Automation in Construction 123 (2021) 103509

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Automated building extraction using satellite remote sensing imagery

Qintao Hu [a,b,c], Liangli Zhen [d], Yao Mao [a,b,*], Xi Zhou [a,b], Guozhong Zhou [a,b]

[a] Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu 610209, China
[b] Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu 610209, China
[c] University of Chinese Academy of Sciences, Beijing 100039, China
[d] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore 138632, Singapore

ARTICLE INFO

Keywords:
Remote sensing
Building extraction
Urban planning
Digital city construction
2019 MSC: 11–29 99–00.

ABSTRACT

Automatic extraction of buildings from remote sensing images plays a critical role in urban planning and digital city construction applications. In real-world applications, however, real scenes can be highly complex (e.g., various building structures and shapes, presence of obstacles, and low contrast between buildings and surrounding regions), making automatic building extraction extremely challenging. To conquer this challenge, we propose a novel method called Deep Automatic Building Extraction Network (DABE-Net). It adopts squeeze-and-excitation (SE) operations and the residual recurrent convolutional neural network (RRCNN) to construct building-blocks. Furthermore, an attention mechanism is introduced into the network to improve segmentation accuracy. Specifically, to handle small buildings, we highlight small buildings and develop a multi-scale segmentation loss function. The theoretical analysis and experimental results show that the proposed method is effective in building extraction and outperforms several peer methods on the dataset of Mapping challenge competition.

1. Introduction

Automatic building extraction, which identifies buildings from the captured images, has been widely applied in many applications, such as urban planning [1,2], geographic information system (GIS) data updating [3,4], damage assessment [5,6] and digital city construction [7,8]. Early research studies on building extraction are usually done based on aerial imagery [9] due to its high spatial resolution. Nevertheless, it is time-consuming to obtain the images of a large area like the whole city.

In recent decades, the availability of high-resolution satellite imaging sensors provides a new data source for automatic building extraction. The high spatial resolution of remote sensing imagery reveals fine details in urban areas and greatly facilitates the automatic building extraction. A large number of methods have been developed using remote sensing imagery. For instance, based on traditional image processing approaches [10–13], traditional methods consider spectra, shape, and texture as the input features. Then they take support vector machine (SVM), random forest (RF), or AdaBoost as the classifier [12–14] to extract buildings. However, designing the feature extractors requires high expertise in the area, and the obtained features may not be suitable for new datasets.

Inspired by the great success of deep learning in image classification [15], speech recognition [16], and machine translation [17], some researchers applied deep learning approaches to remote sensing segmentation tasks [18–23]. There are also a few attempts on building extraction [1,24–27]. For example, Li et al. [3] proposed a U-Net-based semantic segmentation method for the extraction of building footprints from high-resolution multispectral satellite images and multi-source GIS data. Lu et al. [25] proposed a building edge detection model using a richer convolutional features (RCF) network. The RCF-building model could detect building edges more accurately and obtain a significant performance improvement over the baselines. Shrestha et al. [28] proposed a fully connected network-based building extraction approach by combining the exponential linear unit (ELU) and conditional random fields (CRFs). Wu et al. [29] presented a boundary regulated network called BR-Net for accurate aerial image segmentation and building outline extraction. The BR-Net achieves significantly higher performance than the U-Net model.

Although these methods have achieved promising performance in simple scenarios, real-world applications usually involve in highly complex scenes. For instance, the structure and the shape of buildings

* Corresponding author at: Key Laboratory of Optical Engineering, Chinese Academy of Sciences, Chengdu 610209, China.
E-mail address: maoyao@ioe.ac.cn (Y. Mao).

https://doi.org/10.1016/j.autcon.2020.103509
Received 29 November 2019; Received in revised form 23 September 2020; Accepted 29 September 2020
Available online 16 December 2020

