Multiconstraint Transformer-Based Automatic Building Extraction From High-Resolution Remote Sensing Images

Wei Yuan, *Member, IEEE*, Weihang Ran, *Student Member, IEEE*, Xiaodan Shi, *Member, IEEE*, and Ryosuke Shibasaki, *Member, IEEE*

***Abstract*—Building extraction from very high-resolution remote sensing images is a fundamental task and is widely used in applications, such as change detection, disaster assessment, and real-time update of geographic information databases. However, due to the complexity of the geographical environment and the diversity of target features, accurate automatic building extraction remains very challenging. With the fast development of deep learning techniques, convolutional neural networks (CNN) have been widely used in remote sensing research and have achieved considerable results. But for large urban area-based building detection tasks, the CNN-based method usually gets into local optima and generates many false positive detections around building boundaries. To avoid the local optima and be aware of nonlocal information, this article proposes a hybrid feature extraction model based on the combination of the CNN and Transformer to realize the automatic building detection from very high-resolution remote sensing images. Meanwhile, a multiconstraint weighting mechanism is proposed to enhance the ability of the model to recognize the regular geometric boundaries of buildings. Comprehensive experiments are conducted on the three different datasets. The proposed MC-TRANSU achieves the best F1-score and intersection over union, compared with the state-of-the-art methods, such as SegNet, TransUnet, and Swin-Unet, and the detection accuracy improved around 5%. Quantitative and qualitative results verify the superiority and effectiveness of our model.**

***Index Terms*—Building extraction, multiconstraint, remote sensing image, Transformer.**

I. Introduction

BUILDING detection is a fundamental and significant task that aims to locate all buildings in the remote sensing image. It can be an upstream task in various applications, such as urban planning [1], environment monitoring [2], and land resource utilization [3]. In recent years, with the rapid development of smart city construction, the need for updating city maps has also increased, which requires faster and more accurate building detection technology. Therefore, related research on this topic is still indispensable.

Specific object detection technology has been discussed in remote sensing studies for decades, which can be seen in [4] and [5]. Traditional building detection methods are mainly based on handcrafted features. Lin and Nevatia [6] described a method based on shadow and the geometric shape of roofs to detect buildings from monocular aerial images. Zhang [7] proposed a hybrid method that utilizes both multispectral images and optical images for building feature extraction and through the texture filtering to determine the building detection; Li and Wu [8] used an edge descriptor to detect buildings from LiDAR and aerial images. With the fast development of machine learning algorithms, the clustering and classification machine learning methods, such as support vector machine [9], [10] and random forests [11], [12], were proposed for building detection from remote sensing images. However, the accuracy of these kinds of methods is either highly dependent on the adaptation degree of manually constructed features or on the fusion of multisource data, and it is very sensitive to the shape and density changes of the target. The stability and robustness of these methods are poor and cannot meet the needs of large-scale applications.

In the past decades, deep learning-based methods have dominated all kinds of benchmarks. Since LeCun et al. [13] proposed LeNet, convolutional neural networks (CNN) have been widely concerned by researchers for their excellent performance in the task of classification, which led to the born of many effective image semantic segmentation models, such as FCN [14], U-Net [15], SegNet [16], DeconvNet [17], and PSPNet [18]. Because of the good performance of these models on traditional indoor and outdoor scene segmentation, researchers have also applied them to remote sensing images-based urban mapping and scene understanding [19], [20]. Zuo et al. [21] proposed an HF-FCN that employs hierarchically fused FCN to achieve building extraction. Shrestha and Vanneschi [22] added conditional random fields into FCN. Abdollahi and Pradhan [23] carried out a model named MultiRes-UNet that integrates semantic edges with polygons. Compared with traditional machine learning-based methods, these CNN-based methods provide much higher detection accuracy and robustness to the shape

Manuscript received 24 May 2023; revised 17 August 2023; accepted 24 September 2023. Date of publication 27 September 2023; date of current version 13 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 41771479 and in part by the Japan Society for the Promotion of Science under Grant 23K13419. *(Wei Yuan and Weihang Ran contributed equally to this work.) (Corresponding author: Wei Yuan.)*

The authors are with the Center for Spatial Information Science, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: miloyw@iis.u-tokyo.ac.jp; ranweihang@csis.u-tokyo.ac.jp; shixiaodan@csis.u-tokyo.ac.jp; shiba@csis.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/JSTARS.2023.3319826

and density variations of the building instance. However, the CNN-based methods show limitations in building global contexts and long-range dependencies within images, especially limiting the performance in delivering complete building boundaries in building detection tasks.

In 2017, a new kind of deep learning model architecture containing an attention mechanism named Transformer was proposed and started a new wave of enthusiasm in natural language processing fields [24]. These successes encouraged researchers also start applying Transformer models for image data analysis. Vision Transformer (ViT) [25] was first proposed for image data processing. After that, Swin Transformer [26] was proposed to reduce the significant computational complexity of ViT. In order to achieve fine-grained visual tasks, TransUnet [27] combined UNet and ViT and got significant performance improvements in medical image segmentation. After that, Swin-Unet [28] was proposed as the first pure Transformer-based U-shaped model to leverage the power of Transformer for 2-D image segmentation. Encouraged by the progress achieved in medical image research, some researchers also began applying Transformer to remote sensing imagery segmentation. Wang et al. [29] applied ViT for building extraction. Yuan Wei et al. [30] and Xin Chen [31] tried to use Swin Transformer to realize building detection. Qiu et al. [32] tried to use Transformers for cross-domain building detection.

Although both CNN and Transformer have achieved good results in remote sensing image processing, the related research on combining them to realize remote sensing image segmentation is still very limited. Since multihead self-attention (MSA) calculation and convolution calculation exhibit opposite behaviors, indicating that MSA aggregate feature maps but convolution diversify them [33], a hybrid model that combines CNN and Transformer can theoretically achieve better results. Especially for large urban area with high building densities, only using CNN-based feature extractor may cause the extracted features into locally optimal, which may lead to false detections around small buildings and building boundaries. On the other hand, directly deploying the Transformer-based feature extractor into large-size remote sensing images require very large computational resources. Therefore, we propose a model that contains the advantages of both CNN and Transformer in this article. The first part of the encoder in this model is a CNN module with residual connections [34], which gives it the ability to extract multiscale features and local dependencies. After that, a Transformer module containing several MSA layers is added to increase the capacity of extracting global relationships. The decoder comprises CNN modules to help the model fuse multiscale features and resize the prediction to the original image size, achieving end-to-end image segmentation. However, considering that building targets in remote sensing images usually have more regular geometric contours, additional constraints should be added to enhance the model's ability to extract precise boundary contours. Wu et al. [35] proposed the MCFCN model for building detection and change detection [36], which came up with a multiscale constraint mechanism and added it to the architecture of the traditional FCN model to improve its performance in remote sensing image segmentation and outperform the state-of-the-art U-net model at that time. Encouraged by this success, we tried to insert a multiscale constraint mechanism into our CNN-Transformer-based model and proposed MC-TRANSU. We applied this model to three different aerial imagery datasets to do semantic segmentation and examined its performance. Experimental results show that our proposed model outperforms other models on all the datasets, achieving excellent results of 0.8309, 0.7212, 0.7593, and 0.8942 on F1 score, intersection over union (IoU), Kappa coefficient, and accuracy, respectively, on the Tokyo dataset, which covered the central part of Tokyo city with all kinds of high-density buildings. Furthermore, it shows that MC-TRANSU has a reliable performance in aerial image segmentation, indicating that the multiscale constraint mechanism is also effective in the CNN and Transformer hybrid-architecture model. The main contributions of this article are listed as follows.

