

Received June 14, 2021, accepted July 5, 2021, date of publication July 15, 2021, date of current version July 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3097630

HA U-Net: Improved Model for Building Extraction From High Resolution Remote Sensing Imagery

LEILEI XU¹, YUJUN LIU^{2,3}, PENG YANG^{4,5}, HAO CHEN⁶, HANYUE ZHANG⁷, DAN WANG^{3,8}, AND XIN ZHANG⁸

¹School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China

²Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

³Provincial Geomatics Center of Jiangsu, Nanjing 210013, China

⁴Qilu Research Institute, Aerospace Information Research Institute, Chinese Academy of Sciences, Jinan 250100, China

⁵Suzhou Zhe Xin Information Technology Company Ltd., Suzhou 215000, China

⁶Institute of Geodesy and Geoinformation Science, Technische Universität Berlin, 10553 Berlin, Germany

⁷Precision Forestry Key Laboratory of Beijing, Beijing Forestry University, Beijing 100083, China

⁸College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China


Corresponding authors: Yujun Liu (liuyj.20b@igsrr.ac.cn) and Hao Chen (1145871257@qq.com)

ABSTRACT Automatic extraction of buildings from high-resolution remote sensing images becomes an important research. Since the convolutional neural network can perform pixel-level segmentation, this technology has been applied in this field. But the increase in resolution prone to blurry segmentation because the model needs more edge detail and multi-scale detail learning. To solve this problem, a method is proposed in this paper, which consists of three parts: (1) an improved model named Holistically-Nested Attention U-Net (HA U-Net) is designed, which integrates the attention mechanism and multi-scale nested modules to supervise prediction; (2) During model training, an improved weighted loss function is proposed to make the designed model more focused on learning boundary features; (3) watershed algorithm is exploited for image post-processing to optimize segmentation results. The designed HA U-Net performs well on WHU Building Dataset and Urban3d Challenge dataset, and achieves 9.31%, 2.17% better F1-score and 10.78%, 1.77% better IOU than the standard U-Net respectively. The experimental results indicate that the proposed method can well solve the building adhesion problem. The research can serve as updating geographic databases.

INDEX TERMS Deep learning, building extraction, holistically-nested neural network, attention mechanism, weight mapping, watershed algorithm.

I. INTRODUCTION

The widespread of high-resolution remote sensing images makes it possible to accurately identify and locate artificial buildings from images. Such relevant research can provide basic database for related tasks such as old city reconstruction, urban planning, population estimation, and topographic map update [1]–[5]. However, targets usually vary greatly in scale, and many small buildings are displayed in the dense form in remote sensing images. This problem becomes more serious as the resolution of the image increases. This poses a huge challenge for the accurate and instantiated extraction of small buildings, especially for many areas with complex backgrounds [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang .

High-resolution images can intuitively reflect the rich texture structure and spatial semantic relation of the surface, which creates unique conditions for the application of CNN with powerful automatic feature extraction capabilities [7]–[11], in the field of automatic building extraction [12]–[14]. Among them, Mnih [15] used the deep neural network based on RBM network to extract buildings and roads in aerial imagery. Alshehhi *et al.* [10] replaced original fully connected layer in Mnih's model with a global average layer. Huang *et al.* proposed building extraction method based on fully convolutional neural networks. [7] Their research work eliminated the discontinuities caused by blocky areas and improved the predicted accuracy of building segmentation results. Due to the low output resolution of most segmentation models such as FCNs [16], DeepLab [17] and SegNet [18], the detailed information of

the building can be lost during downsampling. To solve this problem, Yang *et al.* [19] proposed a dense attention network called Dan, which integrated the spatial attention module to strengthen the learning of advanced features. In view of the scale diversity of buildings in images, Sun *et al.* proposed MCNN to extract multi-scale features and the features were input to different SVMs for classification [20]. Masouleh *et al.* introduced a new encoder-decoder dilated CNN for multi-scale building segmentation which includes multi-size dilated convolutional layers and modified skip connections to increase the level of abstraction abilities for multi-scale segmentation tasks [21].

Considering that there are many small and dense objects in remote sensing buildings, Hamaguchi *et al.* [22] used the LFE module to reduce the expansion coefficient and local characteristics. The dense circular convolution block and the non-porous convolution layer proposed by Zhang and Wang [23] balanced the relationship between the large receptive field and the small receptive field, it achieved good results in both large and small target extraction. According to the characteristics of remote sensing image, different models suitable for respective segmentation tasks are designed. Regarding the multi-source characteristic of remote sensing image, Pan *et al.* [24] combined lidar data with optical remote sensing data as input to train deep convolutional neural network. Chen *et al.* [25] and Xu *et al.* [4] took ResNet [26] as the backbone network for feature extraction and improved the segmentation accuracy of target by fully convolutional neural networks. In addition, Lin *et al.* [27] proposed the ESFNet, and parameters of the model were reduced by 8 times, which greatly improved the performance of the model without affecting the predicting accuracy.

Regarding the existing researches on deep learning to extract buildings from high-resolution remote sensing images, most of the researches rely on the full convolutional neural network architecture to make improvements and explorations according to specific problems [28], [29]. However, in essence, this type of semantic segmentation models to deal with extraction problems classify pixels into buildings and non-buildings, instead of emphasizing the distinction between individual buildings [30]. In addition, the buildings in remote sensing images have diverse scales, and the buildings are much smaller in many areas. They are mostly arranged in a compact manner with blurred boundaries, which are not conducive to prediction. At this stage, most of the auto instance segmentation methods origin from multi-task models [5], [31]. The multi-task network is composed of three subnets: classification net, detection net, and segmentation net derived from the regional proposal network (PRN). To handle relatively simple building segmentation task, it seems too complicated. Motivated by these limitations, this paper focuses on the system's segmentation capabilities in handling small building and dense building areas. First of all, our method applies an efficient and simple holistically-nested network (HNN) [32], [33] in the U-Net model. Based on the

generated semantic middle-level clues, the HNN architecture can learn the interior and boundary information of the building especially in small size, which is conducive to improving the segmentation and prediction ability. Furthermore, a powerful attention mechanism module is exploited to efficiently integrate multi-scale path information. The final designed network is called HA U-Net in this paper. Meanwhile, to segment adjacent targets, the total number of pixels on the adjacent boundary is much smaller than that in the entire image, which causes great obstacles to segmentation. Inspired by distance transform based weight map [34]–[36], the improved weight map is applied to loss function to assign more weights to the boundaries of small building areas, so that the network can focus on the learning of these areas and strengthen the boundary segmentation of small buildings. Finally, the watershed post-processing method [37] is used to further improve the partition effect between them and optimize the fine adjustment.

The main contributions of this paper are as follows.

(1) HA U-Net is designed by combining U-Net with the holistically-nested network and attention mechanism. The holistically-nested network fuses multiple levels of features at the decoder side, and these features participate in the final classification. The attention mechanism makes the lateral output of each level of the holistically-nested network not only have the detailed information of this level, but also have the semantic information of a higher level. The effect of the network model improvement on building extraction is studied in this paper;

(2) Research on the improvement of weight mapping. In view of the building adhesion problem in building extraction, the background, building boundaries and internal histograms in remote sensing images are obviously different. This paper applies the weight mapping improvement method in model training so that the target boundary in the image is fully learned by model;

(3) Research on image post-processing method based on watershed. The building adhesion problem often appears in densely constructed areas. Also, the deep learning prediction usually output low probability value at the boundary of the building and high probability value inside the building. Considering these two problems, the watershed segmentation method based on internal and external labels is applied.

