers of floors of each building in the nDSM were extracted and estimated in the ArcGIS environment (version 10.2). Lastly,

the total floor area of the homesteads in the study area was obtained by counting the construction areas of each floor. The formula is as follows:

$$Area_{floors} = \sum_{i=1}^{floors_{max}\Sigma} Area_{gird} \times N_i, \quad (9)$$

where $Area_{floors}$ is the total floor area of the homesteads, $Area_{grid}$ is the area of the grid resolution of nDSM, N is the number of floors, and i ranges from 1 to $floors_{max}$ for each building.

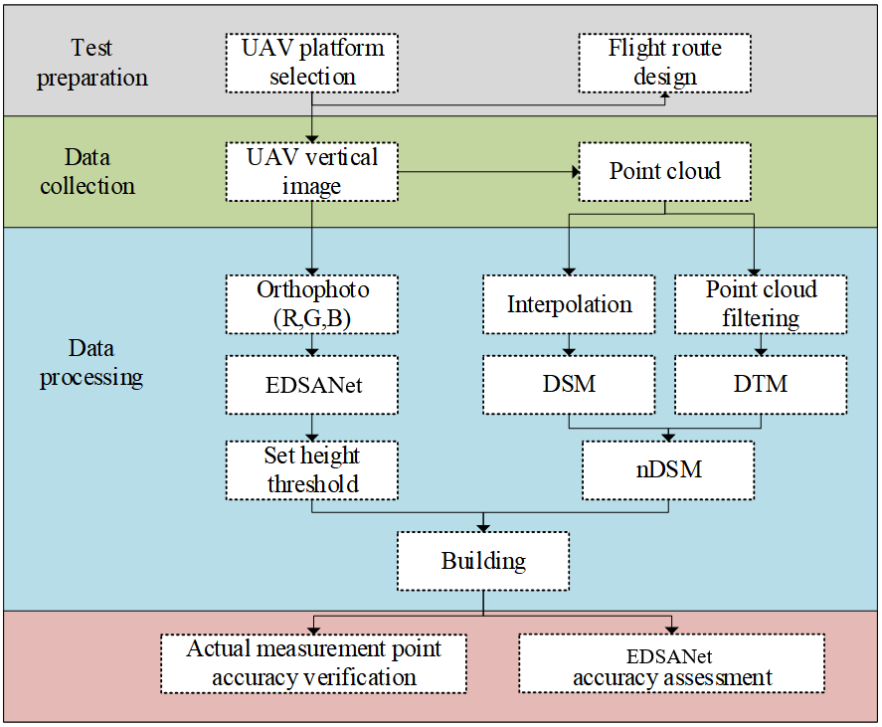


Figure 8. Flowchart of building height and floor area estimation.

4. Results

4.1. Building Extraction Using Deep Learning Models

Five classic and state-of-the-art deep learning models, including SegNet [31], UNet [30,32], Deeplabv3+ [61], MAP-Net [62], ARC-Net [23], and AGs-Unet [21], were compared to verify the performance and efficiency of the proposed EDSANet model with the rural Weinan building dataset. Figure 9 presents the qualitative results of building extraction using different deep learning models. SegNet returned too many false positives and false negatives exhibiting the worst performance with the dataset. Deeplabv3+, MAP-Net, and AGs-Unet presented quite similar performances in building extraction. For the proposed EDSANet model, the building segmentation results were satisfactory, and most buildings were generally well-segmented regardless of the type of roof (e.g., colored steel tile or sloped tile); additionally, the building footprints were very clear. However, the deep learning model could not clearly separate the boundaries between households in connected buildings (Figure 9a,b).

The specific analysis of the figure is as follows: Columns (a)–(d) represent four images randomly selected to show the test results. In (a) and (b), the proposed EDSANet model achieved effective completeness in extracted results for the whole single building. In the second column of buildings in (c), compared with Ags-Unet and ARC-Net, EDSANet clearly extracted the boundary of the buildings and showed the distinct gap between the buildings. Moreover, EDSANet was more advanced in the expression of the surrounding details of the building gap than Unet, MAP-Net, and Deeplabv3+, as shown in the lower

right corner of the image building extraction results in (d), but it was not as good as the boundary smoothness that the ARC-Net model achieved.