1) We propose an MC-TRANSU model for automatic building extraction from very high-resolution remote sensing images, which outperforms the commonly used models, such as U-net, TRANSUNET, and SwinUnet in various metrics.
2) We combine the CNN's ability to extract locality dependencies with Transformer's characteristic of weak inductive bias and capture of long-range dependencies, which can make our model consider both local information and global context information better.
3) We introduce the multiconstraint attention mechanism to better take into account the internal relations between multiscale features and improve the defect that the middle layer of the traditional CNN model has a low semantic contribution to the final result, which makes the extraction results more complete and accurate.

The rest of this article is organized as follows. In Section II, the architecture details of our proposed model are described. Section III introduces the specific process of the experiment, showing the comparison of metrics of each model and the comparison of visual effects. Section IV presents the discussion. Finally, Section V concludes this article and provides the content of future research.

II. Methodology

The overall structure of our model is shown in Fig. 1. First the input image will go through the convolutional layers for feature extraction. This operation can get the locality information from the image. Then, the generated multichannel feature map will be fed into several Transformer blocks. After the MSA computation, the implicit global information within the image will be aware. When this step is finished, the result will be combined with the features generated by the encoder part through the skip connection structure. This mechanism has already been proven effective in UNet. Finally, The multiscale constraints mechanism is applied between the output of each decoder block and the corresponding ground truth images.

A. Encoder Part

Nowadays, most semantic segmentation models follow the encoder–decoder structure. For the proposed method, the encoder is used to extract multiscale information from the original

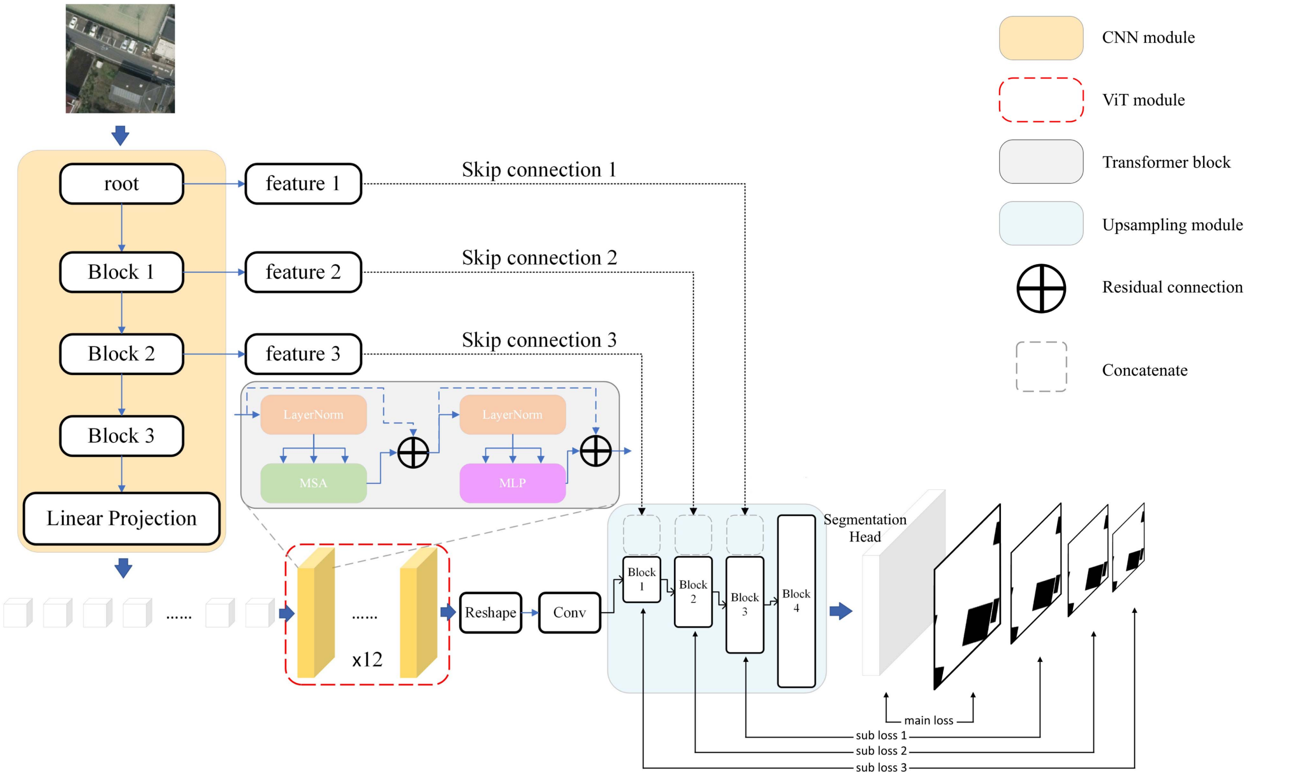


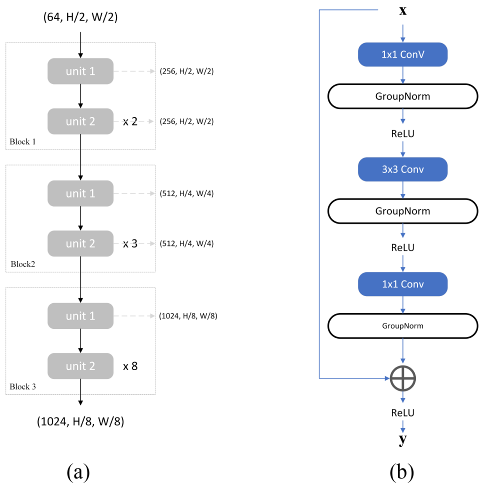
Fig. 1. Architecture of MC-TRANSU.

image, and its performance will directly affect the final prediction result of the entire model. The convolution computation and MSA computation can be seen as opposite behaviors, which can provide complementary information to each other. Based on this, we combine the convolution layers and transformer blocks to form the encoder part of the proposed method. Let $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$be the input of our model, where C, H, and W are the channel dimension, height, and width of the input image, respectively. First, the input data will be processed by a root block, containing a convolution layer that applied weight standardization and a group normalization layer followed by a ReLU activation layer. After that, three encoding blocks are continuously used to extract high-level feature representations with different scales from the result. Each encoding block contains 3, 4, and 9 residual blocks, respectively. The details are shown in Fig. 2.