The rest of this paper is organized as follows. In Section 2, methods are introduced in detail, including the specific structure of the proposed network, the weight mapping method, and the image post-processing method. Introduction to dataset in the experiment, and the experimental details such as model parameters and experimental evaluation indicators are shown in Section 3. In Section 4, experiments are conducted for model improvement, and multiple sets of experiments where methods are combined separately are performed to investigate the optimal scheme. The conclusion is drawn in Section 5.

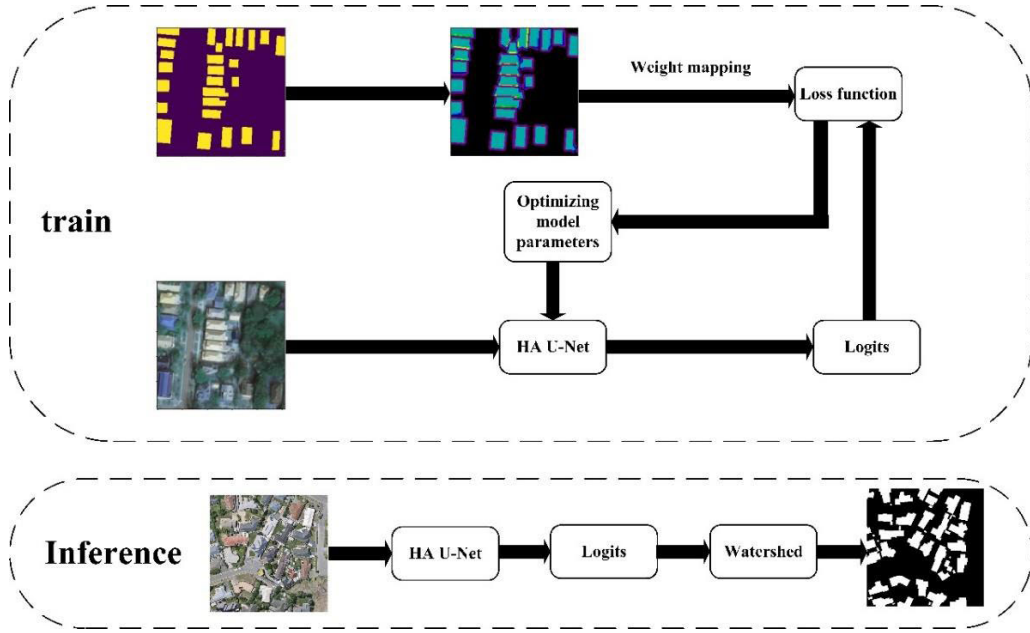


FIGURE 1. Method flowchart. The flowchart is divided into training stage and inference stage. In the training stage, images and corresponding labels are taken as input to train the model; weight mapping is applied to loss function for further optimization. In the inference stage, the image is input into the model, and the output probability map of the last layer of the model is subjected to watershed post-processing to obtain the final binarization map.

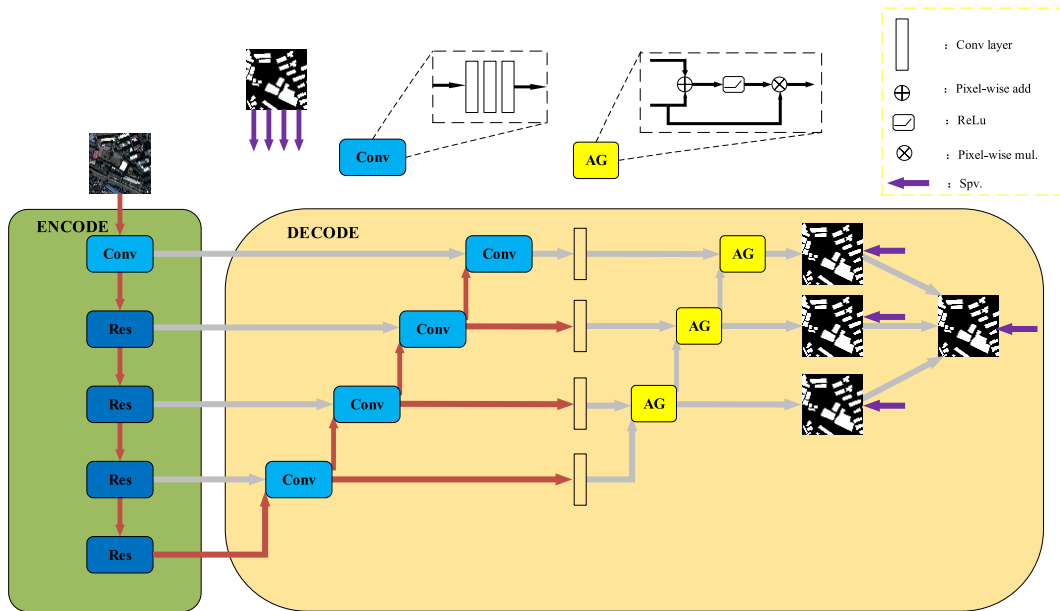


FIGURE 2. The pipeline of HA U-Net. Brown thick lines represent the building feature information flowing between the scales. The purple arrow indicates the back propagation of the loss function. AG: attention module ENCODE: Downsampling module. DECODE: Upsampling module.

II. METHODS

As shown in Fig. 1, the propose method is composed of three parts: HA U-Net based on U-Net which integrates HNN module and attention module, weight mapping applied to loss function in model training process and watershed post-processing in model predicting process. These three parts are described in session 2.1, 2.2, 2.3 respectively.

A. HA U-NET NETWORK ARCHITECTURE

This paper exploits the holistically-nested network and attention mechanism to improve the U-Net network structure, and designs an improved model: Holistically-Nested Attention U-Net (HA U-Net). Shown in Fig. 2 and Table 1, the structure of HA U-Net can be regarded as a combination of encoder and decoder. The encoder of this network adopts Resnet34 [26],

TABLE 1. HA U-Net network structure.

stage	name	Module type	Output size ^a
encode	Input module		64×256×256
	block1	Downsampling	64×128×128
		Residual block1	64×128×128
		Residual block2	128×64×64
		Residual block3	256×32×32
	Residual block4	512×16×16	
decode	block2 (Upsampling)	Convolution block1	256×32×32
		Convolution block2	128×64×64
		Convolution block3	64×128×128
		Convolution block4	32×256×256
	block3 (attention)	Attention block1	256×32×32
		Attention block2	128×64×64
		Attention block3	64×128×128
		Attention block4	32×256×256
	block4 (holistical nesting)	Feature fusion block	1×256×256
	block5 (Auxiliary loss)	Convolution block1	1×256×256
		Convolution block2	1×256×256
		Convolution block3	1×256×256

^aOutput size: take the input image size of 256×256 as an example for calculation, and its format is channel number × height × width.

which is composed of 4 residual blocks to extract feature. The fully connected layer of Resnet34 is replaced with a decoder structure. The decoder can be regarded as four modules: up-sampling module, attention module, overall nesting module, and auxiliary loss module.

The up-sampling module has the same structure as the decoder in the standard U-Net model. The up-sampling recover spatial location information of target and uses the bilinear difference method to restore to the original image size. Upsampled feature map of each layer is concatenated with the corresponding downsampled feature map of the encoder. The advantage of concatenation is that the semantic information of the target can be extracted, so that the model can make prediction at the pixel level. Since the encoder produces a total of four layers with different resolutions to propagate context information, the upsampling operation will also be performed four times.