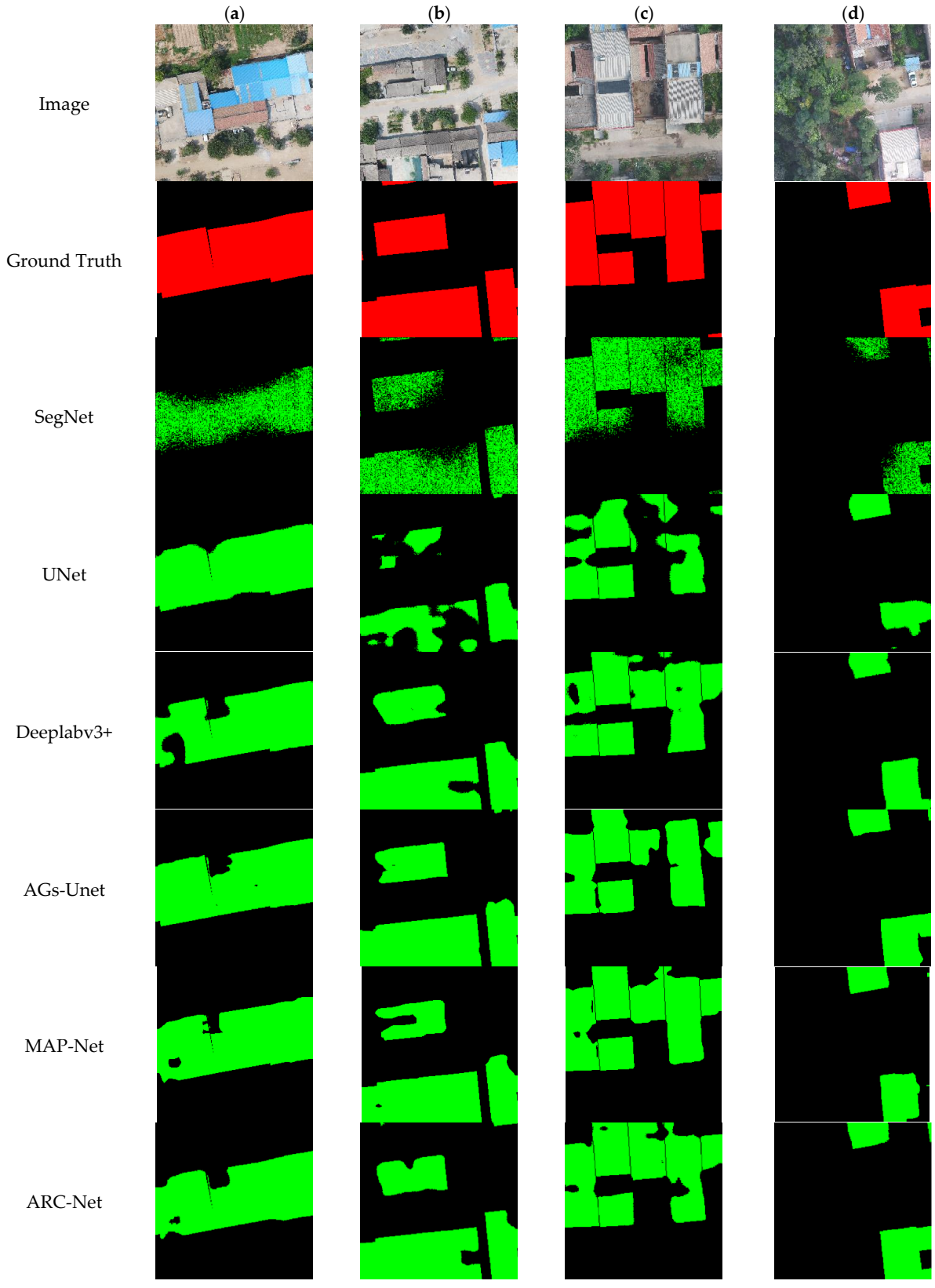


Figure 9. *Cont.*

EDSANet

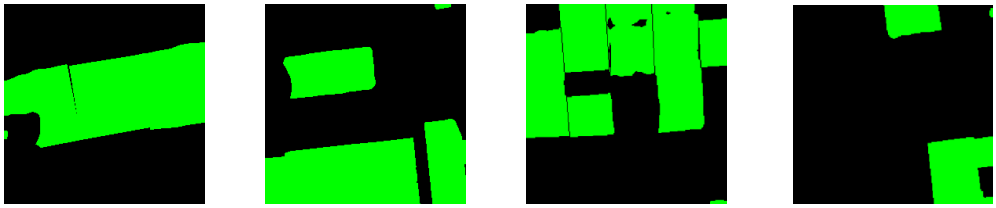


Figure 9. Building extraction results of different deep learning models with the rural Weinan building dataset. (**a**–**d**) Four images randomly selected to show the test results. SegNet, UNet, Deeplabv3+, Ags-Unet, MAP-Net, ARC-Net, and EDSANet, respectively, are represented by the building extraction results from the four groups of comparison experiments. Green represents the buildings and black represents the background. In the ground truth, red represents the buildings and black represents the background.

Table 3 presents the quantitative results of the building segmentation with the rural Weinan building dataset. SegNet obtained an overall accuracy of 0.740, while other models were all above 0.80. ARC-Net obtained an overall accuracy of 0.929 with a precision of 0.876, while EDSANet obtained an overall accuracy of 0.939 with an IoU of 0.848. In the experiments with the rural Weinan dataset, our proposed EDSANet model better balanced efficiency and accuracy compared to the MAP-Net and the ARC-Net models and achieved optimality for four evaluation metrics but not for recall, where Deeplabv3+ held the highest score of 0.946. Both the qualitative and quantitative experiment results demonstrate that EDSANet can effectively extract and fuse the features of rural buildings, improving the extraction accuracy for rural buildings. The results of the building extraction using the EDSANet model in Helan village are presented in Figure 10.