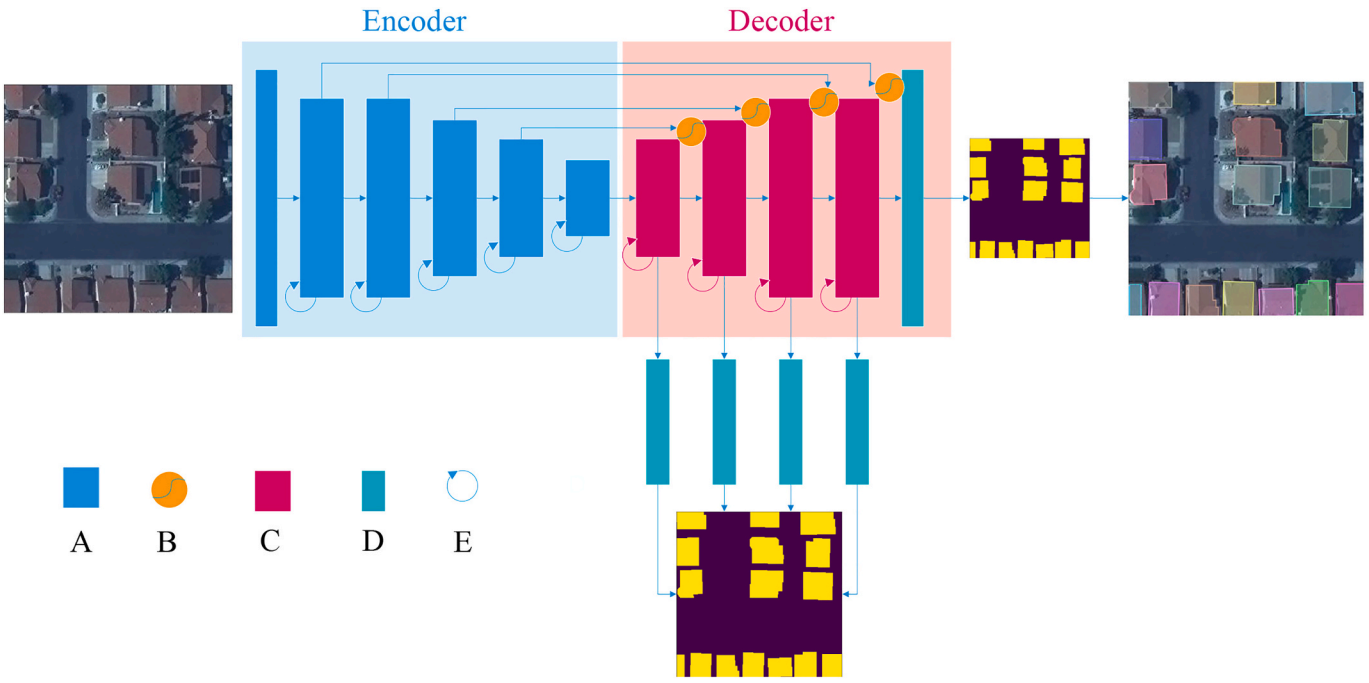

Fig. 1. The flowchart of the proposed DABE-Net. It has a U-Net shape CNN structure and includes an encoder and a decoder, where A is the SERRCNN-block, B is the Attention gates, C is the SERRCNN-block with up-sampling operation, D is the up-sampling operation and E is a recurrent convolutional operation.

vary largely in different countries; the presence of obstacles posed by surrounding objects, like tress and billboards, and the contrast between buildings and surrounding regions may be extremely low. It makes automatic extraction of buildings from remote sensing images challenging, and the performance of existing methods deteriorates sharply.

To conquer the challenges above, in this paper, we propose a novel method called Deep Automatic Building Extraction Network (DABE-

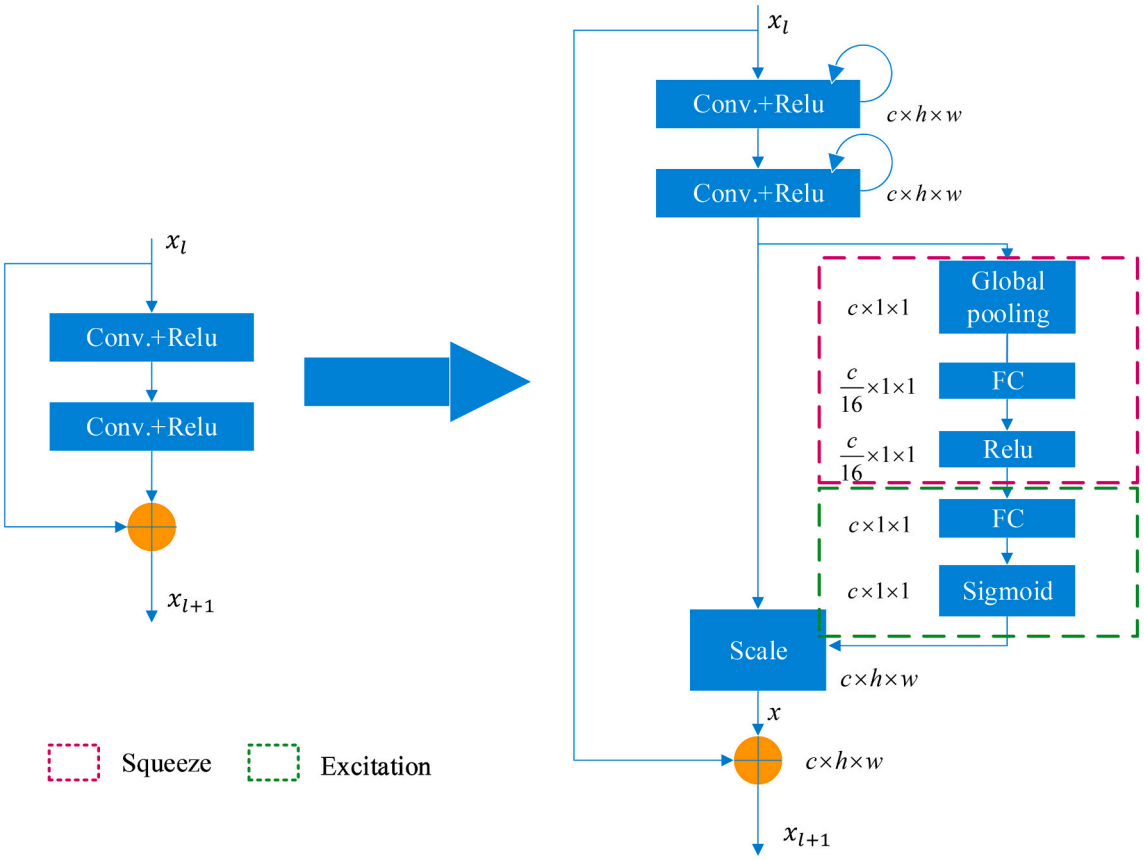

Fig. 2. The proposed SERRCNN-block that combines Squeeze-and-Excitation operations with Recurrent Convolutional Neural Networks on the residual model.

