PCL–PTD Net: Parallel Cross-Learning-Based Pixel Transferred Deconvolutional Network for Building Extraction in Dense Building Areas With Shadow

Wuttichai Boonpook, Yumin Tan, Kritanai Torsri, Patcharin Kamsing, Peerapong Torteeka, and Attawut Nardkulpat

***Abstract*—Urban building segmentation from remote sensed imageries is challenging because there usually exists a variety of building features. Furthermore, very high spatial resolution imagery can provide many details of the urban building, such as styles, small gaps among buildings, building shadows, etc. Hence, satisfactory accuracy in detecting and extracting urban features from highly detailed images still remains. Deep learning semantic segmentation using baseline networks works well on building extraction; however, their ability in building extraction in shadows area, unclear building feature, and narrow gaps among buildings in dense building zone is still limited. In this article, we propose parallel cross-learning-based pixel transferred deconvolutional network (PCL–PTD net), and then is used to segment urban buildings from aerial photographs. The proposed method is evaluated and intercompared with traditional baseline networks. In PCL–PTD net, it is composed of parallel network, cross-learning functions, residual unit in encoder part, and PTD in decoder part. The performance is applied to three datasets (Inria aerial dataset, international society for photogrammetry and remote sensing Potsdam dataset, and UAV building dataset), to evaluate its accuracy and robustness. As a result, we found that PCL–PTD net can improve learning capacities of the supervised learning model in differentiating buildings in dense area and extracting buildings covered by shadows. As compared to the baseline networks, we found that proposed network shows superior performance compared to all eight networks (SegNet, U-net, pyramid scene parsing network, PixelDCL, DeeplabV3+, U-Net++, context feature enhancement networ, and improved ResU-Net). The experiments on three datasets also demonstrate the ability of proposed framework and indicating its performance.**

***Index Terms*—Building extraction, building shadow, dense building, PCL–PTD net, semantic segmentation.**

Manuscript received 28 April 2022; revised 27 August 2022 and 25 November 2022; accepted 13 December 2022. Date of publication 19 December 2022; date of current version 2 January 2023. This work was supported in part by the Chinese National Key R&D Program under Grant 2019YFE0126400, in part by the third-party funds from the Wildfire Risk Area Project under Grant JRA-CO-2564-15158-TH, in part by the Thai Space Consortium Pathfinder Satellite under Grant B05F630115, and in part by the King Mongkut's Institute of Technology Ladkrabang under Grant KREF046603. *(Corresponding author: Yumin Tan.)*

Wuttichai Boonpook and Attawut Nardkulpat are with the Department of Geography, Faculty of Social Sciences, Srinakharinwirot University, Bangkok 10110, Thailand (e-mail: wuttichaib@g.swu.ac.th; attawut.nardkulpat@g.swu.ac.th).

Yumin Tan is with the School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: tanym@buaa.edu.cn).

Kritanai Torsri is with the Ministry of Higher Education, Science, Research and Innovation, Hydro-Informatics Institute, Bangkok 10900, Thailand (e-mail: kritanai@hii.or.th).

Patcharin Kamsing is with the Air-Space Control, Optimization and Management Laboratory, Department of Aeronautical Engineering, International Academy of Aviation Industry, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand (e-mail: patcharin.ka@kmitl.ac.th).

Peerapong Torteeka is with the National Astronomical Research Institute of Thailand, Chiangmai 50180, Thailand (e-mail: peerapong@narit.or.th).

Digital Object Identifier 10.1109/JSTARS.2022.3230149

I. Introduction

ACCURATE and up-to-date building information is essential for urban analysis and management [1], and it can be obtained from pixel-based classification [2] or semantic segmentation [3] in remote sensing images. Typically, pixel-based classifying with low to medium spatial resolution of imagery is widely used and can provide reasonable results [4]. Deep-learning (DL)-based semantic segmentation methods could extract buildings by learning object features and patterns in high spatial resolution imagery [5]. However, there are still some challenging problems in extracting urban buildings from high spatial resolution imagery where there are many details of features, such as, tall buildings, and narrow gaps and shadows between buildings making building boundary unclear, leading to unsatisfactory accuracy of building extraction.

In recent years, numerous studies have used DL techniques and semantic segmentation has become a popular method [6]. Many studies have demonstrated that the DL could yield highly accurate segmentation [7]. The main part of a DL algorithm is network architecture embedded in the system that functions in cultivating features and patterns of objects by convolutional methods and learns multidimensional data by pooling methods [8]. Various novel functions had been proposed to improve the learning abilities, for example, astrous convolution [9], residual learning [10], bottleneck module [11], PixelDCL function [12], attention refinement module [13], Shufflenet unit [14], channel and position attention module [15], etc. These functions are used as parts in many networks for semantic segmentation, such as fully convolutional network [16], DeconvNet [17], U-Net [18], SegNet [19], pyramid scene parsing network (PSPnet) [20], Deeplab [9], and so on. Some of these networks are proposed to accurately extract buildings from remote sensing images. The complexity of building shapes as well as the variety of building features are hot topics in DL-based semantic segmentation. Luo et al. [21] presented a comprehensive review on DL-based building extraction from remote sensing images. In addition, Lin et al. [22] proposed ESFNet to reduce the computational cost