Table 3. Building extraction results with rural Weinan building dataset using different CNN models.

Models	OA	Precision	Recall	F1	IoU
SegNet	0.740	0.759	0.698	0.723	0.568
UNet	0.876	0.774	0.939	0.848	0.738
Deeplabv3+	0.899	0.813	**0.946**	0.872	0.777
AGs-Unet	0.907	0.864	0.911	0.887	0.798
MAP-Net	0.916	0.877	0.888	0.891	0.799
ARC-Net	0.929	0.876	0.921	0.902	0.822
EDSANet	**0.939**	**0.949**	0.887	**0.916**	**0.848** [1]

[1] Bold items in each column indicate the highest value.

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

Figure 10. Spatial distribution of rural buildings in Helan village. (**a**) Ground truth of homesteads and (**b**) identification results based on the EDSANet model.

4.2. Building Height Estimation

Figure 11a shows the DSM extracted from the overlapping aerial images using photogrammetry technology. The DTM, based on morphological filtering, was utilized to obtain the ground area in the DSM (Figure 11b). The pixel values of the nDSM represent

the height of the rural elements above the terrain (Figure 11c) and were calculated using the difference between the DSM and the DTM. The enclosed building area and the vegetation area on the ground cannot be correctly distinguished based only on the difference in height data. In the extraction of ground objects from high-resolution remote sensing data, the building segmentation precision obtained from the combination of spectral and height information is generally higher than that obtained using only spectral information or only height information.

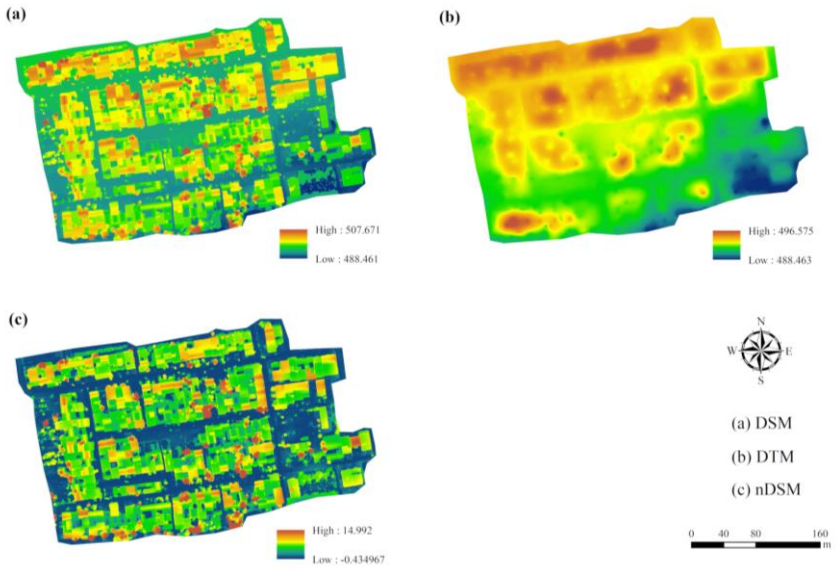


Figure 11. UAV-based estimation of the number of floors in rural buildings. (**a**) The DSM based on the photogrammetry workflow with the overlapping UAV images, (**b**) the DTM based on the point cloud filtering algorithm with the DSM images, and (**c**) the DTM subtracted from the DSM to create the nDSM.

The frequency distribution of the nDSM pixel values (Figure 12) was then calculated to obtain the building height information. Two pixel-value peaks were distributed near the 0.3 m and 4 m height differences. The pixel value of 0.3 m represents farmland crops and country roads, and the height difference of 4 m primarily represents the height of the roofs of one-story buildings or the walls of courtyards. All pixels in the nDSM grid with values less than 0.3 m were removed to avoid interference when extracting the building height. Moreover, because of the low reflectivity of the vegetation in the red band, the vegetation was well-extracted in the red band of the DOM image; the vegetation pixels in the nDSM were then detached with a raster operation in ArcGIS.

The building segmentation results of the deep learning method were a set of architectural and non-architectural images without a spatial reference. To facilitate the calculation and to display the results, the raster-based building footprint was converted into vector data after map projection in ArcGIS software to the coordinate system consistent with the reference image, which also made it possible to further calculate the floor area. Furthermore, the nDSM model of the interference pixels, including vegetation and roads, was removed with the mask of the homestead area. In accordance with the results of the field survey, the floors of the buildings in the study area were set at 4 m intervals. The height difference of 1–12 m was then set as indicating the first, second, and third floors (Table 4). In contrast, areas with floor heights of less than 1 m were set as indicating the courtyard height. Seventeen field survey buildings were randomly selected and utilized to examine the accuracy of the floor classification from the nDSM.