The high-level feature representation produced by CNN has a shape of $1024 \times H' \times W'$, where $H' = \frac{H}{16}, W' = \frac{W}{16}$. Then, the embedding operation is conducted just like in ViT, the original output is first reshaped to patch sequence $\{\mathbf{x}_p^i \in \mathbb{R}^{P^2 \cdot C'} | \quad i = 1, \ldots, N\}$ then projected to a latent D dimensional space, where P is the size of each patch, $N = \frac{H' \cdot W'}{P^2}$ is the number of patches, and D is the constant latent vector size through all of the Transformer layers later. In this process, position information is also added into patches by conducting patch embedding as follows:

$$\mathbf{z_0} = \left[\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}} \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C') \times D}$ is the linear projection and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times D}$ is the position embedding.



(a) (b)

Fig. 2. Details of residual block. (a) Illustration of Blocks 1, 2, 3, and the shape of features generated. (b) Details of each unit.

Then, a Transformer encoder part that contains 12 Transformer blocks is used to obtain long-distance dependency from the patch embeddings. The output can be represented as follows:

$$\mathbf{z'}_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1} \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell \tag{3}$$

where $\mathbf{MSA}$ represents the MSA computation, $\mathbf{MLP}$ represents the multilayer perceptron (MLP), and $\mathbf{LN}$ represents the layer normalization (LN).

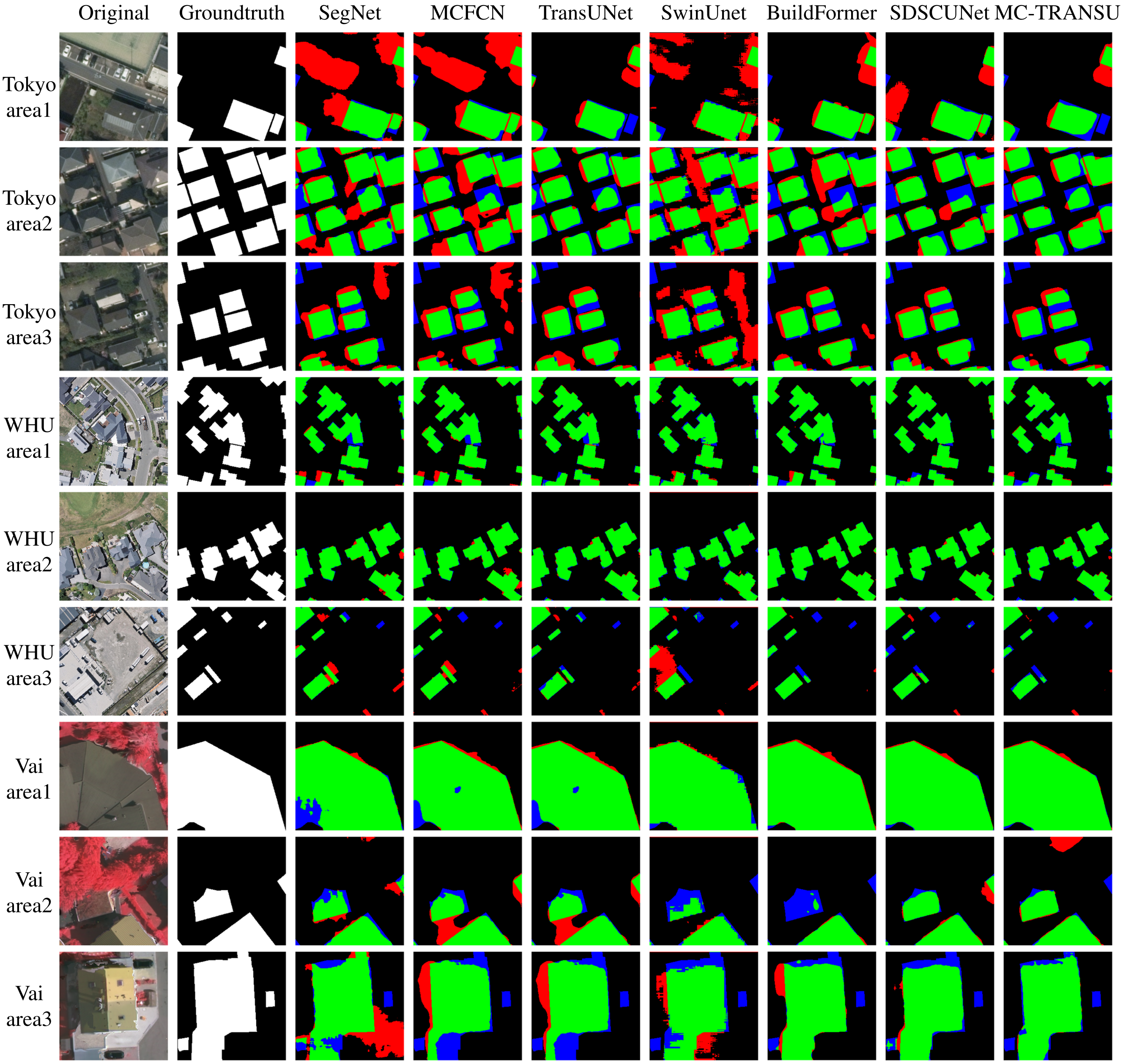


Fig. 3. Segmentation results of different models in small-scale regions. The green, red, blue, and black pixels represent the predictions of true positive, false positive, false negative, and true negative samples, respectively. Tokyo, WHU, and Vai stands for Tokyo dataset, WHU building dataset, and Vaihingen dataset, respectively.

B. Decoder Part

After the computation of Transformer layers, we can get a vector sequence with shape of $\frac{H'W'}{P^2} \times D$, containing both the local and global information of the original image. In order to achieve a pixelwise dense prediction task, this vector sequence should be resized to the original size of the input image. Thus, a decoder structure is needed. In the proposed model, this vector sequence will be reshaped to a 3-D tensor with a shape of $512 \times \frac{H}{16} \times \frac{W}{16}$ and then concatenated with the feature representations generated by the encoder part through a skip-connection structure. The output will be processed by upsampling and repeating the concatenate operation. In addition to the result generated by the last decoder block, which will be processed by a segmentation head and output as the final result, the outputs of the previous decoder blocks are also stored and used for the calculation of the multiscale constraint loss. Each decoder block has an upsampling layer and two convolution layers, followed by a ReLU layer.

C. Multiconstraint Part

In a normal semantic segmentation model, the loss is computed by comparing the predicted results with the ground truth. Then the backpropagation algorithm is applied to update every layer's parameters and achieve the neural network training. But this kind of training strategy is thought to have some inherent drawbacks, given as follows.