In the original U-Net model, the output segmentation map can only be yielded when the feature map is upsampling to the top layer and merge with the corresponding feature map in the encoding layer. In multiple upsampling, the last one is selected as the output. However, the feature maps at different scales in several other upsampling process are not fully utilized, which is not completely beneficial to the extraction of targets in remote sensing images. Meanwhile, the features both inside and at the boundary of small-scale targets are easily lost in the upsampling process. Thus, the repeated use of the low-level feature information is fundamental to obtaining high-resolution and accurate segmentation results. The approach adopted in the proposed model is as follows

(1) The attention mechanism is used between two adjacent output feature maps in the upsampling process, and coarse-scale feature maps supervise the fine-scale feature maps. Then, the 1 × 1 convolution is performed to reduce

the channel number to obtain the lateral output of the corresponding layer.

(2) Inspired by the idea of HNN, several loss functions are calculated in the model's intermediate layers. Based on the incorporation of predictions from different network stages, different levels are nested to enhance the extraction ability of targets at multiple scales, especially small targets.

(3) To ensure that the fusion of the lateral output from each intermediate layer contributes best to the final probability map, appropriate fusion is adopted instead of fusing all different scales. On this basis, the model's capture of target edge information is supervised by lateral loss functions more effectively.

After testing and comparison, it is found that too much or too little nesting has adverse effect on the overall performance of the model. Finally, this paper uses sub-modules 2, 3, and 4 (from the bottom of the decode, sub-module is numbered sequentially starting from 1) as fusion of output feature maps, which is abbreviated as HNN234 for convenience.

An auxiliary loss module is added for model training. The specific location is shown by the purple arrow in Fig. 2. In the auxiliary loss module, the lateral outputs of HNN234 and the final fusion output are 1 × 1 convolved and scaled to the original image size. They are respectively calculated with the ground truth. The main loss function supervises the final output layer of the network model, and the auxiliary loss function is set in each lateral output layer to supervise the feature learning of other scales. The final loss function formula is as follows:

$$\text{FinalLoss} = \text{Loss} + \text{Lossside1} + \text{Lossside2} + \text{Lossside3} \quad (1)$$

B. LOSS FUNCTION COMBINED WITH WEIGHT MAP

The weight map for each ground truth segmentation is pre-computed to compensate different frequency of pixels

from classes in the training dataset and to force the network to learn the small separation borders in the training.

The house adhesion often appears in segmentation results. The biggest challenge comes from the segmentation of compact buildings. Moreover, the number of pixels in the interior, interstitial areas, and boundaries of adjacent buildings is much smaller than the total number of pixels, which increases the difficulty of segmentation and makes the segmentation of compact buildings most challenging. In addition, training the model with extremely unbalanced classes causes network optimization difficulties easily.

Therefore, the weighted cross entropy loss function [38] is used to strengthen model learning for contours of the building, which can be calculated by (2).

$$Loss = -W^{IWM} \sum_i y_i \log\left(\frac{e^{y'_i}}{\sum_i e^{y'_i}}\right) \quad (2)$$

where y_i represents the ground truth, and y'_i represents the predicted values. W^{IWM} is the proposed weight map, and it can be calculated by (3).

$$W^{IWM} = W^{DWM} * \alpha + (1 - \alpha) * W^{UWM} \quad (3)$$

The improved weight mapping (IWM) is a weighted combination of UWM and DWM. $\alpha \in (0, 1)$ is a control parameter, and it is found that $\alpha = 0.6$ contributes to better results on the Urban3D challenge dataset.

$W(p)^{UWM}$ [37] and $W(p)^{DWM}$ (p, β) [39] are two different weight mapping functions separately. $W(p)^{UWM}$ can be calculated by (4).

$$W(p)^{UWM} = W_c(p) + W_0 * \exp\left(-\frac{(d_1(p) + d_2(p))^2}{2\sigma^2}\right) \quad (4)$$

W_c is the weight map to balance the class frequencies; d_1 denotes the distance to the boundary of the nearest target, and d_2 denotes the distance to the boundary of the second nearest target. In our experiments w_0 and σ are set to 10 and 5, respectively. $W^{DWM}(p, \beta)$ can be calculated by (5).

$$W^{DWM}(p, \beta) = W_0(p) * \left(1 - \min\left(\frac{\vartheta_g(p)}{\beta}, 1\right)\right) \quad (5)$$

where ϑ_g represents the Euclidean distance of the closest non-background pixel assigned to the p pixel of the g category; $W_0(p)$ is the class imbalance weight, which is inversely proportional to the number of pixels in the class; β is a control parameter used to decay the contour weight.

The weight mapping of different distance conversion methods is shown in Figure 3. It can be seen that UWM is superior in dealing with the imbalance of categories, and compact targets occupy more weight, and vice versa. Unfortunately, even though UWM has strong ability to segment compact buildings, it does not perform well on the boundary problem of sparsely distributed building areas. In contrast, DWM can smooth the boundary as a whole, but does not have strong ability to identify closely adjacent buildings. A weighted combination of U-Net weight mapping (UWM) and distance

transform-based weight mapping (DWM) is made to enhance the discriminative ability of the network to obtain a more accurate segmentation of individual building.

C. IMAGE POST-PROCESSING BASED ON WATERSHED ALGORITHM

To further improve the edge segmentation effect of buildings, the probability segmentation map of the model output is processed by watershed post-processing operations. The result is as shown in Fig. 4. Considering the individual building from its geometric center, the probability value is generally distributed from high to low, which is pyramid-shaped. Meanwhile, the edge performance is not confident, providing conditions for the tag-based watershed algorithm.

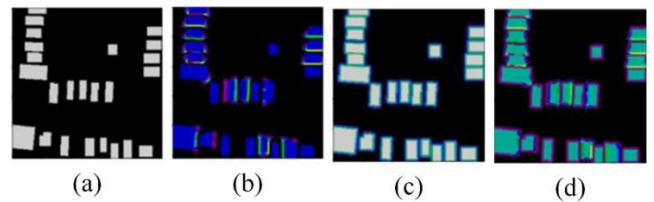


FIGURE 3. Weight mapping of different distance conversion methods. (a) Binarization label; (b) UWM weight label; (c) DWM weight label; (d) IWM weight label.

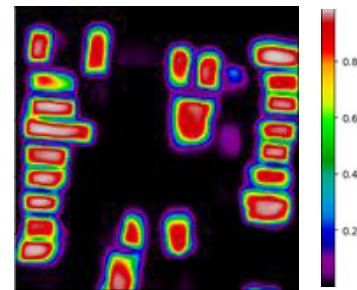


FIGURE 4. Heat map of probability distribution.

The basic idea of the watershed algorithm [40] is to imagine the ladder diagram as a topographic map, and simulate flooding or precipitation in reality. When the water level fluctuates, segmentation areas forms on the image surface, and the area boundary is the desired watershed boundary. The general watershed algorithm first obtains the gradient image, as shown in formula 5. Then, it uses the obtained gradient image as the input image, and finally performs corresponding processing. In this way, pixels with similar spatial positions and similar gray values are connected to each other to form a closed contour.

$$\begin{aligned} g(x, y) &= grad(f(x, y)) \\ &= \{[f(x, y) - f(x - 1, y)]^2 + [f(x, y) \\ &\quad - f(x, y - 1)]^2\}^{0.5} \end{aligned} \quad (6)$$

In (6), $f(x, y)$ is the model output, and $\text{grad}()$ is the gradient image; $g(x, y)$ is the output image processed by the gradient operator.

The general watershed algorithm is prone to over-segmentation. To overcome this defect, this paper performs a double threshold operation on the probability distribution map to obtain both internal and external tags. Specifically, the high threshold corresponds to the internal tag, and the low threshold corresponds to the external tag. Then, the tag-based watershed algorithm is exploited to process and retain the waterline, and finally superimpose the waterline on the prediction result to improve the edge segmentation of the buildings.