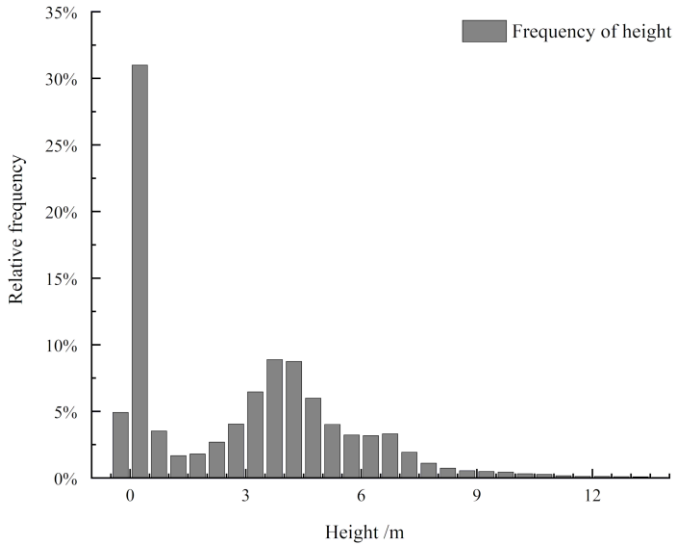


Figure 12. Frequency distribution diagram of the nDSM pixel values.

Table 4. Classification rules for the vegetation and the number of building floors.

Parameter	Threshold	Class
Brightness	≤60	Vegetation
Height	≤1 m	Courtyard
Height	1 m ≤ nDSM ≤ 4 m	One floor
Height	4 m ≤ nDSM ≤ 8 m	Two floors
Height	8 m ≤ nDSM ≤ 12 m	Three floors

Figure 13 displays the classification results for the building floors; 16 buildings were correctly classified and the height of 1 building was overestimated. The accuracy of the floor classification from the nDSM was 0.94, and the RMSE was 0.243 (Table 5). Field verification showed that the abnormal point was a canopy built by residents. The canopy was low and easily covered by the tall arbor canopy on one side. Therefore, the canopy height was calculated as the height of the vegetation canopy.

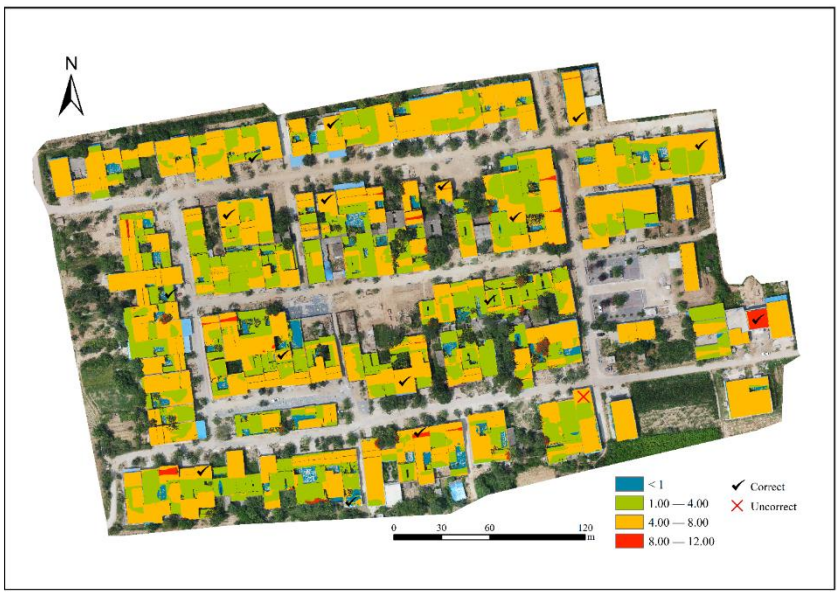


Figure 13. Classification results of building floors with nDSM.

Table 5. Confusion matrix for the number of floors divided by the nDSM model.

		Prediction			
		Courtyard	**Courtyard**	**Courtyard**	**Courtyard**
Actual	Courtyard	1	0	0	0
	One floor	0	3	0	0
	Two floors	0	1	11	0
	Three floors	0	0	0	1

4.3. Floor Area Estimation

The area of the homesteads was computed by multiplying the building area by the number of floors. We calculated the total construction area of the homesteads based on the reclassified results for the building heights from Section 4.2. The results showed that the total area of the homesteads was 3.1×10^5 m^2. Specifically, the homestead area for one-floor buildings was approximately 1.14×10^5 m^2 and accounted for 37.3% of the total homestead area; the homestead area for two-floor buildings with heights of 4–8 m was approximately 1.78×10^5 m^2 and accounted for 58.2% of the total construction area of the homesteads; three-floor buildings with a height difference of more than 8 m had a homestead area of approximately 3.33×10^3 m^2 and accounted for approximately 1.1% of the total construction area of the homesteads. The construction area for courtyards and low-rise shanty households accounted for only 3.4% of the total construction area. Figure 14 displays the frequency histogram for the building height; the average value of the pixels reached 4.45 m, and the standard deviation was 1.62. In conclusion, the average height and pixel frequency distribution indicated that residential buildings in the research area are primarily composed of two floors; this was consistent with the field survey results.