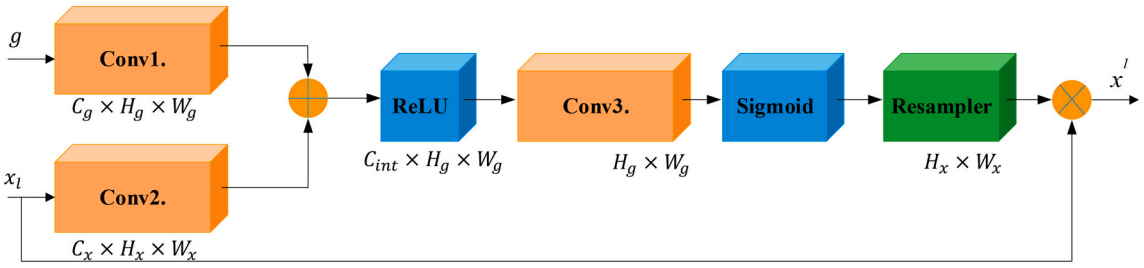

Fig. 3. Schematic of the attention gate (AG). x^l is the input features and α is attention coefficient computed in AG. g is the input of the encoder, x^l is the decoder featuremaps as the same size as g.

Net). It adopts squeeze-and-excitation (SE) operations and the residual recurrent convolutional neural network (RRCNN) to construct building-blocks. Furthermore, an attention mechanism is introduced into the network to improve segmentation accuracy. Specifically, to handle small buildings, we highlight small buildings and develop a multi-scale segmentation loss function. As shown in Fig. 1, we design two loss functions for our proposed model, including the multi-scale segmentation loss and the segmentation loss. Following this learning strategy, our DABE-Net achieves promising performance on the building extraction dataset. The novelty and the main contribution of this work are two-fold: 1) A novel deep model is developed for automatic building extraction from remote sensing images. It includes SE and RRCNN blocks and involves attention gates to attach importance to network channel information and global information, and 2) Unlike existing methods, our method can effectively balance samples and focus on buildings of different scales. Extensive experiments on the Mapping challenge competition dataset have been conducted. The results demonstrate that our method outperforms several peer methods for automatic building extraction, which indicates the effectiveness of the proposed method.

The remainder of this paper is as follows. Section 2 presents our proposed method. Experimental results and discussion are provided in Section 3. We conclude this paper in Section 4.

2. Method

Our proposed deep automatic building extraction network (DABE-Net) has a U-Net shape CNN structure, which involves an encoder and a decoder. We introduce the SE and RRCNN (SERRCNN)-block and the attention gates into the encoder and the decoder. The SERRCNN-block can enhance the discriminative features, and the attention gates can extract semantic contextual information. Meanwhile, we improve the lovÃ¡sz hinge loss to balance the background and the building.

2.1. SERRCNN-block

Recently, the deep convolution neural network is used in the semantic segmentation of high-resolution remote sensing images with excellent performance. However, the information on convolutional features among different channels is not effectively utilized. We introduce the squeeze-and-excitation operations [30] to learn the attention weights of different feature channels automatically. According to the attention weights, discriminative features are enhanced while redundant features for the target tasks are suppressed. Simultaneously, there are rich feature details in the remote sensing image. We further adopt the recurrent convolutional neural network (RCNN) [31] to extract the information about the details from image features. The operation improves building boundary information.

Including the Squeeze-and-Excitation operations and the recurrent convolutional neural networks, we develop the Squeeze-and-Excitation Residual RCNN (SERRCNN)-block, as shown in Fig. 2. RCNN and its variants have already shown superior performance on object recognition tasks using different benchmarks [31], we introduce it into DABE-Net, by following [31], we denote x_l as the input of l^{th} layer of SERRCNN-block and a pixel located at (i,j) in an input sample on the k^{th} feature map in the Recurrent Convolutional Layers (RCL) [31]. The output of the network at the time step t, $O^l_{ijk}(t)$ can be expressed as Eq. (1) [31]:

$$O^l_{ijk}(t) = \left(w^f_k\right)^T \times x^{f(i,j)}_l(t) + \left(w^r_k\right)^T \times x^{r(i,j)}_l(t-1) + b_k \tag{1}$$

where $x^{f(i,j)}_l(t)$ and $x^{r(i,j)}_l(t-1)$ are the inputs of the l^{th} standard convolution layers and the l^{th} RCL, respectively. The values of $(w^f_k)^T$ and $(w^r_k)^T$ are the weights of the standard convolutional layer and the RCL of the k^{th} feature map respectively, and b_k is the corresponding bias.

The output of l^{th} RCL with the activations can be expressed as Eq. (2):

$$u = F(x_l, w_l) = \sigma\left(O^l_{ijk}(t)\right) \tag{2}$$

where F is the recurrent operation, σ refers to the ReLU [32] function, and u is the output of l^{th} RCL.

To tackle the issue of exploiting channel dependencies, we consider the signal to each channel in the output features. We compress the feature along the spatial dimension by turning each two-dimensional feature channel into $z \in \mathbb{R}^c$. It has a global receptive field to some extent, and the output dimension matches the number of input feature channels. z is generated by shrinking u through spatial dimensions $C \times H \times W$ into $C \times 1 \times 1$, where C is the number of channels and $H \times W$ is the size of feature map, the c^{th} element of z is calculated via Eq. (3) by following [30]:

$$Z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{3}$$

To make use of the information aggregated in the squeeze operation, we adopt the following operation to capture channel-wise dependencies.

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \sigma(W_1 z)) \tag{4}$$

where $W_1 \in \mathbb{R}^{\frac{C}{16} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{16}}$.

The final output of the block can be obtained by rescaling the transformation output u with the activations as

$$\tilde{x} = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{5}$$

and

$$x_{l+1} = x_l + \tilde{x}_c \tag{6}$$

The activations act as channel weights adapted to the input. In this regard, SE blocks intrinsically introduce dynamics conditioned on the input, helping to boost feature discriminability [30].