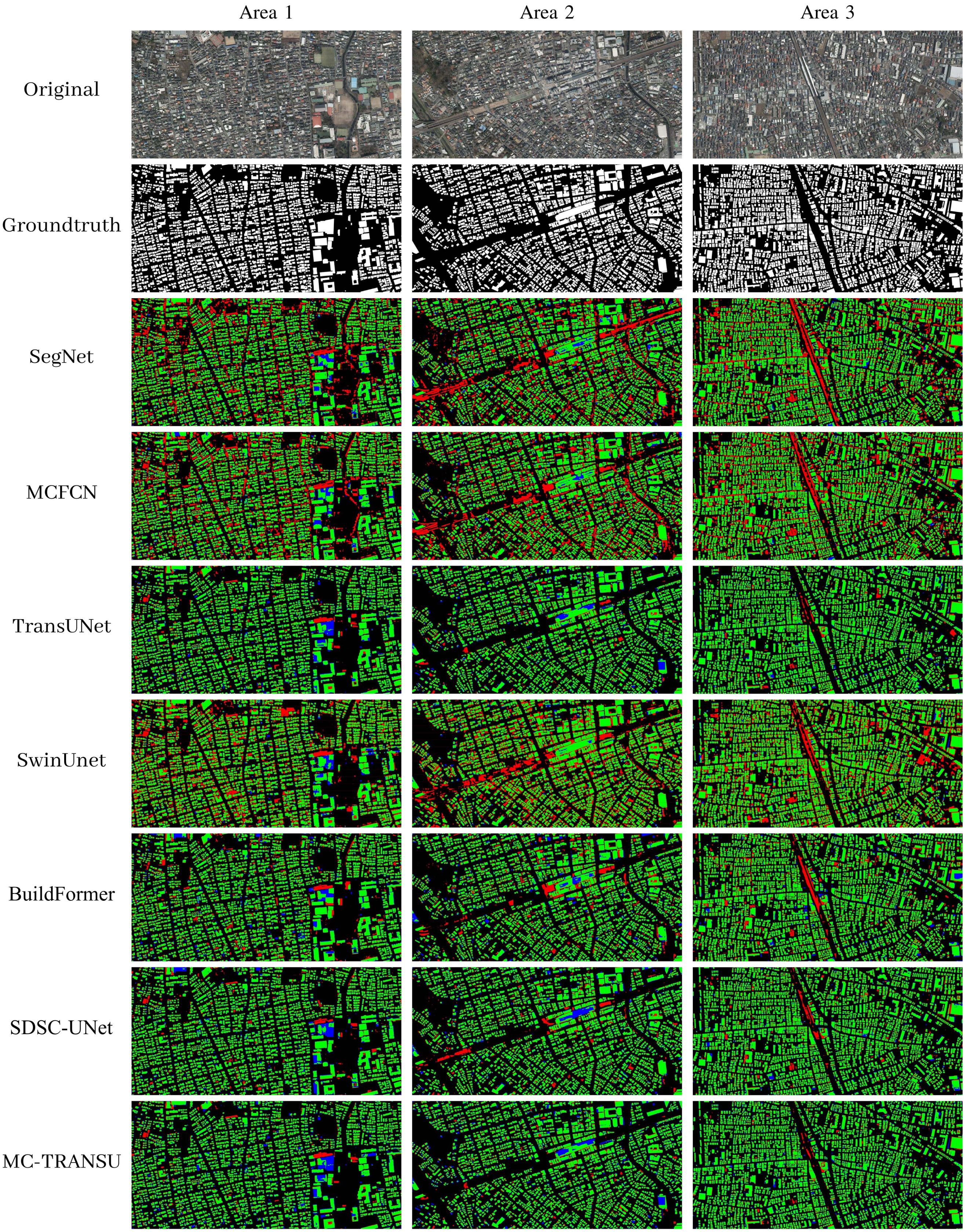


Fig. 4. Segmentation results of different models in large-scale regions. The green, red, blue, and black pixels represent the predictions of true positive, false positive, false negative, and true negative samples, respectively.

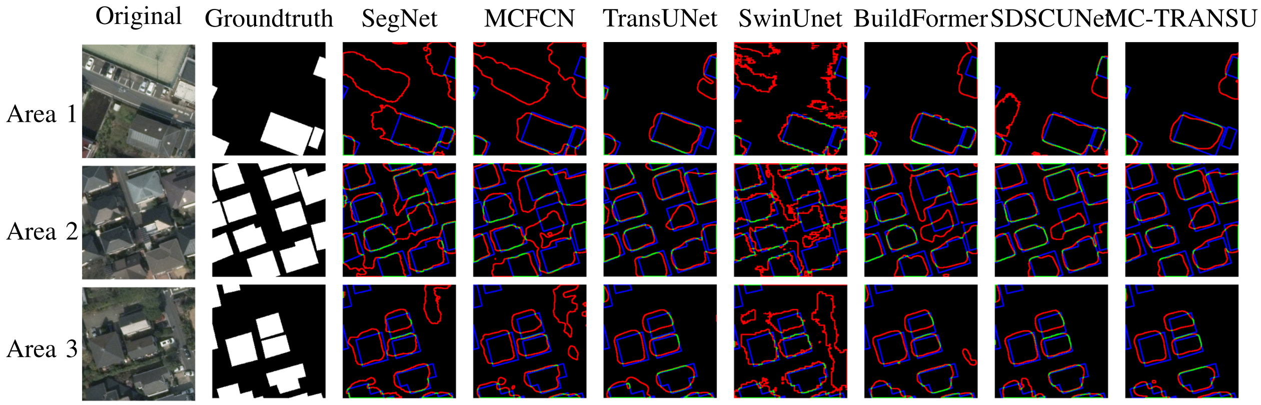


Fig. 5. Boundary extraction results of different models on Tokyo dataset. The green, red, blue, and black pixels represent the predictions of true positive, false positive, false negative, and true negative samples, respectively.

1) Parameters of all the layers are updated only depending on the loss calculated using the result of the last layer. When there are too many layers in the model, the farther the intermediate layer is from the final result, the more its parameter update will be delayed.
2) Backpropagation affects the time it takes to train the model and limits the optimal results the model can achieve. The parameters of all layers are very difficult to reach the best value simultaneously only relying on the output of the last layer and the best performance is restricted as a result.