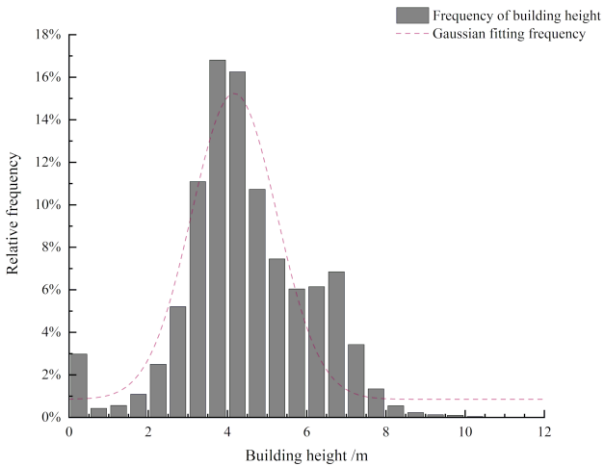
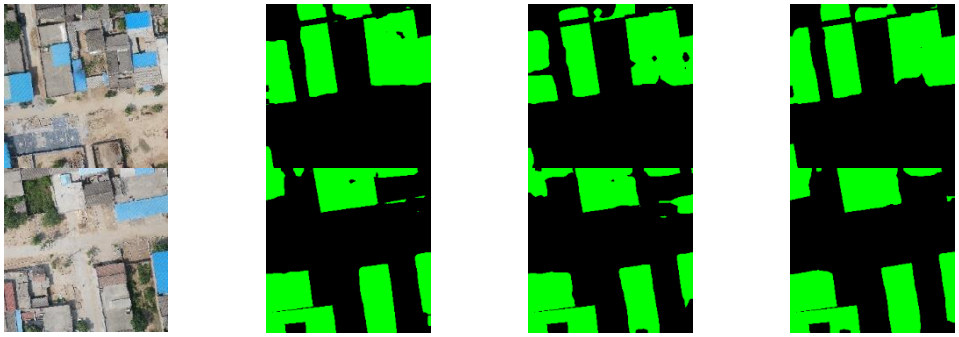


Figure 14. Frequency distribution diagram of building heights.

5. Discussion

5.1. Ablation Experiments

To further verify the feasibility of the DAM consisting of kernel attention and channel attention, the effectiveness of the atrous convolution in the SEM and the extraction precision of the different modules and fusion strategies were evaluated here in ablation experiments. The backbone model included the SIEM, SEM (without atrous convolution), and FFM. The benchmark models and strategies contained the backbone, the backbone + SEM (with atrous convolution), the backbone + DAM, the backbone + AFRM, and the backbone + SEM (with atrous convolution) + DAM + AFRM. Some of the ablation results for building extraction are presented in Figure 15. EDSANet (Figure 15b) showed the best performance in building extraction, with clear boundary identification compared to the model without DAM (Figure 15c) or AFRM (Figure 15d).



(**a**) (**b**) (**c**) (**d**)

Figure 15. Example of extracted results from ablation experiment with the Weinan building dataset. (**a**) The input images, (**b**) results extracted with the proposed EDSANet model, (**c**) EDSANet without DAM, and (**d**) EDSANet without AFRM.

The quantitative comparison results for the different combinations are shown in Table 6. Based on the reference network as the backbone, both the dual attention module consisting of the kernel and the channel attention module and attention feature refinement module improved the representation ability for the features extracted from the network. Compared with the backbone, the accuracy was significantly improved. However, the recall performance of the backbone, at 0.907, was better than that of the backbone + SEM (atrous convolution), the backbone + DAM, and the backbone + AFRM, individually. It can be concluded from Table 4 that adding the DAM or AFRM modules reduced the accuracy of the model based on the backbone, and all the OA, precision, recall, F1, and IoU results for the backbone + SEM (atrous convolution) + DAM + AFRM network, with AFRM, were improved. This indicates that AFRM can adjust the excessive convergence of building feature information after attention extraction with DAM, thereby improving the accuracy of building extraction in remote sensing. The last row in Table 6 shows the results for the proposed EDSANet model, which achieved the best performance in all evaluation metrics except for recall.

Table 6. Building extraction accuracy for modules and variants of the model.

Models	OA	Precision	Recall	F1	IoU
Backbone	0.911	0.862	**0.907**	0.883	0.783
Backbone + SEM (atrous convolution)	0.905	0.855	0.889	0.870	0.771
Backbone + DAM	0.906	0.847	0.899	0.870	0.773
Backbone + AFRM	0.914	0.878	0.882	0.879	0.787
Backbone + SEM (atrous convolution) + DAM + AFRM	**0.939**	**0.949**	0.887	**0.916**	**0.848** [1]

[1] Bold items in each column indicate the highest value.

5.2. Summaries and Limitations

Recent years have witnessed widespread application of deep learning in building extraction and other tasks owing to advancements in automatic learning features and strong adaptability. Previous studies have primarily focused on urban building extraction, which lacks application in rural China. In this study, we proposed the EDSANet model to extract buildings from UAV imagery in rural Weinan, China. The overall accuracy of the building extraction achieved by EDSANet was 0.929, and the precision was 0.876. Buildings were well-identified with clear boundaries regardless of the type of roof (e.g., colored steel tile or sloped tile). Buildings in rural areas mostly have one or two floors and are generally made of adobe, brickwood, and brick-concrete. The rural area selected in this research has fewer and more consistent building structure types than urban areas, which facilitates

building extraction. However, for some irregularly arranged rural areas, the performance of building extraction with the EDSANet should be further analyzed.