III. EXPERIMENT PREPARATION

A. DATASET

The Urban3d Challenge dataset [41] on Topcoder contains roughly 103,000 buildings at the scale in urban settings. There are many complex scenes in dense areas containing many small closely-spaced buildings, which is suitable to verify the feasibility of the method in this paper. The benchmark dataset has a spatial resolution of 0.5m, orthorectified RGB imagery, ground truth, digital surface models (DSM) as well as digital terrain models (DTM) are included in this dataset. Since the ultimate goal of this work is to semantically segment the target, the Class-Level Images are selected as the ground truth, which indicates whether each pixel belongs to the building class or not. The dataset is evenly divided into tiles with the area of 1 square kilometer, and 174 tiles with 2048×2048 pixels are obtained for the experiment. The obtained tiles are randomly divided into three subsets, i.e., training set (128 slices), validation set (32 slices), and testing set (14 slices).

The WHU Building Dataset [42] is a building dataset consisting of satellite imagery dataset and aerial imagery dataset. The subset of aerial imagery is selected for verification in the experiment. It covers 450 square kilometers of Christchurch in New Zealand and contains 18,7000 buildings. After resampling and cropping, the ready-to-use samples include non-overlapping tiles with 512×512 pixels and a spatial resolution of 0.3 m. The dataset is divided into a training set (4736 tiles, containing 130,500 buildings), a validation set (1036 tiles, containing 14,500 buildings) and a test set (2416 tiles, containing 42,000 buildings).

B. EXPERIMENT SETUP

1) TRAINING DETAILS

To make a full use of the dataset, DSM and DTM were also processed accordingly and used as the fourth band during the comparative experiment for Urban3d Challenge dataset. Difference calculations on these two models were done to obtain the normalized digital surface model (nDSM) for model training. Before the original image and the corresponding ground truth were input into the model for training, they were cropped to 256×256 pixel slices with 210 pixels as the step

length to improve the model training efficiency on the two datasets.

In addition, the training data was enhanced during the training process to improve the generalization ability of the model. The enhancement includes randomly missing pixels, sharpening images, random rotation, cropping edge pixels, and mirroring flips.

The proposed HA U-Net was implemented using Pytorch. All models were trained and tested in the Linux platform with a GeForce RTX 3090 (24 GB RAM).

During training, the network model was optimized with the improved algorithm based on Adam, i.e., RAdam [43]. The optimization algorithm RAdam with momentum accumulates the rate of historical gradient movement. When the gradient in a certain direction is too different from the previous one, it will weaken the current gradient. If the gradient in a certain direction is not much different from the previous one, it will increase this time. This makes the network converge faster. Also, the learning rate planning function ReduceLROnPlateau was exploited to update the parameters of the deep learning network model, so that the learning rate was scaled proportionally when the cumulative times exceed the tolerance times. Besides, the momentum value of the training was set to 0.9 and the batch size was set 16.

2) EVALUATION INDEX

To quantitatively evaluate the proposed method for building segmentation, the intersection over union (IOU), kappa coefficient, and instantiated F1-score (Ins F1) were used as the evaluation criteria.

In the segmentation task, IOU is expressed as the degree of coincidence between the truth value and the prediction value, i.e., the pixel-wise intersection and union between Ground Truth (GT) and the prediction (P). IOU can be calculated by the following formula:

$$\text{IOU} = \frac{GT \cap P}{GT \cup P} \quad (7)$$

The Kappa coefficient is a criterion used to test consistency, and it can also be used to evaluate the pixel-classification result. For classification problems, the so-called consistency is whether the actual classification results are consistent with the prediction results. The calculation of the Kappa coefficient is based on the confusion matrix, which is shown+ as follows:

$$\text{kappa} = \frac{P_0 - P_e}{1 - P_e} \quad (8)$$

where P_0 is the sum of the diagonal elements in the confusion matrix divided by the sum of the entire matrix elements, and it is equivalent to accuracy; P_e is the sum of the products of the actual and predicted pixel number corresponding to all categories divided by the square of the total pixel number.

Besides, the Ins F1 is exploited by this paper to further evaluate the instance segmentation ability of the network model. This criterion has been used as an evaluation metric in the Urban 3D challenge in 2018 and the Ali Tianchi Building

Intelligence Census Competition in 2020. The definition and calculation of the Ins F1 are described as follows:

(1) Take all connected components of the true value and the predicted result as the object, and find the component with the highest IOU among the true value components for each predicted component;

(2) Judge each component based on IOU. If IOU is greater than 0.50, the component is classified as TP, otherwise, FP;

(3) If the predicted component does not exist in the true value, it is classified as FP.

F1-score can be instantiated based on TP, FP, and FN results, its calculation method is as follow.

$$prediction = \frac{TP_{IOU>0.5}}{TP_{IOU>0.5} + FP_{IOU>0.5}} \quad (9)$$

$$recall = \frac{TP_{IOU>0.5}}{TP_{IOU>0.5} + FN_{IOU>0.5}} \quad (10)$$

$$F1 = \frac{prediction \times recall}{prediction + recall} \quad (11)$$

IV. RESULT AND DISCUSSION

A. IMPROVED MODEL COMPARATIVE EXPERIMENT

U-Net was used as the benchmark model for reference and comparison. To verify the effect of the HA U-Net model on Urban3d Challenge dataset, two sets of experiments were conducted: (1) Different levels of nesting on the decoder of U-Net were compared to obtain the best nesting scheme; (2) The effect of adding attention mechanism on further improvement of the model segmentation ability was determined. The experimental results were compared with those of only nested U-Net model, U-Net and attention U-Net. The final model HA U-Net is determined through two rounds of comparisons.

The results of experiment 1 are shown in Fig. 5. It can be seen that: (1) compared with the benchmark U-Net model, adding the multi-scale features to U-Net, HNN series greatly improves the segmentation results of multi-scale building areas. (2) Nesting the top three scales obtains the best segmentation results, indicating that appropriate nesting of output features at different scales is beneficial. However, it is not the best way to nest all the features of different scales in the upsampling module. The possible reason is that resolution of the underlying feature map is too low, and sufficient spatial information cannot be recovered, thus reducing the overall performance of the model when the feature map is resampled.

The results of experiment 2 are illustrated in Fig. 6. Compared with U-Net, Attention U-Net does not obtain significantly improved results. The adding of attention module to the module (U-Net+ HNN234) caused a reduced red area and an increased recall rate, indicating that the attention module can use coarse-scale features in nested module to activate regions of interest for higher-resolution features and inhibit the role of its irrelevant areas.

To further quantitatively evaluate the classification effect of the improved model, the three criteria are calculated, including IOU, Kappa coefficient and Ins F1. The results are

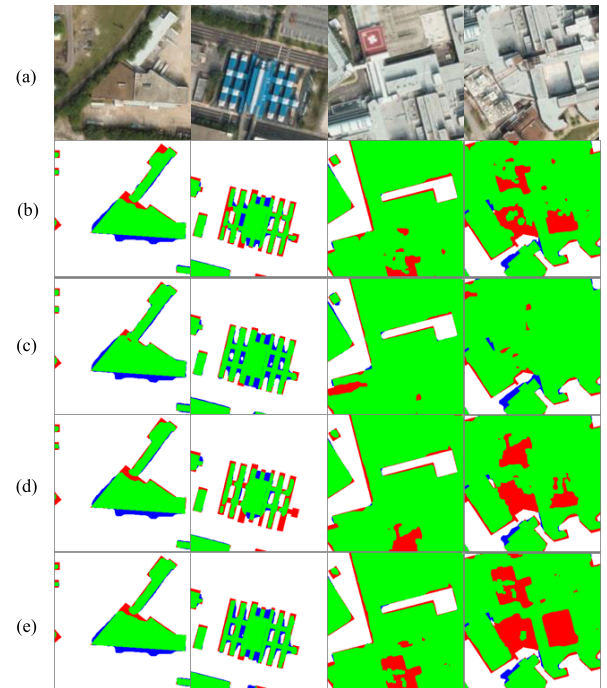


FIGURE 5. Results of the HNN series, (a) remote sensing image, (b) (c) (d) and (e) respectively represent prediction results of HNN34, HNN234, HNN1234, and U-Net where green color stands for TP, blue color stands for FP and "1234" represents four different levels of lateral output.