and memory consumption. Ding et al. [23] presented a spatial pyramid pooling module based on the LinkNet architecture to learn various building features. Chen et al. [24] evaluated the Res2-Unet to enhance its efficiency of small building extraction and confusing background objects. Lei et al. [25] performed Selective Nonlocal ResUNeXt++ (SNLRUX++) to increase the performance of building extraction tasks on high-resolution remote sensing images. Kang et. [26] proposed a novel network (PiCoCo) which comprises pulling intraclass and separating interclass representations in latent space, and imposing the prediction consistency of the model in different augmented unlabeled data for semi-supervised learning building segments with limited data annotations. Wang and Miao [27] demonstrated residual U-Net (RU-Net) architecture to extract the building. The network comprises U-Net architecture, residual learning, and atrous spatial pyramid pooling. Its performance overcomes the sharp, boundary, and multiscale information of the building on remote sensing imagery. Later, Sheikh et al. [28] presented improved ResU-Net (IRU-Net) architecture which integrating spatial pyramid pooling module, atrous convolution, residual connection, and skip connection for building extraction. Chen et al. [29] proposed the context feature enhancement network (CFENet) that comprises the spatial fusion module, the focus enhancement module, and the feature decoder module to overcome the complexity and diversity of buildings. For building boundary extraction, Wu et al. [30] introduced a BR-Net to overcome errors from roof segmentation and outline extraction. Yang et al. [31] demonstrated an end-to-end edge-aware network (EANet) to extract building boundary. For boundary constraint, Wei et al. [32] investigated an automatic building footprint extraction method, and Liu et al. [33] developed a trainable chain fully convolutional neural network in order to fuse ortho images and digital surface model (DSM) in building extraction. The above proposed architectures not only present high accuracy in building segmentation, but also reduce computational parameters and increase learning capacity.

However, building extraction in highly dense urban areas with heavy shadows caused by tall buildings and complex building features is still a challenge for semantic segmentation. Thus, the objective of this article focuses on the building extraction in shadows area, unclear building feature, and narrow gaps among buildings. The main contribution is to propose an adjustment network architecture as called PCL–PTD net, which comprises a parallel network, cross learning, residual unit, and pixel transferred deconvolution, to increase learning features capability over dense urban zone. The performance of proposed the PCL–PTD network is further designed to evaluate on remote sensing datasets [Inria aerial dataset, international society for photogrammetry and remote sensing (ISPRS) Potsdam dataset, and UAV building dataset] and intercompared with several traditional baseline network architectures and adjustment network architectures. After the introduction, methodology is described in Section II, followed by experiment designs and analysis in Section III. Then, discussion and conclusion are presented in Sections IV and V, respectively.

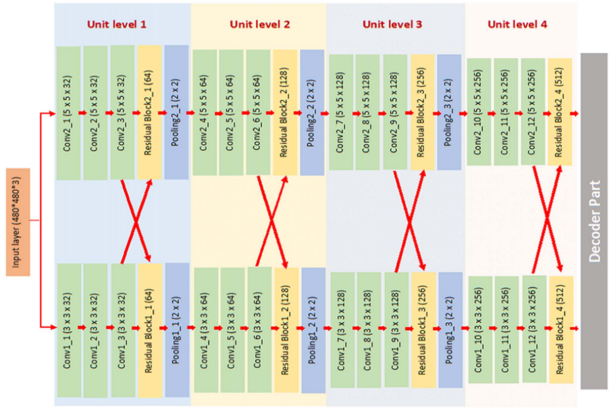

Fig. 1. Parallel deep convolutional network.

II. Methodology

As aforementioned, the proposed semantic segmentation network is based on a parallel cross-learning-based pixel transferred deconvolutional network (PCL–PTD) that comprises of a parallel convolutional network, residual block, cross-learning, pixel transferred deconvolution, and adjusted encoder and decoder networks. Details of each PCL–PTD net's component are described below.

A. Parallel Deep Convolutional Networks

The parallel convolutional networks comprise two deep convolutional networks. Each network performs 12 convolution layers to produce a set of feature maps and 4 max pooling methods to calculate translation invariances over small spatial shifts with 4-unit levels. It helps the model to learn multidimensional features from low to high levels. The numbers of filter banks are 32, 64, 128, and 256 at each stage level separately, as shown in Fig. 1. The increasing number of filter banks is for expanding learning capacity on feature maps to detect and extract target features. To learn object features, the first network (top) applies a receptive field with kernel size of 5×5 and the second network (bottom) is convoluted with a 3×3 kernel size to local operations. These multiple receptive fields can recognize features in different perspectives. Furthermore, we know Max-pooling operations could change feature maps into small translation features. A residual block is introduced to the networks and its operation is described in subsection B.

B. Residual Block

The residual block aims to solve degradation problems and expand feature maps. There are two steps in this process. First is to process the residual framework by adding a skip connection (x) from top layer to bottom layer in a convolutional block F(x). It optimizes residual learning when data are fed layer by layer. Second is to concatenate two sets of feature maps derived from previous layer H(x) and cross layer N(x), as shown in Fig. 2. This concatenation increases the number of feature maps and enlarges the learning capacity of features.

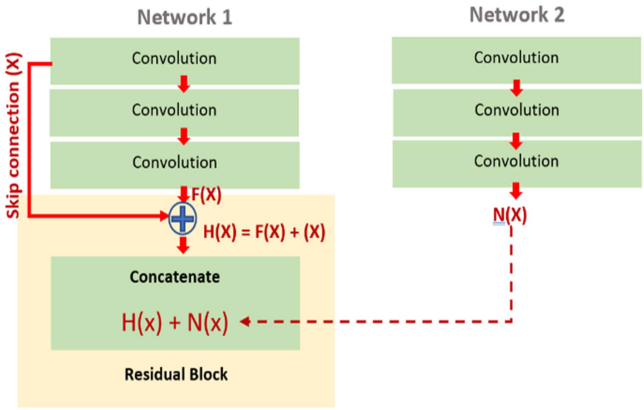

Fig. 2. Residual block.