Assume that a model has a total of ℓ layers without using multiconstraint. The final loss is denoted as C, so the gradient of the last layer can be denoted as $\nabla_{\theta^\ell} C$, where θ_ℓ represent the parameter of layer ℓ. The gradient of the parameters of the penultimate layer can be computed as follows:

$$\nabla_{\theta^{\ell-1}} \mathbf{C} = \left(\frac{\partial \theta^\ell}{\partial \theta^{\ell-1}}\right)^\top \nabla_{\theta^\ell} \mathbf{C} \tag{4}$$

whereas if we introduce the multiconstraint mechanism into the $\ell-1$ layer, the final loss $\mathbf{C} = \alpha \mathbf{C}_\ell + \beta \mathbf{C}_{\ell-\mathbf{1}}$. Now, the gradient of parameters in $\ell-1$ layer can be represented as follows:

$$\begin{aligned}\nabla_{\theta_{\ell-1}} \mathbf{C} &= \alpha \nabla_{\theta^{\ell-1}} \mathbf{C}_\ell + \beta \nabla_{\theta^{\ell-1}} \mathbf{C}_{\ell-1} \\ &= \alpha \left(\frac{\partial \theta^\ell}{\partial \theta^{\ell-1}}\right)^\top \nabla_{\theta^\ell} \mathbf{C}_\ell + \beta \nabla_{\theta^{\ell-1}} \mathbf{C}_{\ell-1}.\end{aligned} \tag{5}$$

We can see that the update of parameters in $\ell-1$ layer can be accelerated by the extra contribution of $\nabla_{\theta^{\ell-1}} \mathbf{C}_{\ell-1}$ and can also converge to the degree closer to the optimal. In the proposed model, we added four multiscale constraints by using the outputs of the first three decoder blocks and the segmentation head to compute the corresponding loss. The final loss is calculated by distributing weight to each loss

$$\mathbf{C} = \alpha \times \mathbf{C}_1 + \beta \times \mathbf{C}_2 + \gamma \times \mathbf{C}_3 + \delta \times \mathbf{C}_4 \tag{6}$$

where $\alpha + \beta + \gamma + \delta = 1$. In our experiments, we used two strategies to choose the weights combination, which will be introduced in Section III.

III. Experiment

A. Dataset Description

To verify the performance of our proposed MC-TRANSU model in the task of building segmentation based on aerial images, we conduct experiments using several aerial image datasets, including Tokyo dataset, Vaihingen dataset, and WHU Building dataset.

1) *Tokyo Dataset:* This dataset contains aerial images covering three districts of Tokyo with a ground sampling distance (GSD) of 16 cm. The original size of each image is 12 500 × 9375 pixels. We manually labeled four images and used two images for training, one for validating the optimal weights combination of MC-TRANSU, and one for testing. In order to save memory of GPU, the original images were cropped into patches with size of 256 × 256 pixels and input to model for training/testing.
2) *Vaihingen Dataset:* This dataset contains 33 orthophoto aerial images with different size that were captured from Vaihingen in Germany. The GSD is 9 cm. From the 16 images with ground truth label we selected 13 images for training, two for validating, and one for testing. They were also cropped to 256 × 256 patches for inputting.
3) *WHU Building Dataset:* This dataset contains original aerial data comes from the New Zealand Land Information Services website with an original GSD of 7.5 cm. These original aerial data were downsampled by WHU researchers to 0.3 m GSD and cropped into 8189 tiles with 512 × 512 pixels. We cropped these tiles into smaller patches with the same 256 × 256 size as the other two datasets. We used 18 944 patches for training, 4144 for validating, and 9664 for testing.

B. Comparison Baselines and Metrics

To evaluate the performance level of our model, we introduce several other classical and commonly used semantic segmentation models here for comparison, including both pure CNN networks and some hybrid CNN and Transformer models. Models we used are listed as follows.

1) *UNet:* One of the most classical end-to-end semantic segmentation networks, proposed the subsequent most used U-shaped structure and skip connection.
2) *ResUnet:* A deep neural network constructed based on UNet and the residual unit put forward by He, which goes beyond vanilla UNet by a large margin.
3) *SegNet:* A semantic segmentation model with encoder–decoder architecture uses poolingindices to perform upsampling instead of deconvolution, which helps to eliminate the number of parameters and achieve tradeoff in memory and accuracy.
4) *MC-FCN:* The first model that proposed multiscale constraint and applied it to FCN for aerial image segmentation.
5) *TransUNet:* A hybrid model that inserts Transformer into UNet and strengthens its ability to model long-range dependency.
6) *SwinUnet:* The first pure Transformer-based U-shaped model, which uses Swin Transformer to extract local-global semantic features.
7) *SDSCUnet [37]*: The recent U-net-based building detection model, which outperforms other pure CNN-based models.
8) *BuildFormer [29]*: The state-of-the-art transformer-based building detection model, which outperforms other transformer-based models.

In order to facilitate comparison and to realistically evaluate the performance of our proposed model, we use the evaluation metrics shown below.

1) *Precision:* This index shows what proportion of positive identification was actually correct.
2) *Recall:* This index represents what proportion of actual positives was identified correctly.
3) *Accuracy:* This index shows the fraction of predictions our model got right.
4) *F1 score:* Since both of precision and recall cannot fully describe the prediction performance of a model, the F1 score is proposed to combine precision and recall by calculating their harmonic average.
5) *IoU:* This measurement is specially designed to check the accuracy of object detection and segmentation, which is calculated by dividing the number of pixels in the overlapping area by the number of pixels in the union area.
6) *Hausdorff Distance (HD):* A index that usually used to measure the distance between two point sets, which is defined as follows:

$$h(\mathbf{A},\mathbf{B}) = \max_{a\in A}\left\{\min_{b\in B} \parallel a-b \parallel\right\} \tag{7}$$

$$H(\mathbf{A},\mathbf{B}) = \mathbf{max}(h(\mathbf{A},\mathbf{B}), h(\mathbf{B},\mathbf{A})) \tag{8}$$

where A and B are two point sets, $\parallel a-b \parallel$ can represent any metric between these points, and we usually use Euclidean distance for simplicity.
7) *Average Symmetric Surface Distance (ASSD):* Average surface distance (ASD) is an unidirectional metric that used to compute the minimal distance for every point from one object to the other, and ASSD is the arithmetic mean of two ASDs. It is defined as follows:

$$\mathbf{ASD}(\mathbf{A},\mathbf{B}) = \sum_{a\in A} \min_{b\in B} \frac{\parallel a-b \parallel}{|A|} \tag{9}$$

$$\mathbf{ASSD}(\mathbf{A},\mathbf{B}) = \frac{\{\mathbf{ASD}(A,B)+\mathbf{ASD}(B,A)\}}{2}. \tag{10}$$

8) *Kappa:* This is a statistic that often used for interrater reliability evaluation and can also be used to measure classification accuracy.

The environment is Python 3.7 and Pytorch 1.7.0 with CUDA 10.1. We use the stochastic gradient descent optimization algorithm with 0.9 momentum and a weight decay of 1e-4 in this experiment to train our model. The base learning rate is uniformly set to 0.005. All of the experiments were conducted on a NVIDIA GeForce GTX 1080Ti 11-GB GPU. The batch size is set to 16. In order to keep consistent with TransUNet, the skeleton of this model, the strategy of gradually decreasing the learning rate is also adopted in this experiment as follows:

$$\mathbf{lr}_\ell = \mathbf{lr}_{\text{base}} \times \left(\frac{\ell}{\mathbf{M}}\right)^{0.9} \tag{11}$$

where the ℓ is the number of rounds of the current iteration, and $\mathbf{M}$ is the maximum number of iterations.