2.2. Attention gates

The complexity of the background and the diversity of the object structures make it easy to cause error extraction and inaccurate segmentation of the semantic boundary of the ground objects. To capture a sufficiently large receptive field and semantic contextual information,

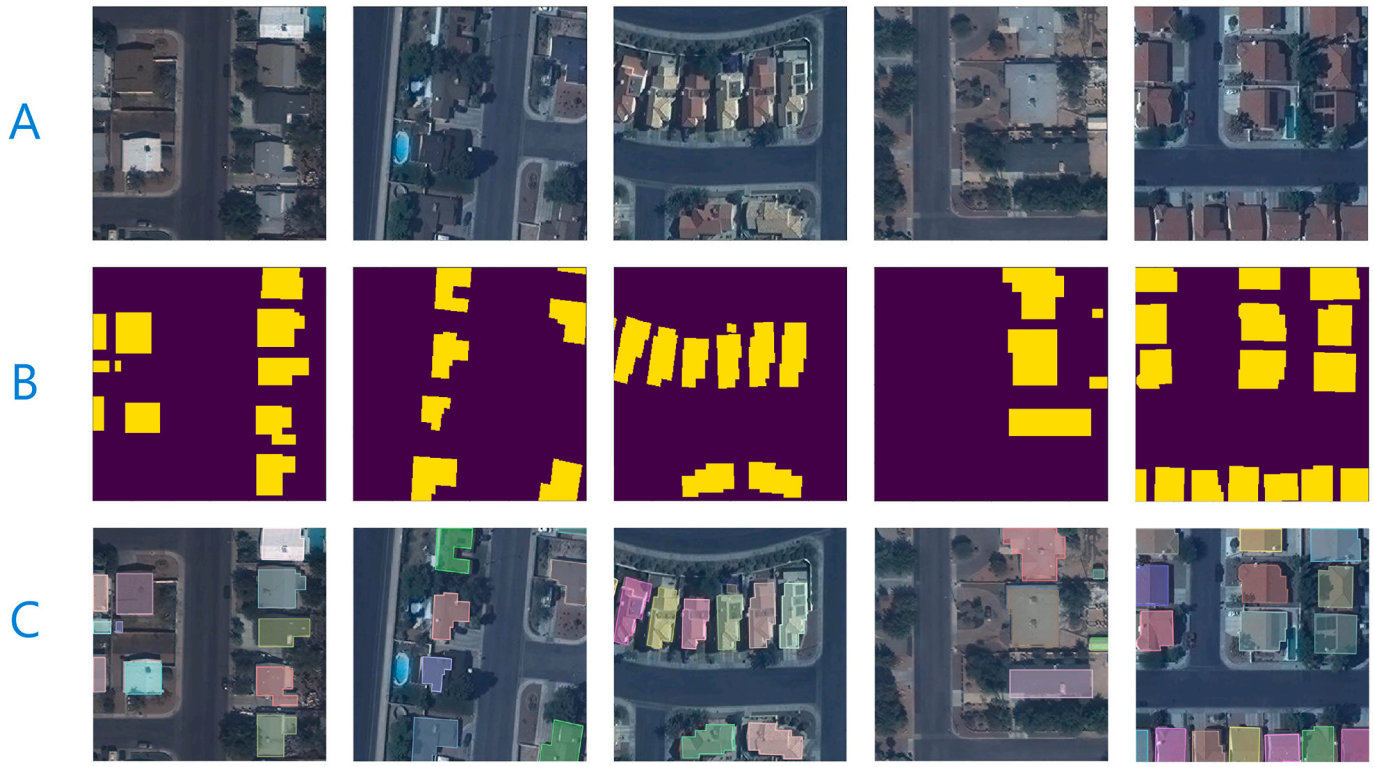

Fig. 4. Dataset on CrowdAI Mapping Challenge, In the figure, Row A shows the remote sensing images, Row B shows the masks, Row C shows the ground-truths.

researchers usually downsample the feature-map grid gradually in standard CNN architectures. Note that the shallow convolutional features are important for extracting some low-level information like colors and edges. Discarding these detailed features may reduce false-positive predictions for building objects. Through attention operation, we progressively suppress feature responses in background regions, such as cars and roads, and improve the network's attention to the building features.

As shown in Fig. 3, we introduce attention coefficient, $\alpha_i \in [0,1]$, identifies salient image regions and restrain feature responses to merely retain activations related to specific tasks. The output of AGs is the element multiplication of input feature mapping and attention coefficient: $\hat{x}_{i,c}^{l} = x_{i,c}^{l} \cdot \alpha_i^l$, from Fig. 3, where α_i^l is a single scalar attention value and computed for each pixel vector $x_i^l \in \mathbb{R}^{F_l}$ and F_l corresponds to the number of feature-maps in layer l. $g_i \in \mathbb{R}^{F_g}$ is used for each pixel i to determine focused regions called gating vector that contains contextual information to prune lower-lever feature responses. Regarding the gating coefficient, we use additive attention in [33], which has been proved being effective in feature extraction. Additive attention is formulated as

$$\alpha_i^l = \sigma_2\left(W_3^T\left(\sigma_1\left(W_1^T s_i^l + W_2^T g_i + b_1 + b_2\right)\right) + b_3\right) \tag{7}$$

where σ_1 refers to the ReLU [34] function and σ_2 is a sigmoid [35] function, W_i, b_i is the parameters and bias of $Conv_i$, respectively. $W_1 \in \mathbb{R}^{F_l \times F_{int}}$, $W_2 \in \mathbb{R}^{F_g \times F_{int}}$, $W_3 \in \mathbb{R}^{F_{int} \times 1}$, $b_{1,\,2} \in \mathbb{R}^{F_{int}}$, and $b_3 \in \mathbb{R}$.