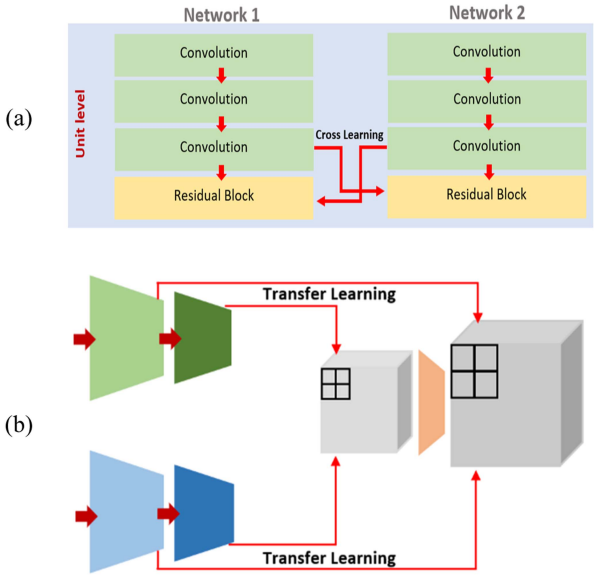

Fig. 3. Cross-learning framework: (a) cross learning and (b) transfer learning.

C. Cross-Learning Framework

This cross-learning framework consists of interconnected networks in order to transfer learning features, and it has two types: cross-learning encoder network and cross-learning between encoder–decoder network, as shown in Fig. 3. Fig. 3(a) shows the architecture of a cross-learning encoder network, which shared features map between parallel tracks of convolution layers. In this step, feature maps of the last convolution layer in each unit level transfer to the residual block located in opposite track to concatenate with other sets of feature maps. Fig. 3(b) illustrates a transfer learning framework, which implements cross-learning between encoder–decoder network, and we can see it builds direct relationships between encoder to decoder parts in order to solve checkerboard problems and deal with spatial features with edges and shapes suffered from regular convolution operations. Outputs of the residual block from parallel tracks will be sent to the upsampling function to shuffle the feature maps in encoder part, and then they will be combined in a deconvolutional operation of decoder part.

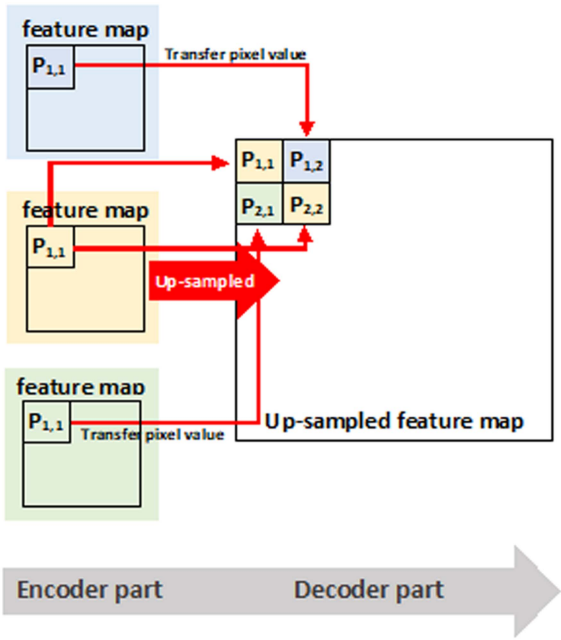

Fig. 4. Pixel transferred deconvolution.

D. Pixel Transferred Deconvolution

This proposed PCL–PTD net takes advantages of pixel deconvolution and transfer learning methods in order to unsampled the size of feature maps. Since simple deconvolution methods may cause checkerboard artifacts over the upsampled features map, which will produce inaccurate object features, edges, or shapes. Thus, the combination of pixel transferred deconvolution is proposed to take the benefits of feature relationship between encoder and decoder in maintaining spatial features suffered from periodical shuffling operations. This pixel transferred deconvolution method generates an up-sampled feature map, as shows in Fig. 4. In the process, a feature map with 1×1-unit pixel is upsampled to a feature map with 2×2-unit pixels. The transfer learning from convolutional layer in parallel encoder networks is applied to build direct relationships among encoder and decoder networks. The upsample processing uses values from transfer learnings and previous convolutional layers to add dependencies among indices (11), (22) and unit pixels (12), (21) in feature map, respectively.

E. Parallel Cross-Learning Based and Pixel Transferred Deconvolutional Network

The PCL–PTD aims to improve learning capacity of semantic segmentation on remote sensing images, as shown in Fig. 5. This network inherits the depth of convolutional neural network to detect and extract the various pattern features of object by generating invariant and abstract feature maps. The encoder part provides learning abilities on feature map(s), which takes advantage of the parallel deep convolutional network, residual block, and cross-learning framework. A corresponding decoder part upsamples feature map(s) into proper size that applies to transfer learning and pixel deconvolutional layer, as shown in Fig. 4.