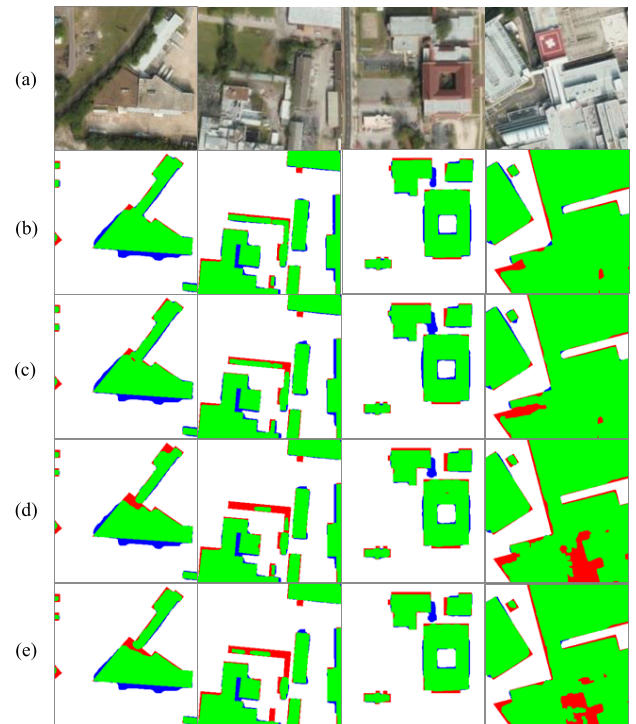


FIGURE 6. Results of U-Net with AG + HNN234 (HA U-Net) and other references (a) remote sensing image; (b), (c), (d) and (e) represent the results of the model HA U-Net, U-Net + HNN234, attention U-Net, and U-Net respectively.

listed in Table 2. It can be seen from the table that the HNN module can improve the extraction of multi-scale buildings. According to the results, HNN234 achieves the best result. Compared with those of U-Net, values of IOU, Kappa and

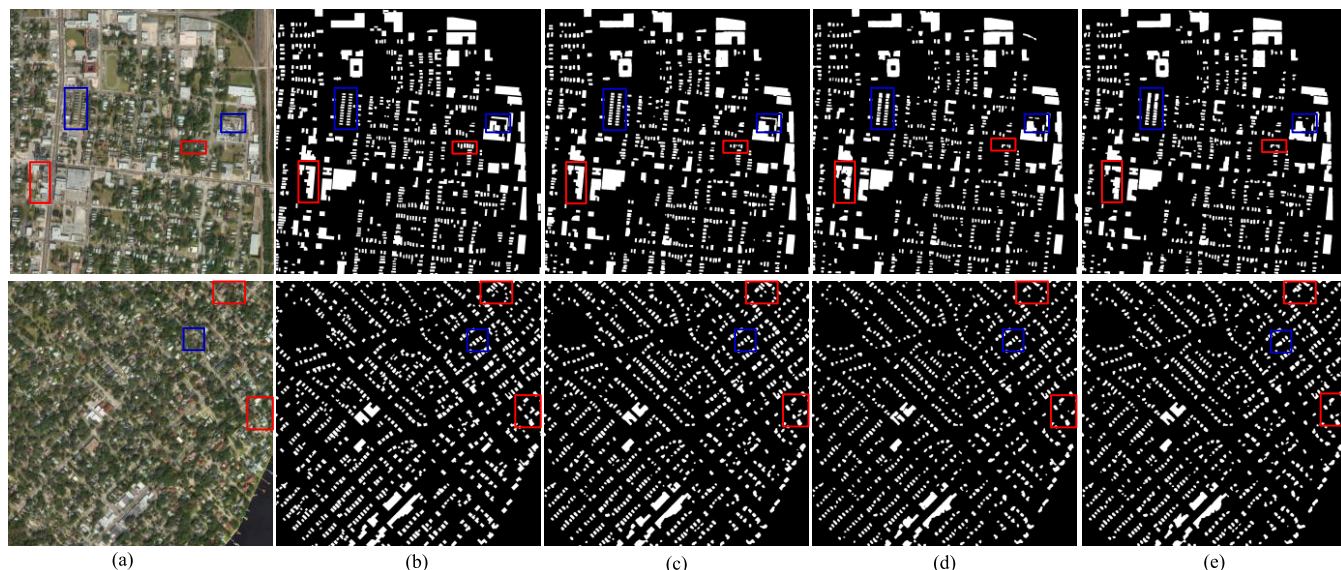


FIGURE 7. Segmentation results of different models, (a) original image, (b) GT, (c) HA U-Net, (d) HRNet, (e) D-LinkNet. The blue boxes illustrate the problem of building adhesion, and the red boxes illustrate the problem of building omission.

TABLE 2. Network experiment comparison (%).

Model\Metric	IOU	Kappa	Ins F1
U-Net+ HNN1234	73.61	82.47	85.73
U-Net+ HNN234	74.79	83.27	87.38
U-Net+ HNN34	73.91	82.70	86.80
U-Net	73.19	82.16	85.76
attention U-Net [44]	73.59	82.45	86.16
HA U-Net	74.96	83.43	87.93

Ins F1 increased by 1.60%, 1.11%, and 1.62%, respectively. The Ins F1 increases most, indicating that the combination of multi-scale features is beneficial to improving the model’s ability to identify individual buildings. Based on the results of experiment 1, HNN234 is selected as the best nesting scheme, which is then integrated with the attention mechanism module (the proposed HA U-Net). Compared with those of U-Net, the IOU, Kappa and Ins F1 have been increased by 1.77%, 1.27%, and 2.17%, respectively. The attention module is exploited to prominently utilize the lower-resolution feature maps and further improve the accuracy of the model, which contributes to a significant improvement in the index of Ins F1. It can be seen that embedding the attention module into the overall nested module further improves the model’s ability to segment individual buildings.

Meanwhile, the latest two models including D-LinkNet [45] and HRNet [46] were selected to compare with the proposed model. The central part of D-LinkNet uses a hollow convolutional layer to store spatial information, while HRNet uses parallel connections to connect high resolution feature map to low resolution feature map to maintain high resolution representation and repeated multi-scale fusion to avoid loss of information as much as possible. The same training strategy was used, and some of the final

results are shown in Fig. 7. The test data with pixels of 256×256 were re-spliced into the original size 2048×2048 . In Fig. 7, although the selected two models use different techniques to strengthen the model’s ability to segment objects, the problems of building adhesion between small buildings and incomplete building recognition still exist. It can be seen that the proposed model in this work obtains more complete boundary extraction and preserves some key pixels of buildings, showing better performance for segmenting the small buildings in dense areas.