Consumer-grade drones are flexible and have high spatial resolution, which can ensure the clear boundaries of buildings and the accuracy of the three-dimensional point cloud model. Series of 3D products based on UAV flight data, including DSM and DTM, were here generated using oblique photogrammetry technology. The nDSM was used to remove the vacant rural plots, and the heights of the buildings were extracted from the nDSM model. However, classifying different types of ground objects with complex spectral information from high-resolution UAV images is difficult. In addition, 17 buildings field-surveyed on the ground, accounting for 8.3% of buildings and covering all numbers of floors in this village, were randomly selected and employed to verify the classification results of building heights in this study. As mentioned in Section 4.2, the height of the building at one sample point was overestimated because the roof was covered by the vegetation canopy. The overall accuracy of the classification results was 0.94. As the property rights and structures of the rural buildings were investigated and confirmed in the field survey, the building height error was primarily due to the instability in the drone flight conditions and the overestimation of the roof height caused by trees. In future studies, we will adopt mathematical morphological methods to eliminate interference factors and to further optimize the accuracy of the building boundaries identified by deep learning methods and elevation extraction using UAV oblique photogrammetry.

6. Conclusions

Rapid and accurate building extraction and floor area estimation at the village level are of great significance for the overall planning of rural development and intensive land use. In this study, we proposed a comprehensive method to estimate village-level homestead areas by combining UAV remote sensing and deep learning technology. First, the building footprints were identified using the proposed EDSANet model, which merged dual attention extraction and attention feature refinement to aggregate multi-level semantics and enhance the performance of building extraction, especially for high-spatial-resolution images. Then, the number of floors of each building was estimated using the nDSM model generated from UAV oblique photogrammetry. The floor area of the entire village was estimated by multiplying the floor area of each building by the number of floors in the village. The case study was conducted in Helan village, Shaanxi province, China. The results show that the overall accuracy of the building extraction with the EDSANet model from UAV images was 0.929, with the precision reaching 0.876. The buildings in Helan village are primarily composed of two stories and have a total floor area of 3.1×10^5 m^2. The field survey verified that the accuracy of the nDSM model was 0.94; the RMSE was 0.243. The experimental results demonstrate that the proposed workflow, combining UAV remote sensing and deep learning technology, can aid in rapid and efficient building extraction and floor area estimation at the village level in China, as well as worldwide.

Author Contributions: Conceptualization, J.Z. and Y.L.; methodology, J.Z.; software, X.C.; validation, X.Y.; formal analysis, X.Y.; investigation, X.Y.; resources, H.C.; data curation, H.C.; writing—original draft preparation, J.Z.; writing—review and editing, Y.L.; visualization, X.C. and L.G.; supervision, L.G.; project administration, Y.L. and G.N.; funding acquisition, Y.L. and G.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly supported by the National Natural Science Foundation of China, grant numbers 42201077 and 42177453; the Natural Science Foundation of Shandong Province, grant number ZR2021QD074; the Shandong Top Talent Special Foundation; and the National Nonprofit Fundamental Research Grant of China, Institute of Geology, China Earthquake Administration, grant number IGCEA2106.

Data Availability Statement: The codes are available at: https://github.com/Avery1991/2022EDSANet (accessed on 4 September 2022).

Acknowledgments: We would like to thank the editors and the anonymous reviewers for their insightful comments and suggestions.