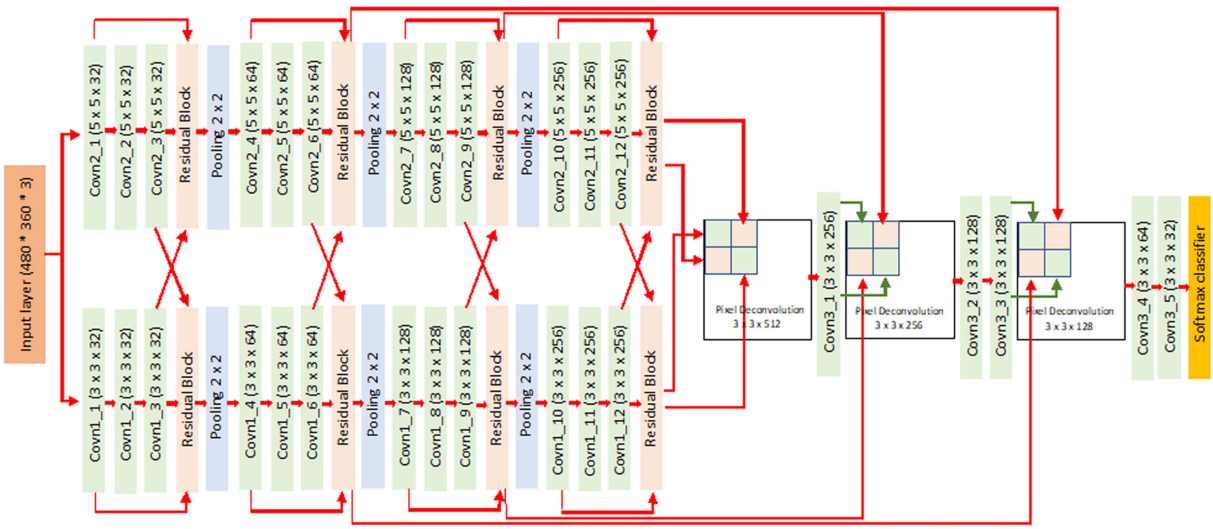

Fig. 5. Cross-learning-based and pixel transferred deconvolutional network.

The encoder part comprises parallel deep convolutional networks, which includes 24 convolutional layers, 6 max-pooling layers, and 8 residual blocks. The first layer feeds input data, which consists of three feature bands (red, green, and blue) with a size of 480 $\times$ 360 pixels to parallel networks. These networks have attractive properties to learn interesting patterns from SegNet network. The structure creates parallel learning and sharing features. Each network is convoluted by sets of filter banks from low-level to high-level features to generate smooth learning and various features. The top network has 12 convolutional layers with a 3 $\times$ 3 filtering kernel, 3 max-pooling layers, and 4 residual blocks. The bottom network is with the same structure as the top network, but the convolutional operation is applied with a 5 $\times$ 5 filtering kernel. The larger filtering kernels or receptive fields can increase the computation of its statistical efficiency in learning object features. In addition, there are 4-unit levels in encoder networks to generate feature maps in various perspectives in different spatial resolution. A unit level includes three convolution layers and a residual block, which has the cross-learning and residual unit methods. The residual block expands the learning feature maps by sharing a set of features between the networks, and it solves the degrading features from convolution method by skip connection. In detail, the third convolutional layer in each unit level is shared to a residual block in another network. The residual unit is implemented to each unit level by using skip connection from the first convolutional layer to the residual block. The output of residual block is fed to the next layer and pixel transferred deconvolution layer is located in decoder part. Then, the set of feature maps are fed to Max-pooling operation, which reduces the feature map resolution by a kernel with size of 2 $\times$ 2. It reduces the memory requirements of the model in storing the parameters and adds an infinitely strong capacity prior to learning small translations over object features. As a part of the decoder process, the last layers of parallel networks pass through the first pixel transferred deconvolution layer. This unsampled layer is to expand the feature map resolution with a factor of 2. In the adjustment process, the operation builds relationship with previous layer and corresponding layer in encoder part. It upsamples the feature map by adding a unit value from top network with kernel indices (21), (12) and a unit value from bottom network with kernel indices (11), (22). The output is then sent to the next convolution network. Furthermore, the second pixel transferred deconvolution layer upsamples feature maps and fills the unit values from residual block in top network to kernel index (12), the unit value from residual block in bottom network to kernel index (21), and the unit value from previous convolutional layer to kernel indices (11), (22). Thus, this decoder part consists of three pixel transferred deconvolution layers and five convolutional layers. Last, the feature maps are fed to a soft-max classifier to produce class probabilities. In total, this network architecture comprises 29 convolutional layers, 6 max-pooling layers, 8 residual blocks, 3 PTD layers, and 1 soft-max layer for building extraction on very high spatial resolution images.

F. Training

This adjustment network architecture is placed in supervised learning model for DL semantic segmentation. It is implemented based on four algorithms to achieve segmentation accuracy. The model optimizes the weight training in convolution layers by stochastic gradient descent in backpropagation algorithm. Hyperparameters including adaptive learning rates, momentum, and weight decay parameters are set to 0.001, 0.9, and 0.0005, respectively. The maximum round of iteration is defined as 100 000 times. The step size of learning is set to every 50 iterations with a factor of 10. To prevent modeling errors in statistics (overfitting or underfitting), batch normalization and dropout functions are introduced to the model. The cost function is set by early step techniques and L2 regularized logistic regression. This DL algorithm is implemented by TensorFlow with python on a PC with CPU of Intel Core i7 (3.4 GHz), RAM of 48 GB, and GPU NVIDIA GeForce RTX$^{\mathrm{TM}}$ 3060 Ti with 8 GB memory.

G. Evaluation Metrics

To evaluate the performance of supervised learning model, some quantitative accuracy metrics have been introduced to assess the learning procedure with evaluation and test datasets. Accuracy assessment is conducted in two steps. The first is to assess the learning procedure during its iterations with an evaluation dataset, and the second is to evaluate the performance of supervised learning model with a test dataset. The quantitative metrics used are overall accuracy (OA), mean intersection over union (mIoU), precision, recall, and per class IoU, as described below.