2.3. Symmetric extension of the lovÃ¡sz hinge loss function

In building extraction, largely different scales of input images may make the network hard to handle. The imbalance between the building and the background further worsens the effect of the network. To conquer this issue, we improve the lovÃ¡sz hinge loss to balance the background and buildings.

Traditionally, the logic regression loss function is used to optimize cross-entropy loss for semantic segmentation. However, due to the measure of cross-entropy loss on a validation set is constantly a poor indicator of the quality of segmentation. The intersection-over-union (IoU) score, namely the Jaccard index is widely adopted to evaluate segmentation masks. According to definition in [36], $F_i(x)$ is the i-th element of the output of the network, given a vector of ground truth labels y^* and predicted labels $\tilde{y}$, the Jaccard index of class c is defined as

$$J_c(y^*, \tilde{y}) = \frac{|\{y^* = c\} \cap \{\tilde{y} = c\}|}{|\{y^* = c\} \cup \{\tilde{y} = c\}|} \tag{8}$$

The corresponding loss function used in empirical risk minimization to train the network as follows:

$$\Delta J_c(y^*, \tilde{y}) = 1 - J_c(y^*, \tilde{y}) \tag{9}$$

where,

$$\tilde{y}_i = sign(F_i(x)) \tag{10}$$

To use a max margin classifier, following [36], in the binary case, the original lovász hinge loss associated with the prediction of the pixel i is computed as

$$m_i = max\left(1 - F_i(x) y_i^*, 0\right) \tag{11}$$

Considering the imbalance between the background and the building, we evolve it by symmetric extension into the new symmetric lovász hinge loss to solve the problem of data imbalance in the binary segmentation:

$$m_i^* = \left(max\left(1 - F_i(x) y_i^*, 0\right) + max\left(F_i(x)\left(1 - y_i^*\right), 0\right)\right)/2 \tag{12}$$

where m_i and $m_i^* \in \mathbb{R}^+$ are the vector of hinge losses.

3. Experiment and discussion

In this section, we first introduce the used dataset. We then introduce the experiment setups and evaluation metrics. At last, we evaluate the

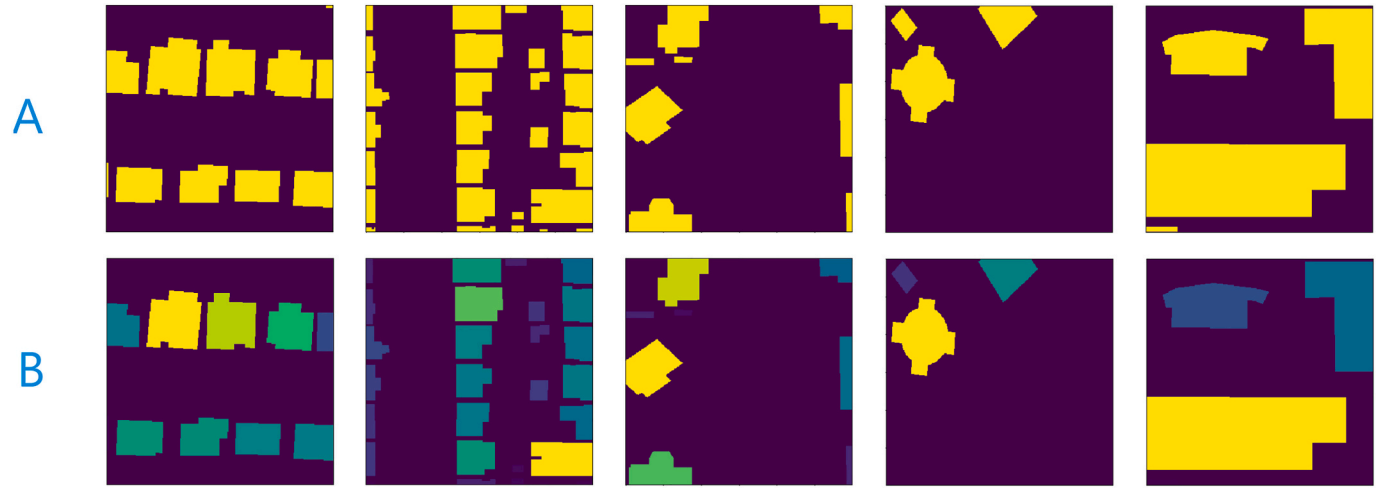

Fig. 5. The weights of different buildings. We use the color to represent the weights of buildings, and the smaller buildings have larger weights, the darker the color. Row A shows the masks and Row B displays the weights.

DABE-Net on the mapping challenge competition dataset and provide some discussions.

3.1. Dataset and preprocessing

The dataset is derived from the mapping challenge competition dataset on crowdai consisting of 280,741 training data and 60,317 Validation data. The image includes 300 × 300 pixels; each pixel is divided into buildings and backgrounds.

Since image segmentation is a data-driven algorithm, accurate results can be obtained according to the high diversity and quality of datasets. Data expansion is an effective method to improve performance by using the same amount of data. In this study, we take several specific methods to expand the training dataset, such as rotation, flipping, inserting random color jitter, and randomly clipping to preprocess each input image. Each image is zoomed to 256 × 256. Some examples of the original images and labels of the dataset are shown in Fig. 4.

To improve the accuracy of extracting small buildings through the network, we design the attention weight of learning, and give larger weights to small buildings, so that the network can focus on small buildings. As shown in Fig. 5, we use the color to represent the weight of buildings, and the smaller buildings have larger weights, the darker the color.