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

References

1. Li, X.; Li, Z.; Yang, J.; Li, H.; Liu, Y.; Fu, B.; Yang, F. Seismic vulnerability comparison between rural Weinan and other rural areas in Western China. *Int. J. Disaster Risk Reduct.* **2020**, *48*, 101576. [CrossRef]
2. Liu, Y.; So, E.; Li, Z.; Su, G.; Gross, L.; Li, X.; Qi, W.; Yang, F.; Fu, B.; Yalikun, A.; et al. Scenario-based seismic vulnerability and hazard analyses to help direct disaster risk reduction in rural Weinan, China. *Int. J. Disaster Risk Reduct.* **2020**, *48*, 101577. [CrossRef]
3. Zhu, Q.; Li, Z.; Zhang, Y.; Guan, Q. Building Extraction from High Spatial Resolution Remote Sensing Images via Multiscale-Aware and Segmentation-Prior Conditional Random Fields. *Remote Sens.* **2020**, *12*, 3983. [CrossRef]
4. Liu, S.Y.; Xiong, X.F. Property rights and regulation: Evolution and reform of China's homestead system. *China Econ. Stud.* **2019**, *6*, 17–27.
5. Liu, Y.; Fang, F.; Li, Y. Key issues of land use in China and implications for policy making. *Land Use Policy* **2014**, *40*, 6–12. [CrossRef]
6. Yu, B.; Liu, H.; Wu, J.; Hu, Y.; Zhang, L. Automated derivation of urban building density information using airborne LiDAR data and object-based method. *Landsc. Urban Plan.* **2010**, *98*, 210–219. [CrossRef]
7. Liu, Y.; Zheng, X.; Ai, G.; Zhang, Y.; Zuo, Y. Generating a High-Precision True Digital Orthophoto Map Based on UAV Images. *ISPRS Int. J. Geo Inf.* **2018**, *7*, 333. [CrossRef]
8. Allouche, M.K.; Moulin, B. Amalgamation in cartographic generalization using Kohonen's feature nets. *Int. J. Geogr. Inf. Sci.* **2005**, *19*, 899–914. [CrossRef]
9. Dandabathula, G.; Sitiraju, S.R.; Jha, C.S. Retrieval of building heights from ICESat-2 photon data and evaluation with field measurements. *Environ. Res. Infrastruct. Sustain.* **2021**, *1*, 011003. [CrossRef]
10. Kamath, H.G.; Singh, M.; Magruder, L.A.; Yang, Z.-L.; Niyogi, D.J. GLOBUS: GLObal Building heights for Urban Studies. *arXiv* **2022**, arXiv:2205.12224.
11. Weidner, U.; Förstner, W. Towards automatic building extraction from high-resolution digital elevation models. *ISPRS J. Photogramm. Remote Sens.* **1995**, *50*, 38–49. [CrossRef]
12. Sefercik, U.G.; Karakis, S.; Bayik, C.; Alkan, M.; Yastikli, N. Contribution of Normalized DSM to Automatic Building Extraction from HR Mono Optical Satellite Imagery. *Eur. J. Remote Sens.* **2014**, *47*, 575–591. [CrossRef]
13. Ji, C.; Tang, H. Gross Floor Area Estimation from Monocular Optical Image Using the NoS R-CNN. *Remote Sens.* **2022**, *14*, 1567. [CrossRef]
14. Toth, C.; Jozkow, G. Remote sensing platforms and sensors: A survey. *Isprs J. Photogramm. Remote Sens.* **2016**, *115*, 22–36. [CrossRef]
15. Colomina, I.; Molina, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 79–97. [CrossRef]
16. Wang, J.Z.; Lin, Z.J.; Li, C.M.; Hong, Z.G. 3D Reconstruction of Buildings with Single UAV Image. *Remote Sens. Inf.* **2004**, *4*, 11–15.
17. Ma, Y.; Wu, H.; Wang, L.; Huang, B.; Ranjan, R.; Zomaya, A.; Jie, W. Remote sensing big data computing: Challenges and opportunities. *Futur. Gener. Comput. Syst.* **2015**, *51*, 47–60. [CrossRef]
18. Zhong, Y.; Ma, A.; Ong, Y.S.; Zhu, Z.; Zhang, L. Computational intelligence in optical remote sensing image processing. *Appl. Soft Comput.* **2018**, *64*, 75–93. [CrossRef]
19. Meng, Y.; Peng, S. Object-Oriented Building Extraction from High-Resolution Imagery Based on Fuzzy SVM. In Proceedings of the 2009 International Conference on Information Engineering and Computer Science, Wuhan, China, 19–20 December 2009.
20. Dahiya, S.; Garg, P.K.; Jat, M.K. Object Oriented Approach for Building Extraction from High Resolution Satellite Images. In Proceedings of the 2013 3rd IEEE International Advance Computing Conference (IACC), Ghaziabad, India, 22–23 February 2013.
21. Yu, M.; Chen, X.; Zhang, W.; Liu, Y. AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network. *Sensors* **2022**, *22*, 2932. [CrossRef]
22. Liu, Y.; Zhang, W.; Chen, X.; Yu, M.; Sun, Y.; Meng, F.; Fan, X. Landslide Detection of High-Resolution Satellite Images Using Asymmetric Dual-Channel Network. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 4091–4094. [CrossRef]
23. Liu, Y.; Zhou, J.; Qi, W.; Li, X.; Gross, L.; Shao, Q.; Zhao, Z.; Ni, L.; Fan, X.; Li, Z. ARC-Net: An Efficient Network for Building Extraction From High-Resolution Aerial Images. *IEEE Access* **2020**, *8*, 154997–155010. [CrossRef]
24. Boonpook, W.; Tan, Y.; Xu, B. Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *Int. J. Remote Sens.* **2020**, *42*, 1–19. [CrossRef]
25. Trevisiol, F.; Lambertini, A.; Franci, F.; Mandanici, E. An Object-Oriented Approach to the Classification of Roofing Materials Using Very High-Resolution Satellite Stereo-Pairs. *Remote Sens.* **2022**, *14*, 849. [CrossRef]
26. Yuan, J. Learning Building Extraction in Aerial Scenes with Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2793–2798. [CrossRef] [PubMed]