OA calculates the percentage of properly classified pixels [true positive (TP) and true negative (TN)] in the total number of pixels [TP, TN, false positive (FP), and false negative (FN)] as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \quad (1)$$

mIoU calculates the average IoU of all classes. The intersection over union (IoU) or the Jaccard index evaluates the ratio of intersection between all correctly classified pixels (TP) and the union of all correctly classified pixels (TP) and all falsely classified pixels (FP + FN), as follows:

$$\mathrm{mIoU} = J(A,B) = \frac{|A\cap B|}{|A\cup B|} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \quad (2)$$

Precision is calculated by the ratio of TP to the sum of a TP and FP, as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \quad (3)$$

Recall is expressed by the ratio of TP to the sum of a TP and FN, as follows:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \quad (4)$$

IoU or the Jaccard index computes the ratio of the intersection value (the number of TPs) to the union value (the sum of FPs, FNs, and TPs), as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \quad (5)$$

III. Experiments and Analysis

To verify the performance of building segmentation from remote sensing imagery, this article conducted the ablation experiment based on six proposed functions as described in Experiment 1. The adjustment network architectures, which were added on each proposed function, were evaluated on three different datasets: Inria aerial dataset, ISPRS Potsdam dataset, and UAV building dataset. Furthermore, the PCL–PTD net is compared with other state-of-the-art networks and adjustment networks, such as SegNet, U-Net, PSPnet, PixelDCL, DeeplabV3+, U-Net++, CFENet, and IRU-Net, to evaluate its effectiveness in building segmentation as demonstrated in Experiment 2–4. All experiments are computed based on six evaluation metrics: OA, mIoU, per class IoU, Precision, and Recall.

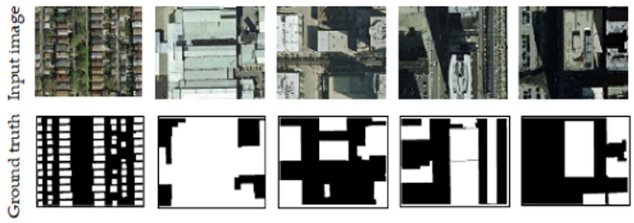

Fig. 6. Sample images from Inria aerial dataset: Upper is RGB images and lower is labeled images (white areas are buildings and black areas are nonbuilding).

TABLE I
Number of Samples for Training, Validating, and Testing From Inria Aerial Dataset

Dataset	**Training**	**Validating**	**Testing**
Inria aerial dataset	23 100	3850	770

A. Experiment 1: Ablation Experiments of Proposed Functions on the Inria Aerial Dataset

Inria aerial image labeling dataset is an open dataset released by [34]. It is generated from aerial photographs with very high spatial resolution captured over the USA and Austria. This dataset shows very dense and high building structures in Austin (TX), Chicago (IL), Kitsap County (WA), Tyrol, and Vienna. Fig. 6 presents samples of aerial photographs in this dataset: RGB orthoimages and labeled images with spatial resolution of 30 cm. There are 24 densely annotated image tiles with the original image size is 6000 × 6000 pixels. A total of 18 tiles are used for training, with 20% of training images being randomly selected for validating set. The other six tiles are used for testing. Every image is cropped into small pieces with an image size of 480 × 360 pixels. The annotated images have two classes: building and nonbuilding. The total numbers of image patches from Inria aerial dataset are provided in Table I.

This ablation experiment illustrates the improvement of building extraction over Inria aerial dataset. The network architectures are adjusted by the proposed functions including SegNet-based network, parallel network, cross-learning function, residual unit function, pixel deconvolution function, and pixel transferred deconvolution function. The performance of ablation experiments is listed in Table II. The quantitative comparisons of the adjustment network show that these networks can accurately detect and extract the buildings over remote sensing data. The SegNet based network (EX1) performs well in building extraction with 86.43% of OA and 72.90 of mIoU. When the SegNet-based network adds residual unit function (EX2) to solve degradation problems. The performance of building extraction increases to 86.89% of OA and 73.10% of mIoU. The pixel deconvolution function proposed to decoder part (EX3) in order to make direct relationship in adjacent pixels to perform upsampling feature maps. It can improve the performance of building extraction with 87.43% of OA and 73.76% of mIoU. Then, the pixel transferred deconvolution function is applied to the network (EX4) to solve the checkerboard artifacts. It can increase OA up to 87.77%

TABLE II
ABLATION EXPERIMENTS ON INRIA AERIAL DATASET

Methods	Functions						mIoU	OA
	SegNet-based network	Parallel network	Cross learning function	Residual unit function	Pixel deconvolution function	Pixel Transferred deconvolution		
EX1	✓						72.90	86.43
EX2	✓			✓			73.10	86.89
EX3	✓			✓	✓		73.76	87.43
EX4	✓			✓	✓	✓	73.89	87.77
EX5	✓				✓		73.28	87.02
EX6	✓				✓	✓	74.01	87.67
EX7	✓	✓					77.80	88.89
EX8	✓	✓		✓			77.78	88.91
EX9	✓	✓		✓	✓		77.97	89.03
EX10	✓	✓		✓	✓	✓	83.65	90.87
EX11	✓	✓	✓				76.80	88.38
EX12	✓	✓	✓	✓			78.80	89.39
EX13	✓	✓	✓	✓	✓		83.80	91.92
EX14	✓	✓	✓	✓	✓	✓	85.90	92.93