We also adopt a pretraining strategy similar to TransUNet to give full play to the performance of Transformer. Pretraining parameters are used in the encoder part of the model, ResNet-50 in the CNN part, and ViT-B in the Transformer. Both are pretrained on the Imagenet21 K dataset.

C. Weights Combination Selection Strategy

Since different weights combination can have a significant effect on the performance of MC-TRANSU, in our experiment we used two strategies to select weight combination, including random searching and dynamic minimizing.

1) *Random Searching:* Random searching is a strategy that is often used in the hyperparameter optimization of machine learning models. Compared with grid searching, random searching sacrificed a little precision but saved a lot of time. In our experiment, we validated randomly selected weights combination on an aerial image chosen from Tokyo dataset. The results are shown in Table I. So in the formal experiment on Tokyo dataset, the weights combination of MC-TRANSU is set to be $\alpha = 0.4, \beta = 0.2, \gamma = 0.2,$ and $\sigma = 0.2$.
2) *Dynamic Minimizing:* Although random searching strategy improves some efficiency, it is still time-consuming when validating different weights combinations on a very large dataset. So we used a dynamic minimizing method in our experiments on WHU building dataset and Vaihingen dataset. After getting four losses in each training iteration, we traverse the value of $\alpha, \beta, \gamma,$ and σ from 0.1 to 1 at intervals of 0.1 and keep their sum equal to 1. The combination that can minimize the final loss will be chosen.

TABLE I
COMPARISON OF DIFFERENT WEIGHTS COMBINATION OF MC-TRANSU

	F1	IoU	HD	ASSD
$\alpha = 0.5, \beta = 0.5$	0.7913	0.6767	57.9508	8.4451
$\alpha = 1.0$	0.7980	0.6808	56.8231	8.6470
$\alpha = 0.5, \beta = 0.3, \gamma = 0.1, \sigma = 0.1$	0.7981	0.6821	56.2202	7.5037
$\alpha = 0.7, \beta = 0.1, \gamma = 0.1, \sigma = 0.1$	0.8001	0.6849	55.9500	7.4461
$\alpha = 0.4, \beta = 0.2, \gamma = 0.2, \sigma = 0.2$	**0.8017**	**0.6856**	**55.0517**	**7.2887**
$\alpha = 0.1, \beta = 0.3, \gamma = 0.3, \sigma = 0.3$	0.7945	0.6779	57.3361	7.7233
$\alpha = 0.5, \gamma = 0.5$	0.8006	0.6844	56.3318	7.7656
$\alpha = 0.5, \sigma = 0.5$	0.7943	0.6771	58.0913	8.7194
$\alpha = 0.2, \beta = 0.2, \gamma = 0.3, \sigma = 0.3$	0.7960	0.6780	56.7535	7.7032
$\alpha = 0.5, \beta = 0.1, \gamma = 0.2, \sigma = 0.2$	0.7924	0.6757	58.3634	7.8230
$\alpha = 0.6, \gamma = 0.2, \sigma = 0.2$	0.7938	0.6760	57.3871	8.2468
$\alpha = 0.3, \gamma = 0.3, \gamma = 0.2, \sigma = 0.2$	0.7932	0.6765	57.6476	7.9561
$\alpha = 0.4, \beta = 0.2, \gamma = 0.1, \sigma = 0.3$	0.7930	0.6755	57.9945	7.7480
$\alpha = 0.4, \beta = 0.2, \gamma = 0.3, \sigma = 0.1$	0.7938	0.6765	57.4798	7.8477

The bold entities means the best index value between the comparison methods

D. Results Analysis

The evaluation indexes of each baseline model and our proposed model are calculated and shown in Table II. On the Tokyo dataset and Vaihingen dataset, we can see that our proposed MC-TRANSU has reached the optimal level in almost all indexes except recall. Compared with the stat-of-the-art CNN and transformer-based methods, our proposed methods show better results on datasets containing different types of buildings, buildformer shows better results when the building roof is more regular, such as the WHU dataset. Although the original TransUNet had already achieved quite good prediction results, our model improved on this basis, which demonstrated the effectiveness of the multiconstraint mechanism. On the WHU Building dataset, almost all of the models have achieved excellent performance due to the large amount of data. Our MC-TRANSU reached the best level in five indexes among eight, which means it is still the best-performing model compared with others.

To further explore the advantages of our proposed model compared with other prediction models and visually compare the differences, we compared their segmentation results with label files in different scale experimental areas. Fig. 3 shows the segmentation results of the model in patch size. It can be seen from the figure that all models can extract the main body of the building in the target area, but models based on pure CNN have a significantly high false positive rate. For example, in Area 1, all of the pure-CNN models identified the playground in the upper left corner as a building, and the state-of-the-art methods, such as SDSCUnet and buildformer are shown false positive around the small building, and in Area 3 they also misclassified the road in the upper-right corner as a building. However, TransUnet, SDSCUnet, and MC-TRANSU, which are based on the hybrid architecture of CNN and Transformer successfully distinguished these parts correctly. SwinUnet and BuildFormer also misidentified these parts and we thought the reason is the Tokyo dataset is too small for SwinUnet to converge. Meanwhile, we can also find that the prediction results of MC-TRANSU on Vaihingen are obviously better than others, meaning our model can perform better on a small dataset with less ground truth data.

To evaluate the overall performance of each model within a large-scale range, we spliced 512 output patches on the Tokyo dataset and got a large-scale image with a size of 8192×4096. Fig. 4 shows the visualization results. We can see that the MC-TRANSU model we proposed has achieved significantly better results than other models in a large-scale region when performing building extraction, with a high true positive rate and a low false positive rate. Most of the main parts of the buildings are identified accurately. Although the stat-of-the-art methods also basically achieved this, MC-TRANSU still reduced the false positive rate on this basis and obtained more accurate results.

In addition to evaluating the segmentation accuracy from the number of pixels or the area, the contour accuracy of the segmentation results is also very important. Therefore, we extract the contour of the segmentation results of each model on the Tokyo dataset and calculate its accuracy. From Table III, our MC-TRANSU model still achieves the highest accuracy in contour recognition. Fig. 5 shows the visual effect of contour extraction. Compared with the messy contours proposed by other models, the contour extracted by MC-TRANSU is more regular and closer to the actual boundary of the building itself. On the other hand, we have calculated the parameter and flops of each method shown in Table III. Our proposed methods have the same parameters and FLOPs as TransUnet, but we have the best performance among all the compared methods. This is due to the fact that we have used the structure of U-net combined with transformer, which makes the network parameters larger, but because of this feature extraction structure, the extracted features are more effective. The detailed feature representation map is shown in Fig. 6.