3.2. Training

We implement our model in PyTorch. The network is randomly initialized under the default setting of PyTorch with no pretraining on any external dataset. All experiments are run on a desktop computer equipped with 2 Intel Xeon E5-2678v3 CPU, 64 GB memory, 4 NVIDIA 2080Ti GPUs (with 11 GB*4 video memory) and Ubuntu 16.04 OS. We use the Adam optimization algorithm for training the network and set default learning parameters. The initial learning rate is $2e-4$ and decays to $1e-5$ after 20 epochs. To fairly compare different methods, the batch size and epochs for training are fixed to 40 and 60, respectively.

3.3. Evaluation metrics

We adopt the evaluation metric of the COCO 2012 dataset, and the segmentation task is assessed by the precision and recall. Segmentations are determined as true or false positives according to the area of overlap with ground-truth. Eq. (13) shows the definition of IoU for evaluating whether a detected building polygon is accurate, which equals the overlap region of a detected building polygon (denoted by B_{gt}) and a ground truth building polygon (denoted by B_{gt}) divided by the union area of B_p and B_{gt} [3].

$$IoU = \frac{aera(B_p \cap B_{gt})}{aera(B_p \cup B_{gt})} \tag{13}$$

If IoU between a detected building polygon and a ground truth building polygon is larger than 0.5, we consider the building polygon as correctly detected. Precision (Eq. (14)) and recall (Eq. (15)) are the common evaluation metrics for COCO dataset. The results of each image are evaluated independently, and the final F1-score is the average value of F1-scores (Eq. (16)) for each image.

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{16}$$

where true positive (TP) indicates the number of building polygons that are detected correctly, false positive (FP) is the number of other objects that are detected as building polygons by mistake, and false negative (FN) represents the number of building polygons not detected.

Table 1
Ablation study of DABE-Net on the Mapping challenge competition dataset. We compare the baseline and analyze the impact of each module and the combination of modules.

Model	DABE-Net	Improved loss	Precision	Recall	F1-score
U-Net			92.1	92.5	92.3
U-Net_i		✓	92.8	93.1	92.9
DABE-Net	✓		93.1	93.2	93.1
DABE-Net_i	✓	✓	94.1	95.2	94.6

3.4. Ablation experiment

To verify the effectiveness of different components, we construct and evaluate the following variants:

U-Net: U-Net + original loss function;
U-Net_i: U-Net + improved loss function;
DABE-Net: DABE-Net + original loss function;
DABE-Net_i: DABE-Net + improved loss function.