and 73.89% of mIoU. However, the SegNet-based network adds only pixel deconvolution function (EX5). The performance of OA drops to 87.02% of OA and 73.28% of mIoU. When the network in EX5 compounds pixel transferred deconvolution function (EX6), it can improve the accuracy up to 87.67% of OA and 74.01% of mIoU. Furthermore, parallel network inspired by SegNet based (EX7), which expands learning capacity from multiple receptive fields, shows better performance of building extraction. It has 88.89% of OA and 77.80% of mIoU. When parallel network applies the residual unit function (EX8). The performance presents a little improvement with 88.91% of OA and 77.78% of mIoU. The integrated network (EX9) of SegNet-based network, parallel network, residual unit function, and pixel deconvolution function shows good performance on building extraction with 89.03% of OA and 77.97% of mIoU. When previous network applies pixel transferred deconvolution function in decoder part (EX10), the model illustrates the improvement of OA up to 90.87% and 83.65%. Moreover, the improvement of encoder part presents the combinations of SegNet-based network, parallel network, and cross-learning function (EX11). The performance is about 88.38% of OA and 76.80% of mIoU. When the previous model (EX11) adds the residual unit function (EX12), the OA and mIoU increase to 89.39% and 78.80%, respectively. The improvement of the network by adding pixel deconvolution function (EX13) shows better performance with 91.92% of OA and 83.80% of mIoU. Last, with our proposed network (PCL–PTD net), the combination of six proposed functions (EX14) presents the highest OA and mIoU with 92.93% and 85.90% and outperforms other adjustment networks (EX 1–EX13). It shows that the integrated functions yield advantages in learning ability in order to segment the buildings over remote sensing data.

B. Experiment 2: Quantitative and Qualitative Results on the Inria Aerial Dataset

This experiment is to evaluate the improvement of the proposed PCL–PTD net, when the adjustment networks are added on each function for extracting buildings in dense urban areas with building shadows and unclear building features. Table III lists results of accuracy assessment and Fig. 7 shows the segmentation results of experiment 2. The baseline SegNet network presents an accuracy result with 86.43% of OA and 72.9% of mIoU. The segmented building by SegNet illustrates that the network can segment building accurately, but it does not work well in dense building area, as shown in Fig. 7 (EX1). The parallel SegNet-based network achieves a very good segmentation accuracy with 88.89% of OA and 77.8% of mIoU, and it also shows that parallel network can learn complex features in dense building area, but it introduces segmentation errors in shadow area, as shown in Fig. 7 (EX2). Furthermore, cross-learning function is applied to parallel network in order to share learning ability between networks, but this network (EX3) shows worse performance than previous network (EX2), as listed in Table III, with only a little better result in per class IoU (building). The improvement of parallel network and cross-learning by adding residual unit function works better with 89.39% of OA and 78.8% of mIoU in building extraction, as shown in Fig. 7 (EX4). It can be seen that dense buildings are segmented in a fairly accurate way, but it presents some errors in areas with unclear building features and shadows. Then a PixelDCL function is applied to the decoder part, where parallel network, cross-learning, and residual unit are implemented in encoder part. This function outputs upsampled feature map and could solve checkerboard artifacts. This adjusted network (EX5) presents an increase in accuracy with 89.39% of OA and 78.8% of mIoU. This network works fairly well in dealing with building segmentation in dense building areas, and it also improves building detection and extraction in shadow areas, as shown in Fig. 7 (EX5). For our proposed adjustment network PCL–PTD net, which consists of a parallel network, cross-learning, residual unit, and pixel transferred deconvolution, it performs best with 92.93% of OA and 85.9% of mIoU. The segmented results (EX6) are accurate in dense building areas where gaps among buildings are very small, as shown in Fig. 7 [EX6(a), (b)], and building under shadows are extracted accurately, as shown in Fig. 7 [EX6(c), (d)]. However, it also does not work when buildings are totally covered by dark shadows.

TABLE III
ACCURACY RESULT (%) OF COMPARATIVE MODEL ON INRIA AERIAL DATASET

Methods	Functions						Per class IoU		Precision	Recall	mIoU	OA
	SegNet-based network	Parallel network	Cross-learning function	Residual unit function	Pixel deconvolution function	Pixel Transferred deconvolution	Building	Non-building				
EX1	✓						86.43	87.00	86.74	85.86	72.90	86.43
EX2	✓	✓					87.01	89.53	89.69	87.88	77.80	88.89
EX3	✓	✓	✓				87.88	88.90	88.78	87.86	76.80	88.38
EX4	✓	✓	✓	✓			88.47	89.63	89.80	88.90	78.80	89.39
EX5	✓	✓	✓	✓	✓		90.67	92.43	92.78	90.91	83.80	91.92
EX6	✓	✓	✓	✓	✓	✓	93.94	91.92	92.08	93.94	85.90	92.93