IV. DISCUSSION

This article proposes a combined CNN and transformer hybrid neural network for precision building detection from high-resolution remote sensing images. Through the comprehensive comparison experiments on three different-sized datasets with different building types. We demonstrate that the proposed hybrid CNN and transformer-based feature extraction model can generate the most accurate building detection results. It is surprising to find that the traditional CNN-based model can achieve almost the same building detection accuracy as the recent ViT-based methods. But the hybrid CNN and ViT combined methods have greatly improved the quality of the result, such as SDSCUnet and our proposed MC-TRANSU. The building outline extraction experiments partially support the conclusion of our work that combining utilizing CNN and Transformer together may get much more complete detection results in one object's structures and boundaries. On the other hand, the multiconstraint mechanism can further improve the detailed feature extraction on small objects and around the object boundaries.

We chose some test image patches from the Vaihingen dataset and visualized the 128×128 medium feature maps of CNN, TransUnet, and our MC-TRANSU for comparison.

TABLE II
COMPARISON OF MC-TRANSU AND BASELINE MODEL

Dataset	Model	F1	IoU	HD	ASSD	Kappa	Precision	Recall	Accuracy
Tokyo dataset	**SegNet**	0.7336	0.5963	75.5729	10.3254	0.5666	0.6611	0.8631	0.8011
	MCFCN	0.7366	0.5989	72.5155	9.8506	0.5649	0.6571	0.8739	0.7997
	TransUNet	0.8287	0.7188	51.5795	5.9489	0.7335	0.8370	0.8383	0.8881
	SwinUnet	0.7483	0.6123	74.6395	10.2193	0.5729	0.6648	**0.8941**	0.8054
	BuildFormer	0.8068	0.6890	42.7681	4.9973	0.6937	0.7923	0.8444	0.8683
	SDSCUNet	0.8270	0.7179	**37.3479**	**4.1291**	0.7331	0.8323	0.8403	0.8878
	MC-TRANSU	**0.8309**	**0.7212**	49.8731	5.7611	**0.7593**	**0.8676**	0.8452	**0.8942**
Vaihingen dataset	**SegNet**	0.8250	0.7312	42.6783	3.7490	0.7746	0.8401	0.8516	0.9335
	MCFCN	0.8410	0.7627	36.7675	2.8377	0.7993	0.8630	0.8577	0.9439
	TransUNet	0.8857	0.8176	25.8234	1.9097	0.8580	0.8951	0.9130	0.9627
	SwinUnet	0.7333	0.6372	56.3805	4.9894	0.6782	0.8089	0.7492	0.9128
	BuildFormer	**0.8937**	**0.8288**	**23.3868**	1.8047	0.8646	0.9095	0.9112	0.9614
	SDSCUNet	0.8907	0.8189	58.8126	2.2411	0.8590	**0.9306**	0.8816	0.9566
	MC-TRANSU	0.8918	0.8248	25.4608	**1.7975**	**0.8652**	0.9022	**0.9143**	**0.9639**
WHU dataset	**SegNet**	0.9172	0.8610	22.7258	1.0574	0.9012	0.9310	0.9180	0.9769
	MCFCN	0.9171	0.8601	22.0424	1.0166	0.9019	0.9361	0.9103	0.9778
	TransUNet	0.9170	0.8601	21.1159	0.9153	0.9018	0.9357	0.9124	0.9779
	SwinUnet	0.8528	0.7681	28.7306	1.7332	0.8284	0.8785	0.8433	0.9639
	BuildFormer	**0.9325**	**0.8855**	**17.7492**	**0.7356**	**0.9204**	**0.9519**	**0.9271**	**0.9823**
	SDSCUNet	0.9316	0.8841	18.2484	0.7560	0.9193	0.9497	0.9268	0.9819
	MC-TRANSU	0.9271	0.8723	20.5912	0.9082	0.9065	0.9421	0.9215	0.9801

The bold entities means the best index value between the comparison methods

TABLE III
COMPARISON OF OUTLINE EXTRACTION ACCURACY ON TOKYO DATASET

	Accuracy	FLOPs	Param
SegNet	0.9493	10.77 G	5.63 M
MCFCN	0.9506	5.35 G	3.41 M
TransUNet	0.9580	38.55 G	105.32 M
SwinUnet	0.9358	8.09 G	27.18 M
BuildFormer	0.9558	29.28 G	40.52 M
SDSCUNet	0.9568	5.89 G	21.31 M
MC-TRANSU	**0.9587**	38.55 G	105.32 M

The bold entities means the best index value between the comparison methods.

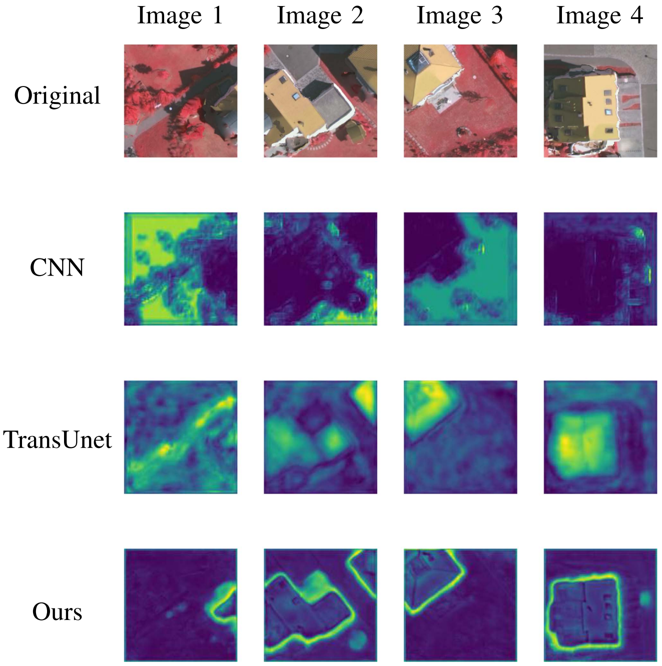


Fig. 6. Visualization results of the 128 × 128 size feature maps generated by CNN, TransUnet, and our MC-TRANSU, respectively.

Fig. 6 shows that objects and backgrounds are more distinct in feature maps output by a hybrid model like TransUnet and MC-TRANSU compared with feature maps output by pure CNN model. The results also demonstrated the significance and completeness of the outlines in the feature maps output by our MC-TRANSU compared with TransUnet, which explained why our model performed better on outline precision.

One of the limitations of this work is data selection and data distribution. Although high-accurate data samples covering Tokyo city and Vahingen City are some parts of New Zealand utilized for training and validation, more data around the world would be better for diversity. When more open-source and high-accurate datasets are available, we will further testify to the performance of the proposed hybrid feature extraction module and multiconstraints mechanism.