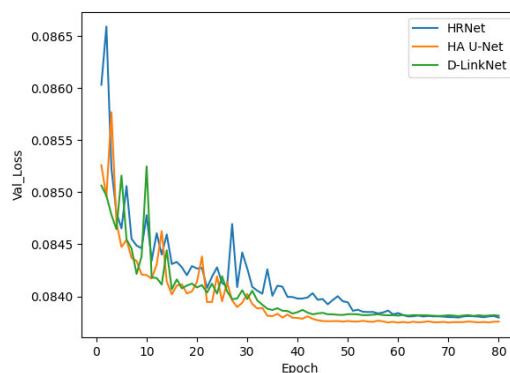


FIGURE 8. Loss in different models on Urban 3D challenge dataset.

Plots of val_loss for training different models on Urban 3D challenge dataset are shown in Fig. 8. Our model has the same convergence rate as the D-LinkNet model. Before the training epoch is 40, the loss function of the model drops sharply, and then the model parameters stabilize. From the three indicators listed in Table 3, the proposed model obtains the best result. Especially, the Ins F1 indicator of the proposed model is almost 6% higher than that of other two models,

TABLE 3. Results on the Urban 3D challenge dataset (%).

Model \Metrics	IOU	Kappa	Ins F1
HA U-Net	74.96	83.43	87.93
HRNet	69.92	79.63	81.97
D-LinkNet	70.66	80.21	81.24

indicating the better instance segmentation ability of the HA U-Net model. Besides, the proposed method also achieves the highest IOU of 74.96% and the highest kappa of 83.43

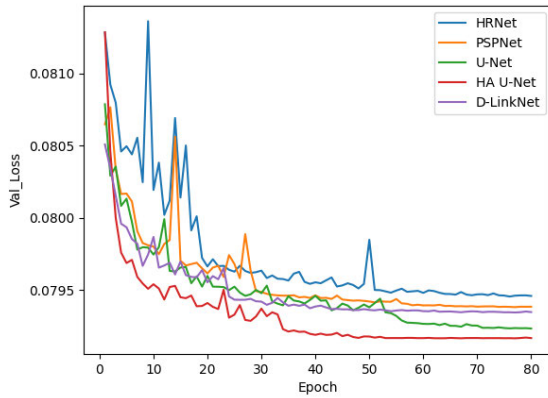


FIGURE 9. Loss in different models on WHU building dataset.

TABLE 4. Results on the WHU building dataset (%).

Model \Metrics	IOU	Kappa	Ins F1
HRNet	58.90	65.69	66.74
U-Net	61.91	69.27	70.01
PSPNet	63.52	71.02	70.93
D-LinkNet	66.78	73.98	74.01
HA U-Net	72.74	79.42	79.32

As shown in Fig 9, Our model tends to be stable when the epoch is about 35 during the training process, and the convergence speed has an advantage over several other models except the PSPNet model. The performance of the models was compared on the WHU Building Dataset, and the results are listed in Table 4. It can be seen that HA U-Net obtains the best result on the three indicators, i.e., Kappa, IoU and Instantiated F1, which are at most 13.73%, 13.84% and 12.58% higher than the worst one.

As for verification of the models on the test set, the model parameters are retained, and the average inference time of a single image (256*256 pixels) is calculated at the same time. The number of parameters and the corresponding FPS (frames per second) are listed in Table 5. It can be seen that the HA U-Net achieves an improved performance compared to the standard U-Net.

The extraction results on the WHU Building Dataset are shown in Fig. 10. As for the prediction results of each model in different regions, there is a significant improvement in the white parts of the prediction results of each model in different

TABLE 5. The inference performance of the models.

Model	Parameters(M)	FPS
HA U-Net	96.3	91
U-Net	95.4	77
D-LinkNet	118	67
HRNet	120	111
PSPNet	81.9	143

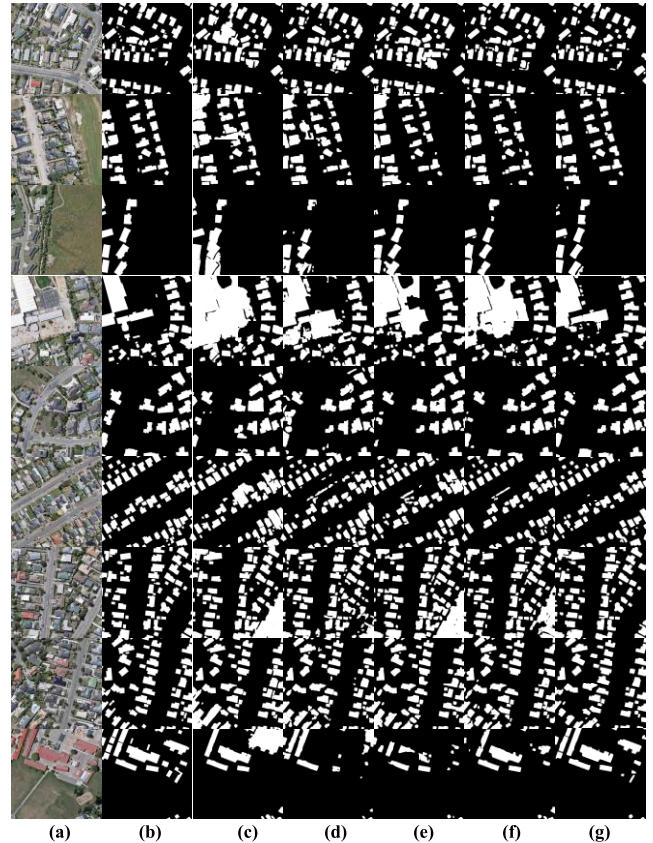


FIGURE 10. Segmentation results of different models, (a) original image, (b) GT, (c) HRNet, (d) U-Net, (e) PSPNet, (f) D-LinkNet, (g) HA U-Net.

regions. Compared with the proposed model, the other four compared models have unsatisfied performance at different objects. Firstly, the overall building maps extracted by Fig. 10. c are with much noises. That is because the parallel network of HRNet greatly increases the complexity of the network, making it difficult to train the network, which affects the model’s ability to segment objects. Besides in Fig. 10. d, buildings are comparatively not complete at small scale. This is mainly because U-Net model does not well integrate the feature information of different scale. Some misclassifications can also be seen in Figure 10. e, which mainly indicate that PSPNet [47] treats features at different scales has some limitation. Finally, as illustrated in Fig. 10. f,g, although both D-LinkNet and HA U-Net retain multi-scale features, it is not sensitive to some small and narrow areas and cannot correctly identify the building gap area due to the large receptive field in D-LinkNet.

B. WEIGHT MAPPING COMPARATIVE EXPERIMENT

The HA U-Net model on Urban3d Challenge dataset without weight mapping was taken as baseline. DWM, UWM and IWM refer to HA U-Net network with w^{DWM} , w^{UWM} and the proposed w^{IWM} weights, respectively. In the first experiment, only three-band RGB data was input to the model; in the second experiment, in addition to applying IWM, nDSM data was added to the input of the model as the fourth band to improve the robustness by learning other band information. All networks were equally initialized.