27. Vakalopoulou, M.; Karantzalos, K.; Komodakis, N.; Paragios, N. Building Detection in Very High Resolution Multispectral Data with Deep Learning Features. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015.
28. Touzani, S.; Granderson, J. Open Data and Deep Semantic Segmentation for Automated Extraction of Building Footprints. *Remote Sens.* **2021**, *13*, 2578. [CrossRef]
29. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1194–1206. [CrossRef]
30. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
31. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
32. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
33. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272. [CrossRef]
34. Liu, Y.; Gross, L.; Li, Z.; Li, X.; Fan, X.; Qi, W. Automatic Building Extraction on High-Resolution Remote Sensing Imagery Using Deep Convolutional Encoder-Decoder With Spatial Pyramid Pooling. *IEEE Access* **2019**, *7*, 128774–128786. [CrossRef]
35. Konstantinidis, D.; Argyriou, V.; Stathaki, T.; Grammalidis, N. A modular CNN-based building detector for remote sensing images. *Comput. Netw.* **2020**, *168*, 107034. [CrossRef]
36. Zhang, X. Village-Level Homestead and Building Floor Area Estimates Based on UAV Imagery and U-Net Algorithm. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 403. [CrossRef]
37. Liao, C.; Hu, H.; Li, H.; Ge, X.; Chen, M.; Li, C.; Zhu, Q. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049. [CrossRef]
38. Xiao, X.; Guo, W.; Chen, R.; Hui, Y.; Wang, J.; Zhao, H. A Swin Transformer-Based Encoding Booster Integrated in U-Shaped Network for Building Extraction. *Remote Sens.* **2022**, *14*, 2611. [CrossRef]
39. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic Segmentation Network With Spatial and Channel Attention Mechanism for High-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 905–909. [CrossRef]
40. Wei, R.; Fan, B.; Wang, Y.; Zhou, A.; Zhao, Z. MBNet: Multi-Branch Network for Extraction of Rural Homesteads Based on Aerial Images. *Remote Sens.* **2022**, *14*, 2443. [CrossRef]
41. Jing, W.; Lin, J.; Lu, H.; Chen, G.; Song, H. Learning holistic and discriminative features via an efficient external memory module for building extraction in remote sensing images. *Build. Environ.* **2022**, *222*, 109332. [CrossRef]
42. Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; Tao, C. Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618014. [CrossRef]
43. Lin, J.; Jing, W.; Song, H.; Chen, G. ESFNet: Efficient Network for Building Extraction from High-Resolution Aerial Images. *IEEE Access* **2019**, *7*, 54285–54294. [CrossRef]
44. Elhassan, M.A.; Huang, C.; Yang, C.; Munea, T.L. DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes. *Expert Syst. Appl.* **2021**, *183*, 115090. [CrossRef]
45. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
46. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical Guidelines for Efficient Cnn Architecture Design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
47. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
48. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification. In Proceedings of the International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 1026–1034.
50. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
51. Yu, M.; Zhang, W.; Chen, X.; Liu, Y.; Niu, J. An End-to-End Atrous Spatial Pyramid Pooling and Skip-Connections Generative Adversarial Segmentation Network for Building Extraction from High-Resolution Aerial Images. *Appl. Sci.* **2022**, *12*, 5151. [CrossRef]
52. De Boer, P.-T.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A Tutorial on the Cross-Entropy Method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]

53. Zhang, Z.; Wang, Y. JointNet: A Common Neural Network for Road and Building Extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
54. Krause, S.; Sanders, T.G.M.; Mund, J.-P.; Greve, K. UAV-Based Photogrammetric Tree Height Measurement for Intensive Forest Monitoring. *Remote Sens.* **2019**, *11*, 758. [CrossRef]
55. Kameyama, S.; Sugiura, K. Effects of Differences in Structure from Motion Software on Image Processing of Unmanned Aerial Vehicle Photography and Estimation of Crown Area and Tree Height in Forests. *Remote Sens.* **2021**, *13*, 626. [CrossRef]
56. Karantzalos, K.; Koutsourakis, P.; Kalisperakis, I.; Grammatikopoulos, L. Model-based building detection from low-cost optical sensors onboard unmanned aerial vehicles. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *XL-1/W4*, 293–297. [CrossRef]
57. Gevaert, C.; Persello, C.; Nex, F.; Vosselman, G. A deep learning approach to DTM extraction from imagery using rule-based training labels. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 106–123. [CrossRef]
58. Özcan, A.H.; Ünsalan, C.; Reinartz, P. Ground filtering and DTM generation from DSM data using probabilistic voting and segmentation. *Int. J. Remote Sens.* **2018**, *39*, 2860–2883. [CrossRef]
59. Serifoglu Yilmaz, C.; Gungor, O. Comparison of the performances of ground filtering algorithms and DTM generation from a UAV-based point cloud. *Geocarto Int.* **2018**, *33*, 522–537. [CrossRef]
60. Shukla, A.; Jain, K. Automatic extraction of urban land information from unmanned aerial vehicle (UAV) data. *Earth Sci. Inform.* **2020**, *13*, 1225–1236. [CrossRef]
61. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848.
62. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction from Remote Sensed Imagery. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *59*, 6169–6181. [CrossRef]