V. CONCLUSION

In this article, we propose a novel MC-TRANSU model architecture based on improving the efficiency of parameter backpropagation in a deep neural network. By introducing the multiscale constraint mechanism into the hybrid model of CNN and Transformer and designing two strategies to obtain the optimal combination of weight parameters, we applied the model to a building extraction task based on an aerial imagery dataset of the Tokyo area and achieved good segmentation results, achieving mean values of F1-score, IoU, and Kappa coefficient at 0.8309, 0.7212, and 0.7593, respectively. Its performance exceeds many classic pure CNN models and the current mainstream CNN and Transformer hybrid models, which indicates that the multiscale constraint mechanism is effective.

In future research, more relevant experiments will be conducted. By adding the proposed mechanism to different segmentation and detection models and applying it to other data types besides aerial images, the effectiveness and robustness of the proposed mechanism can be fully verified. In addition, the theoretical explanation of the specific action mechanism of multiscale constraints is worth researching. We believe that the multiscale constraint mechanism can further improve the accuracy of existing segmentation networks.

REFERENCES

[1] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1873–1876.

[2] C. Piciarelli, C. Micheloni, N. Martinel, M. Vernier, and G. L. Foresti, "Outdoor environment monitoring with unmanned aerial vehicles," in *Image Analysis and Processing*, A. Petrosino, ed. Berlin, Germany: Springer, 2013, pp. 279–287.

[3] Z. Lv, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 44–63, Mar. 2023.

[4] H. Mayer, "Automatic object extraction from aerial imagery–A survey focusing on buildings," *Comput. Vis. Image Understanding*, vol. 74, no. 2, pp. 138–149, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314299907506

[5] B. Guindon, "Computer-based aerial image understanding: A review and assessment of its application to planimetric information extraction from very high resolution satellite images," *Can. J. Remote Sens.*, vol. 23, no. 1, pp. 38–47, 1997. [Online]. Available: https://doi.org/10.1080/07038992.1997.10874676

[6] C. Lin and R. Nevatia, "Building detection and description from a single intensity image," *Comput. Vis. Image Understanding*, vol. 72, no. 2, pp. 101–121, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731429890724X

[7] Y. Zhang, "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering," *ISPRS J. Photogrammetry Remote Sens.*, vol. 54, no. 1, pp. 50–60, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271698000276

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

[9] Y. Meng and S. Peng, "Object-oriented building extraction from high-resolution imagery based on fuzzy SVM," in *Proc. Int. Conf. Inf. Eng. Comput. Sci.*, 2009, pp. 1–6.

[10] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS J. Photogrammetry Remote Sens.*, vol. 62, no. 3, pp. 236–248, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427160700055X

[11] S. Du, F. Zhang, and X. Zhang, "Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach," *ISPRS J. Photogrammetry Remote Sens.*, vol. 105, pp. 107–119, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427161500091X

[12] R. Hänsch and O. Hellwich, "Random forests for building detection in polarimetric SAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 460–463.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

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

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

[16] 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.

[17] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[19] D. He, Q. Shi, X. Liu, Y. Zhong, and X. Zhang, "Deep subpixel mapping based on semantic information modulated network for urban land use mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10628–10646, Dec. 2021.

[20] D. He, Q. Shi, X. Liu, Y. Zhong, G. Xia, and L. Zhang, "Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using landsat imagery," *GIScience Remote Sens.*, vol. 59, no. 1, pp. 2036–2067, 2022.

[21] T. Zuo, J. Feng, and X. Chen, "HF-FCN: Hierarchically fused fully convolutional network for robust building extraction," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 291–302.

[22] S. Shrestha and L. Vanneschi, "Improved fully convolutional network with conditional random fields for building extraction," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1135. [Online]. Available: https://www.mdpi.com/2072-4292/10/7/1135

[23] A. Abdollahi and B. Pradhan, "Integrating semantic edges and segmentation information for building extraction from aerial images using unet," *Mach. Learn. Appl.*, vol. 6, 2021, Art. no. 100194. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000979

[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.

[26] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[27] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[28] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 205–218.

[29] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[30] W. Yuan and W. Xu, "MSST-Net: A multi-scale adaptive network for building extraction from remote sensing images based on swin transformer," *Remote Sens.*, vol. 13, no. 23, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/23/4743

[31] X. Chen, C. Qiu, W. Guo, A. Yu, X. Tong, and M. Schmitt, "Multiscale feature learning by transformer for building extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2503605.

[32] C. Qiu et al., "Transferring transformer-based models for cross-area building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4104–4116, 2022.

[33] N. Park and S. Kim, "How do vision transformers work?," in *Proc. Int. Conf. Learn. Representations*, 2022.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[35] G. Wu et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 407. [Online]. Available: https://www.mdpi.com/2072-4292/10/3/407

[36] W. Yuan et al., "Graph neural network based multi-feature fusion for building change detection," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 377–382, 2021.

[37] R. Zhang, Q. Zhang, and G. Zhang, "SDSC-UNet: Dual skip connection ViT-based U-shaped model for building extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6005005.

Wei Yuan (Member, IEEE) received the M.E. and Ph.D. degrees in civil engineering from The University of Tokyo, Tokyo, Japan, in 2015 and 2018, respectively, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2020.

He joined as a Researcher with the Center for Spatial Information Science, Tokyo, in 2018, where he became an Assistant Professor in 2021. His research interests include photogrammetry and remote sensing, GIS, and computer vision, especially in image matching, 3-D reconstruction, and change detection.

Weihang Ran (Student Member, IEEE) received the bachelor's degree in architecture from Wuhan University, Wuhan, China, and the master's degree in social cultural environmental studies from The University of Tokyo, Tokyo, Japan, where he is currently working toward the Ph.D. degree in mechano informatics.

His research interests include computer vision, remote sensing, and photogrammetry, especially in image analysis, object detection, and scene understanding.

Xiaodan Shi (Member, IEEE) received the B.E. and M.S. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, and the Ph.D. degree in social cultural environmental studies from the Center for Spatial Information Science, The University of Tokyo, Kashiwa, Japan.

She is currently a Postdoctoral Researcher with Future Energy Center, Mälardalen University, Västerås, Sweden, and a Visiting Researcher with the Center for Spatial Information Science, The University of Tokyo. Her current research interests include computer vision, time series prediction, deep learning and its applications in pedestrian trajectory prediction, human flow monitoring, data mining in GPS, smart city, and smart building.

Ryosuke Shibasaki (Member, IEEE) is currently a Professor with the Center for Spatial Information Science, The University of Tokyo, Tokyo, Japan. His research interests include the integration of data and models based on GIS to reconstruct spatial temporal dynamics of objects, microsimulation modeling, human behavior understanding and modeling, the analysis of mobile phone data, and urban informatics and their applications.

Prof. Shibasaki was the President of the Asian GIS Association and the GIS Association of Japan, a Board Member of the Japanese Society of Photogrammetry and Remote Sensing, and a Member of the Scientific Committee of World Data System, the International Council of Scientific Union, and the Space Strategic Policy Committee of Japanese Government (Cabinet Office of Prime Minister).