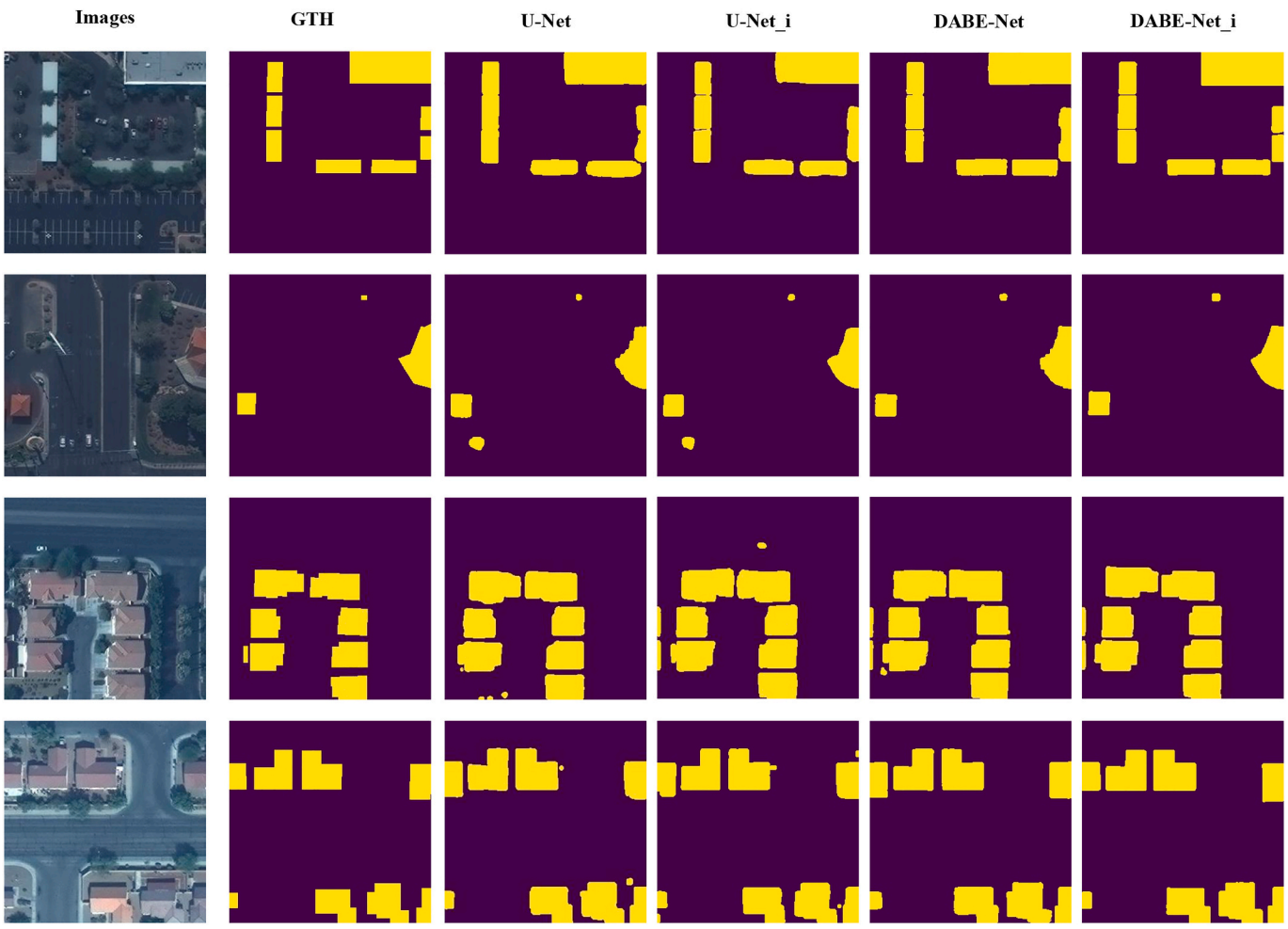

Fig. 6. Comparison of experimental results of four different models, our method (last column) can enrich enhance the details so that buildings in output have more distinct, margin and small buildings are easier to be detected.

Table 1 shows the precision, recall, and F1-scores of different tested methods in building extraction on validation datasets. In the four experiments, the full proposed model achieves the highest scores. A comparison of the results of our DABE-Net and other methods on some examples are shown in Fig. 6, from which we can also see that the full model can achieve much better results.

3.5. Dscussion

From Table 1, by comparing between the U-Net model and the U-Net_i model, we can see that the precision value is increased by 0.7%, the recall rate is increased by 0.6%. From the result of DABE-Net and DABE-Net_i, we observe that the precision is increased by 1.0%, the recall rate is increased by 2.0%, and the advantage of symmetric extension lovÃ¡sz hinge loss is concluded. Similarly, the comparison between the results of U-Net and DABE-Net demonstrates the performance of ARRSEU-Net by the increment of 1.0% for precision and 0.7% for recall. Meanwhile, the comparison between the results of U-Net_i and DABE-Net_i demonstrates the performance of ARRSEU-Net by the increment of 1.3% in terms of precision and 2.1% in terms of recall. Finally, the comparison between the results of U-Net and DABE-Net_i indicates that our proposed method raises precision by 2% and recall by 2.7%.

As shown in Fig. 6, our method (last column) can enhance the details of the buildings. Furthermore, the proposed method visibly reduces the interference of aground vehicles and other buildings. To some extent, the method is even able to correct a few mistakenly marked areas in ground-truth labels.

4. Conclusion

In this paper, we proposed a new U-shaped network called DABE-Net to extract buildings from remote sensing images. We developed the SERCNN-block by integrating squeeze-and-excitation operations and RRCNN, which can extract image features more accurately and capture image details and network channel features. Furthermore, the attention gates are introduced into the network to enlarge the network's weight for important features and improve symmetric extension lovÃ¡sz hinge loss function. To enhance the accuracy of the network to small buildings, we increased the attention weights of small buildings and developed multi-scale segmentation loss for the learning process. We used the Mapping challenge competition dataset on CrowdAi. The experimental results show that our network achieved 94.1% precision, 95.2% recall and 94.6% F1-score. Compared with U-Net, it increased 2.3% F1-score. Experimental results support our conclusions and prove the effectiveness of the network. In the future, we will focus on accelerating our method's speed and extending it to handle more complex remote sensing image segmentation tasks.

Declaration of Competing Interest

None.

Acknowledgments

The work is partially supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A1b0045).