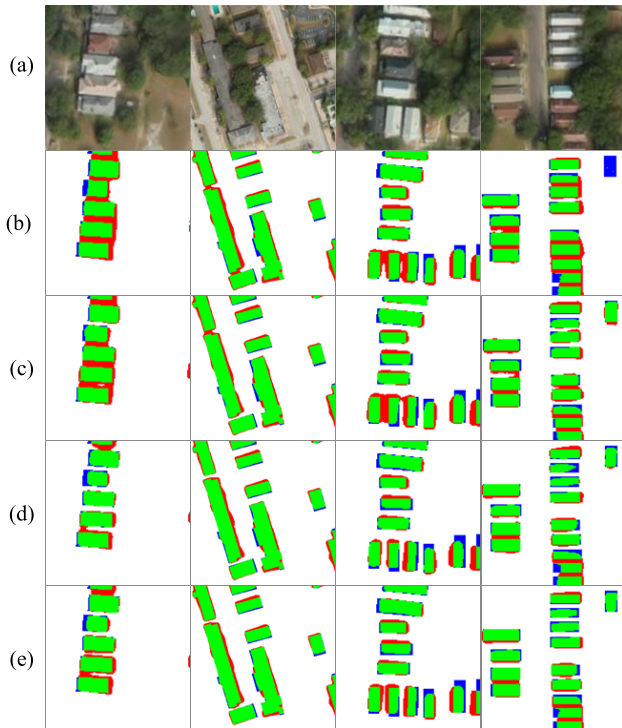


FIGURE 11. Comparison of the weight mapping results, (a) remote sensing images; (b) (c) (d) (e) represent the results of baseline, DWM, UWM, IWM, respectively.

The results of experiment 1 are illuminated in Fig. 11. Compared with the result of the baseline, the use of DWM contributes to a smoother segmentation boundary of the building. Meanwhile, the UWM method exhibits more strong segmentation ability in the interstitial area of closely adjacent buildings. In all cases, the best performance was obtained using by using IWM. The results indicate that IWM can make the model have better capability for instantiated small building extraction.

The results of different weight mapping methods are listed in Table 6. It can be seen that several weight mapping methods achieve good performance on the model. (1) Only using RGB images, IWM obtains 0.66% higher IOU and 0.79% higher Kappa coefficient than DWM, and 0.40% higher IOU and 0.22% higher Kappa coefficient than UWM. The higher performance indicates that IWM has certain advantage in the pixel-classification accuracy of the model. (2) After nDSM

TABLE 6. Experimental results of weight mapping (%).

Methods\Metrics	IOU	Kappa	Ins F1
Baseline(RGB)	73.69	82.44	86.56
DWM	74.08	82.83	86.97
UWM	74.34	82.95	88.23
IWM	74.74	83.17	88.14
RGB+nDSM	74.96	83.43	87.93
RGB+nDSM+IWM	75.32	83.71	89.36

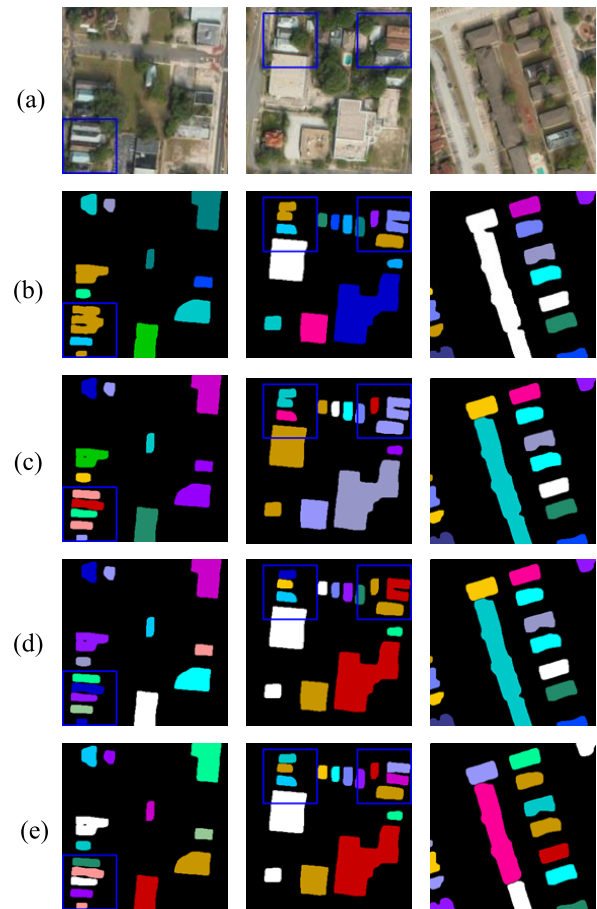


FIGURE 12. Prediction results based on watershed algorithm (a) mean image (b) (c) (d) (e) (f) means the results of baseline, watershed (0.5,0.7), watershed (0.5,0.8), watershed (0.5,0.9), respectively.

data is added as fourth band to train model and IWM is applied in model loss function, all indicators have been greatly improved, especially the Ins F1. The model combined with IWM and nDSM obtains better prediction results than the model combined with nDSM but not IWM. The IOU, Kappa, and the Ins F1 increases by 0.36%, 0.29%, and 1.43%, respectively. This shows that IWM improves both building segmentation and total pixel-classification accuracy to a certain extent. (3) Only using RGB images, the IWM method improves the IOU, Kappa coefficient, and Ins F1 by 1.05%, 0.73% and 1.58%, respectively.

TABLE 7. Experimental results of watershed algorithm (%).

Methods\Metrics	IOU	Kappa	Ins F1
HA U-Net+IWM	75.32	83.71	89.36
HA U-Net+IWM+Watershed (0.5, 0.7)	75.39	83.54	89.38
HA U-Net+IWM+Watershed (0.5, 0.8)	75.34	83.73	89.65
HA U-Net+IWM+Watershed (0.5, 0.9)	75.28	83.69	89.90

C. THE IMPACT OF WATERSHED ALGORITHM ON INSTANCE SEGMENTATION

The HA U-Net model was trained on Urban3d Challenge dataset with two types of data (RGB and nsDM). Meanwhile, IWM was used on the loss function. This combination of methods achieves better result on the urban 3D dataset. Besides, the watershed method was used to post-process the binary output of the model to achieve better instance segmentation result. To verify the effectiveness of the watershed algorithm and determine the optimal threshold, the watershed algorithm was configured with different high thresholds to perform image post-processing. To facilitate experimental analysis, the default low threshold is 0.5, and the threshold in this section defaults to the high threshold.

As shown in Fig. 12, it can be seen that (1) As the threshold increases, the house adhesion problem is gradually alleviated (Fig. 12 d). (2) The watershed algorithm refines the building boundary of the segmentation, and it leaves a watershed line inside the segmentation result of the building, which usually appears on the “house adhesion” (Figs. 12 e f). (3) Among the different threshold results, the high threshold of 0.9 corresponds to the best building instance segmentation (Fig. 12 f).

It can be seen from Table 7 that the use of watershed algorithm for image post-processing has no effect on the accuracy of the binary pixel-wise classification. But the Ins F1 steadily increases with the threshold. Compared with HA U-Net + IWM, the use of watershed algorithm with a threshold of 0.9 increases the Ins F1 by 0.54%, indicating that the use of label-based watershed algorithm for image post-processing can better solve the problem of house adhesion.

V. CONCLUSION

Regarding the existing researches on building extraction from high-resolution remote sensing images based on deep learning technology, the model’s ability to distinguish individual buildings is less concerned. The multi-scale characteristics of buildings require the model to be adjusted accordingly. Also, the remote sensing image classification in dense areas is prone to the “house adhesion” problem, where the boundary of buildings is not predicted well. Based on U-Net, this paper proposes a model called HA U-Net, which aggregates multi-scale feature maps to supervise output predictions and improve the model’s ability to recognize buildings. IWM weight mapping is introduced to make the model focus on the learning of building boundaries during model training. In addition, watershed post-processing algorithm is performed after model prediction to improve the instance

segmentation. The main research conclusions are as follows. As for model design, the best solution combining the standard U-Net with the holistically-nested network and attention mechanism is realized. The constructed network retains information of different levels, and the important semantic information at multiple scales is preserved; IWM weight mapping is performed on loss function, which integrates prior knowledge into the model and makes the model focus on the boundary area of the building and the gap area of closely adjacent buildings; the watershed post-processing algorithm further improves the instance segmentation ability of the model. The proposed model can achieve better performance than the standard U-Net and other models. It is worth noting that ResNet is as encode in the entire network design process. If the current best encoder network is used instead, it is believed that the performance of the proposed model can be further improved.