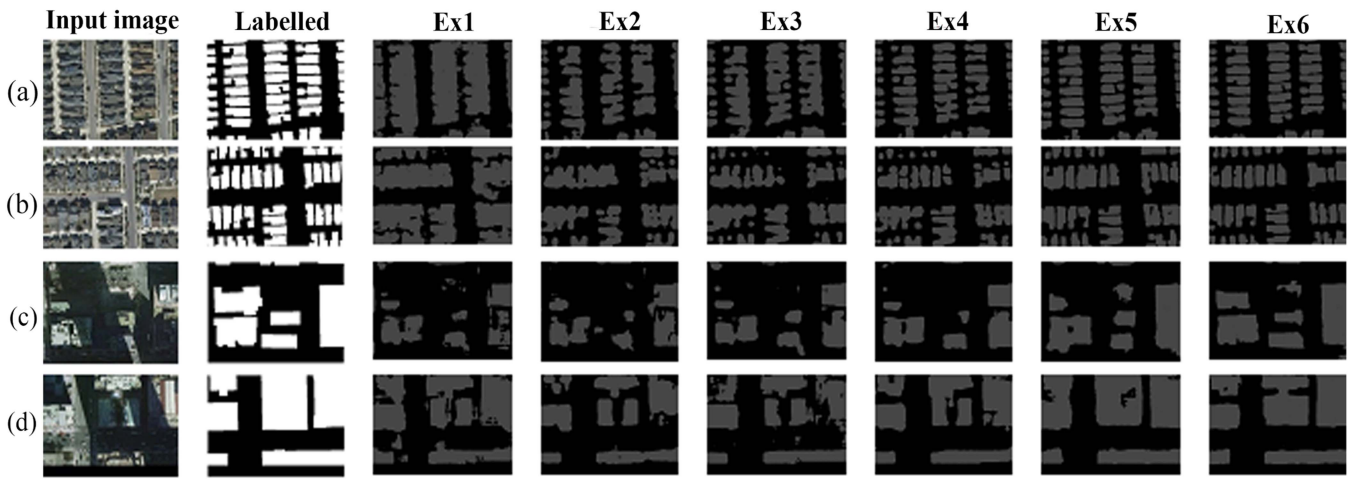

Fig. 7. Segmentation results of comparative model on Inria aerial dataset.

TABLE IV
ACCURACY RESULT (%) OF PROPOSED NETWORK AND EIGHT COMPARATIVE METHODS ON INRIA AERIAL DATASET

Network	Per class IoU		Precision	Recall	mIoU	OA
	Building	Non building				
SegNet	86.43	87.00	86.74	85.86	72.90	86.43
U-Net	88.45	87.89	88.00	88.90	76.80	88.38
PSPnet	87.99	89.01	89.70	87.87	77.80	88.90
PixelDCL	90.01	87.77	88.23	90.39	78.80	89.35
DeeplabV3+	91.74	88.08	89.37	92.83	79.10	90.41
U-Net++	90.55	86.23	87.37	90.90	77.80	88.89
CFENet	92.15	88.54	89.32	92.93	81.80	90.09
IRU-Net	92.83	89.63	90.20	92.93	82.80	91.42
PCL-PTD net	93.94	91.92	92.08	93.94	85.90	92.93

The quantitative and qualitative comparisons of the different networks for the testing set is presented the performance of the proposed network architecture (PCL–PTD net) with other six baseline networks and two adjustment networks. Table IV lists the accuracy results of all networks. Though all these networks can detect and extract buildings with a fairly high accuracy, as shown in Fig. 8(a), some networks show their weakness in Fig. 8(b)–(d). To be specific to our proposed network, it could be seen that the PCL–PTD net works well in dealing with building extraction from so complex scenes with the highest accuracy. The SegNet presents the lowest accuracy with 86.43% of OA and 72.9% of mIoU in building extraction among the eight networks. It can be used to detect and extract buildings on high spatial resolution images, but it is weak in areas with dense buildings shadows, as shown in Fig. 8 [SegNet(b)–(d)]. The most commonly used network architecture, the U-Net, is with a U-shaped encoder–decoder network architecture, and shows better performance with 88.38% of OA and 76.8% of mIoU. This network can work well in extracting building shapes in dense building areas, but it is weak in areas where buildings are covered by shadows, as shown in Fig. 8 [U-Net(b)–(d)]. The performance of PSPnet, a global context aggregation by pyramid pooling module in different region based, is higher than the U-Net with 88.90% of OA and 77.8% of mIoU. Whereas, its result in per class IoU is less accurate than that of the U-Net network, as shown in Fig. 8 [PSPnet (b)–(d)]. The PixelDCL architecture works well on building segmentation with 89.35% of OA and 78.8% of mIoU, and it also presents better result in per class IoU of building class with 90.01% accuracy. The segmented result shows its good performance in detecting and extracting buildings in dense building areas, but not in areas with high buildings and shadows, as shown in Fig. 8 [PixelDCL(b)–(d)]. The DeeplabV3+ is the latest version of DeepLab series that comprises multiple atrous convolutional rates and aligned Xception model. The performance has scored the best value 90.41% of OA and 79.01% of mIoU ahead of the PixelDCL. The segmented images show an accurate of building segmentation over the gaps among buildings and unclear building features. However, it lacks in building shape and shadow building, as shown in Fig. 8 [DeeplabV3+(a)–(d)]. The improvement network, the U-Net++, is an essentially encoder and decoder subnetwork, which is connected through a series of nested and

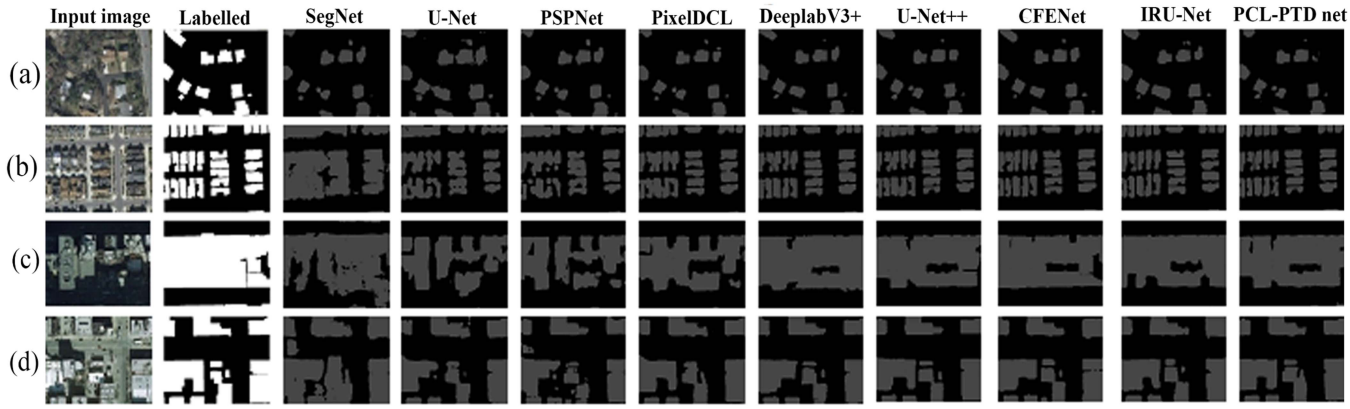

Fig. 8. Segmentation results of comparative networks on Inria aerial dataset.