References

[1] R. Alshehhi, P.R. Marpu, W.L. Woon, M. Dalla Mura, Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks, ISPRS J. Photogramm. Remote Sens. 130 (2017) 139–149, https://doi.org/10.1016/j.isprsjprs.2017.05.002.
[2] X. Gao, M. Wang, Y. Yang, G. Li, Building extraction from rgb vhr images using shifted shadow algorithm, IEEE Access 6 (2018) 22034–22045, https://doi.org/10.1109/ACCESS.2018.2819705.
[3] W. Li, C. He, J. Fang, J. Zheng, H. Fu, L. Yu, Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data, Remote Sens. 11 (4) (2019) 403, https://doi.org/10.3390/rs11040403.
[4] Q. Bi, K. Qin, H. Zhang, Y. Zhang, Z. Li, K. Xu, A multi-scale filtering building index for building extraction in very high-resolution satellite imagery, Remote Sens. 11 (5) (2019) 482, https://doi.org/10.3390/rs11050482.
[5] C. Xiong, Q. Li, X. Lu, Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network, Autom. Constr. 109 (2020) 102994, https://doi.org/10.1016/j.autcon.2019.102994.
[6] A.J. Cooner, Y. Shao, J.B. Campbell, Detection of urban damage using remote sensing and machine learning algorithms: revisiting the 2010 Haiti earthquake, Remote Sens. 8 (10) (2016) 868, https://doi.org/10.3390/rs8100868.
[7] B. Zhang, C. Wang, Y. Shen, Y. Liu, Fully connected conditional random fields for high-resolution remote sensing land use/land cover classification with convolutional neural networks, Remote Sens. 10 (12) (2018) 1889, https://doi.org/10.3390/rs10121889.
[8] W. Li, H. Fu, L. Yu, A.P. Cracknell, Deep learning based oil palm tree detection and counting for high-resolution remote sensing images, Remote Sens. 9 (1) (2016) 22, https://doi.org/10.3390/rs9010022.
[9] E. Tarantino, B. Figorito, Extracting buildings from true color stereo aerial images using a decision making strategy, Remote Sens. 3 (8) (2011) 1553–1567, https://doi.org/10.3390/rs3081553.
[10] G. Cheng, J. Han, A survey on object detection in optical remote sensing images, ISPRS J. Photogramm. Remote Sens. 117 (2016) 11–28, https://doi.org/10.1016/j.isprsjprs.2016.03.014.
[11] S. Ahmadi, M.V. Zoej, H. Ebadi, H.A. Moghaddam, A. Mohammadzadeh, Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours, Int. J. Appl. Earth Obs. Geoinf. 12 (3) (2010) 150–157, https://doi.org/10.1016/j.jag.2010.02.001.
[12] M. Belgiu, L. Drăguţ, Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery, ISPRS J. Photogramm. Remote Sens. 96 (2014) 67–75, https://doi.org/10.1016/j.isprsjprs.2014.07.002.
[13] X. Huang, L. Zhang, Morphological building/shadow index for building extraction from high-resolution imagery over urban areas, IEEE J. Select. Topics Appl. Earth Observ. Remote Sens. 5 (1) (2011) 161–172, https://doi.org/10.1109/JSTARS.2011.2168195.
[14] M. Awrangjeb, C. Fraser, Automatic segmentation of raw lidar data for extraction of building roofs, Remote Sens. 6 (5) (2014) 3716–3751, https://doi.org/10.3390/rs6053716.
[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Iformation Processing Systems, 2012, pp. 1097–1105, https://doi.org/10.1145/3065386.
[16] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke, The Microsoft 2017 Conversational Speech Recognition System, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018, pp. 5934–5938, https://doi.org/10.1109/icassp.2018.8461870.
[17] M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, et al., Google's multilingual neural machine translation system: enabling zero-shot translation, Trans. Assoc. Comput. Linguistics 5 (2017) 339–351, https://doi.org/10.1162/tacl_a_00065.
[18] W. Li, C. He, J. Fang, H. Fu, Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 238–241, https://doi.org/10.1109/CVPRW.2018.00043.
[19] H. Lin, Z. Shi, Z. Zou, Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network, Remote Sens. 9 (5) (2017) 480, https://doi.org/10.3390/rs9050480.
[20] Y. Bai, E. Mas, S. Koshimura, Towards operational satellite-based damage-mapping using u-net convolutional network: a case study of 2011 tohoku earthquake-tsunami, Remote Sens. 10 (10) (2018) 1626, https://doi.org/10.3390/rs10101626.
[21] I. Sa, M. Popovic, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, R. Siegwart, Weedmap: a large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming, Remote Sens. 10 (9) (2018) 1423, https://doi.org/10.3390/rs10091423.
[22] R. Liu, Q. Miao, B. Huang, J. Song, J. Debayle, Improved road centerlines extraction in high-resolution remote sensing images using shear transform, directional morphological filtering and enhanced broken lines connection, J. Vis. Commun. Image Represent. 40 (2016) 300–311, https://doi.org/10.1016/j.jvcir.2016.06.024.
[23] G. Cheng, F. Zhu, S. Xiang, Y. Wang, C. Pan, Accurate urban road centerline extraction from vhr imagery via multiscale segmentation and tensor voting, Neurocomputing 205 (2016) 407–420, https://doi.org/10.1016/j.neucom.2016.04.026.
[24] S. Ji, S. Wei, M. Lu, A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery, Int. J. Remote Sens. 40 (9) (2019) 3308–3322, https://doi.org/10.1080/01431161.2018.1528024.
[25] T. Lu, D. Ming, X. Lin, Z. Hong, X. Bai, J. Fang, Detecting building edges from high spatial resolution remote sensing imagery using richer convolution features network, Remote Sens. 10 (9) (2018) 1496, https://doi.org/10.3390/rs10091496.
[26] B. Huang, K. Lu, N. Audeberr, A. Khalel, Y. Tarabalka, J. Malof, A. Boulch, B. Le Saux, L. Collins, K. Bradbury, et al., Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark, in: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 6947–6950, https://doi.org/10.1109/IGARSS.2018.8518525.
[27] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, Y. Xu, Building extraction in very high resolution imagery by dense-attention networks, Remote Sens. 10 (11) (2018) 1768, https://doi.org/10.3390/rs10111768.
[28] S. Shrestha, L. Vanneschi, Improved fully convolutional network with conditional random fields for building extraction, Remote Sens. 10 (7) (2018) 1135, https://doi.org/10.3390/rs10071135.
[29] G. Wu, Z. Guo, X. Shi, Q. Chen, Y. Xu, R. Shibasaki, X. Shao, A boundary regulated network for accurate roof segmentation and outline extraction, Remote Sens. 10 (8) (2018) 1195, https://doi.org/10.3390/rs10091496.
[30] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, https://doi.org/10.1109/TPAMI.2019.2913372.
[31] S. Jetley, N.A. Lord, N. Lee, P.H. Torr, Learn to pay attention, in: ArXiv, 2018, pp. 1–15. https://arxiv.org/abs/1804.02391.
[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848, https://doi.org/10.1109/TPAMI.2017.2699184.
[33] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: ArXiv, 2014, pp. 1–14. https://arxiv.org/abs/1409.0473.
[34] J. Han, C. Moraga, The influence of the sigmoid function parameters on the speed of backpropagation learning, in: International Workshop on Artificial Neural Networks, Springer, 1995, pp. 195–201, https://doi.org/10.1007/3-540-59497-3_175.
[35] Y. Ito, Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory, Neural Netw. 4 (3) (1991) 385–394, https://doi.org/10.1016/0893-6080(91)90075-G.
[36] M. Berman, A. Rannen Triki, M.B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4413–4421, https://doi.org/10.1109/CVPR.2018.00464.