ACKNOWLEDGMENT

Thanks to the Group of Photogrammetry and Computer Vision (GPCV), Wuhan University for providing WHU Building Dataset and Goldberg *et al.* for providing Urban3D challenge dataset.

REFERENCES

- [1] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using deep learning: A survey,” 2020, *arXiv:2001.05566*. [Online]. Available: <http://arxiv.org/abs/2001.05566>
- [2] Q. Meng and X. Duan, “Scene classification of high-resolution remote sensing images based on deep convolutional neural network,” *J. Central China Normal Univ. (Natural Sci.)*, vol. 53, no. 4, pp. 568–574, Aug. 2019.
- [3] B. Huang, B. Zhao, and Y. Song, “Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery,” *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.
- [4] Y. Xu, L. Wu, Z. Xie, and Z. Chen, “Building extraction in very high resolution remote sensing imagery using deep learning and guided filters,” *Remote Sens.*, vol. 10, no. 1, p. 144, Jan. 2018.
- [5] K. Chen, W. Ouyang, C. C. Loy, D. Lin, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, and J. Shi, “Hybrid task cascade for instance segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4974–4983.
- [6] M. K. Masouleh and R. Shah-Hosseini, “Fusion of deep learning with adaptive bilateral filter for building outline extraction from remote sensing imagery,” *J. Appl. Remote Sens.*, vol. 12, no. 4, p. 1, Nov. 2018.
- [7] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, “Building extraction from multi-source remote sensing images via deep deconvolution neural networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Beijing, China, Jul. 2016, pp. 1835–1838.
- [8] L. Li, J. Liang, M. Weng, and H. Zhu, “A multiple-feature reuse network to extract buildings from remote sensing imagery,” *Remote Sens.*, vol. 10, no. 9, p. 1350, Aug. 2018.
- [9] X. Huang, W. Yuan, J. Li, and L. Zhang, “A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017.
- [10] R. Alshelhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks,” *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.
- [11] S. Ji, S. Wei, and M. Lu, “A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery,” *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3308–3322, May 2019.

- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, vol. 2012, pp. 1097–1105.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [15] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.
- [16] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [19] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11, p. 1768, Nov. 2018.
- [20] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sens.*, vol. 11, no. 3, p. 227, Jan. 2019.
- [21] M. Khoshboresh-Masouleh, F. Alidoost, and H. Arefi, "Multiscale building segmentation based on deep learning for remote sensing RGB images from different sensors," *J. Appl. Remote Sens.*, vol. 14, no. 3, p. 1, Jul. 2020.
- [22] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 1442–1450.
- [23] Z. Zhang and Y. Wang, "JointNet: A common neural network for road and building extraction," *Remote Sens.*, vol. 11, no. 6, p. 696, Mar. 2019.
- [24] X. Pan, L. Gao, A. Marinoni, B. Zhang, F. Yang, and P. Gamba, "Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network," *Remote Sens.*, vol. 10, no. 5, p. 743, May 2018.
- [25] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "TEMPORARY REMOVAL: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 42–55, Jan. 2019.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [27] J. Lin, W. Jing, H. Song, and G. Chen, "ESFNet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019.
- [28] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.
- [29] M. Aamir, Y.-F. Pu, Z. Rahman, M. Tahir, H. Naeem, and Q. Dai, "A framework for automatic building detection from low-contrast satellite images," *Symmetry*, vol. 11, no. 1, p. 3, Dec. 2018.
- [30] P. Schuegraf and K. Bittner, "Automatic building footprint extraction from multi-resolution remote sensing images using a hybrid FCN," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 4, p. 191, Apr. 2019.
- [31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [32] L. Yu, P. Wang, X. Yu, Y. Yan, and Y. Xia, "A holistically-nested U-net: Surgical instrument segmentation based on convolutional neural network," *J. Digit. Imag.*, vol. 33, no. 2, pp. 341–347, Apr. 2020.
- [33] Y. Zhuge, A. V. Krauze, H. Ning, J. Y. Cheng, B. C. Arora, K. Camphausen, and R. W. Miller, "Brain tumor segmentation using holistically-nested neural networks in MRI images," *Med. Phys.*, vol. 44, no. 10, pp. 5234–5243, Oct. 2017.
- [34] F. A. Guerrero-Pena, P. D. Marrero Fernandez, T. Ing Ren, M. Yui, E. Rothenberg, and A. Cunha, "Multiclass weighted loss for instance segmentation of cluttered cells," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2451–2455.
- [35] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.
- [36] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. Rose, and B. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the United States," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.
- [37] Y. Yu, B. F. Li, X. W. Zhang, Y. P. Liu, and H. Q. Li, "Division, V. Marked watershed segmentation algorithm for RGBD images," *J. Image Graph.*, vol. 21, pp. 145–154, Mar. 2016.
- [38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
- [39] Y. Xu, Y. Li, Y. Wang, M. Liu, Y. Fan, M. Lai, I. Eric, and C. Chang, "Gland instance segmentation using deep multichannel neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2901–2912, Mar. 2017.
- [40] A. Bieniek and A. Moga, "An efficient watershed algorithm based on connected components," *Pattern Recognit.*, vol. 33, no. 6, pp. 907–916, Jun. 2000.
- [41] H. Goldberg, M. Brown, and S. Wang, "A benchmark for building footprint classification using orthorectified RGB imagery and digital surface models from commercial satellites," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Washington, DC, USA, Oct. 2017, pp. 1–7.
- [42] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
- [43] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*. [Online]. Available: <http://arxiv.org/abs/1908.03265>
- [44] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: <http://arxiv.org/abs/1804.03999>
- [45] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 182–186.
- [46] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.
- [47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.



LEILEI XU received the M.S. degree in surveying and mapping engineering from Hohai University, Nanjing, China. His research interests include object detection and semantic segmentation.



YUJUN LIU is currently pursuing the Ph.D. degree with the Institute of Geographic Sciences and Natural Resources Research, CAS, with a focus on deep learning applied to remote sensing imagery processing. His research interests include computer vision and machine learning.



PENG YANG received the B.S. degree from the East China University of Technology, Nanchang, China, in 2018, and the M.S. degree in surveying and mapping engineering from Hohai University, Nanjing, China, in 2020. His major research interests include high-resolution remote sensing imagery semantic segmentation and computer vision.



DAN WANG was born in 1987. She received the M.S. degree in cartography and geographic information system from Nanjing Normal University, in 2012. She is currently a Senior Engineer with the Jiangsu Provincial Geomatics Centre. She has mainly engaged in technical research and engineering practice in smart cities, spatial data processing, and deep learning.



HAO CHEN was born in 1994. He received the M.Sc. degree from Tongji University, China, in 2020. He is currently pursuing the Ph.D. degree with the Technical University of Berlin. From July 2020 to January 2021, he worked as a Research Assistant with the College of Surveying and Geo-Informatics, Tongji University. His research interests include spatial data processing, planetary mapping, and deep learning.



HANYUE ZHANG was born in 1994. She is currently pursuing the Ph.D. degree with Beijing Forestry University. Her research interests include spatial data processing, forest management, and data mining.



XIN ZHANG was born in 1994. He received the B.E. degree from Tongji University, Shanghai, China, in 2016, where he is currently pursuing the Ph.D. degree with the College of Surveying and Geo-Informatics.

...