dense skip pathways. This network presents its performance with 88.89% of OA and 77.80 of mIoU that lower accuracy than the DeeplabV3+, the PixelDCL, the PSPnet, respectively. The segmentation results illustrate errors of building shape and narrow gaps among buildings, as shown in Fig. 8 [U-Net++(b)–(d)]. For adjustment networks which designed for building extraction, the context feature enhancement network, the CFENet [29], is selected to this comparison. The performance is about 90.09% of OA and 81.80 of mIoU, which is lower accuracy than the DeeplabV3+ (0.32%), but higher than other state-of-the-art networks. The segmented images show better results in building shape. But it is an error in shadow buildings, as shown in Fig. 8 [CFENet(c) and (d)]. Furthermore, the IRU-Net [28] is integrating the residual learning and atrous spatial pyramid pooling methods, skip connection for automatic building extraction. This network achieves high accuracy with 91.92% of OA and 83.80% of mIoU. The model can detect the complex building features and extract the building shape accurately, as shown in Fig. 8 [IRU-Net(a) and (b)]. Its performance is claimed the same as [27]. But the shadow area shows an error of building extraction, as shown in Fig. 8 [IRU-Net(c) and (d)]. While our proposed network architecture (PCL–PTD net) performs best with 92.93% of OA and 84.3% of mIoU, together with 93.94% of per class IoU in building class. Its segmentation outperforms the other eight baseline networks in detecting and extracting narrow gaps among buildings in dense buildings areas, as shown in Fig. 8 [PCL-PTD net(b)]. Furthermore, it also works well in segmenting buildings under shadows, as shown in Fig. 8 [PCL–PTD net(c)], as well as in dense areas with tall buildings, as shown in Fig. 8 [PCL–PTD net(d)].

C. Experiment 3: Quantitative and Qualitative Results on the ISPRS Potsdam Dataset

The ISPRS Potsdam dataset is an open dataset provided by the commission III of ISPRS, which is available online [35]. It is a very high-resolution aerial photograph with spatial resolution of 5 cm. The images captured over the Potsdam city in Germany, where there are the dense settlement structures. The dataset consists of 36 images tiles, while 30 tiles were used for training set, 20% of training set were randomly selected for validating

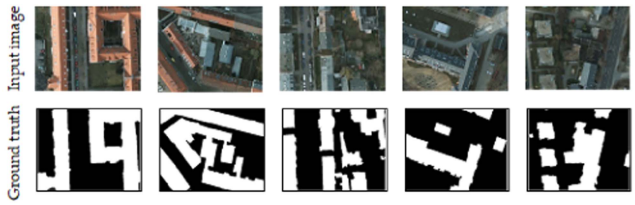

Fig. 9. Sample images from ISPRS Potsdam dataset: Upper is RGB images and lower is labeled images (white areas are buildings and black areas are nonbuilding).

TABLE V
NUMBER OF SAMPLES FOR TRAINING, VALIDATING, AND TESTING FROM ISPRS POTSDAM DATASET

Dataset	Training	Validating	Testing
ISPRS Potsdam dataset	320	80	80

set. The remaining six tiles were used for a testing set. An image comprises 1500 × 1500 pixels. The annotated image was labeled into two classes: building and nonbuilding. Each image tile was clipped and split to 480 × 360 pixels, as shown in Fig. 9. The number of samples for training, validating, and testing from ISPRS Potsdam dataset shown in Table V.

This dataset is to verify the performance of the proposed PCL–PTDnet to detect and extract the building in high building and building shadow. Table VI and Fig. 10 show the accuracy result and building segmentation over the ISPRS Potsdam dataset. The EX1 shows its performance with 89.39% of OA and 78.8% of mIoU. The segmentation results work well on building segmentation, but it lacks in building shadow, as shown in Fig. 10 [EX1(f)–(h)], and unclear building feature as shown in Fig. 10 [EX1(f) and (h)]. EX2 works better than EX1 with increasing 89.90% (+0.51%) of OA and 79.8% (+1%) of mIoU. The results overcome the building shadow problem, as shown in Fig. 10 [EX2(g) and (h)]. However, it worse in unclear building features, as shown in Fig. 10 [EX2(f)]. Furthermore, the EX3 achieves high performance over the EX2 with 90.91% (+1.01%) of OA and 81.8% (+2%) of mIoU. This adjustment network can learn and segment unclear building features, as shown in Fig. 10

TABLE VI
ACCURACY RESULT (%) OF COMPARATIVE MODEL ON ISPRS POTSDAM DATASET

Methods	Functions						Per class IoU		Precision	Recall	mIoU	OA
	SegNet-based network	Parallel network	Cross-learning function	Residual unit function	Pixel deconvolution function	Pixel transferred deconvolution	Building	Nonbuilding				
EX 1	✓						87.45	90.48	90.63	87.39	78.80	89.39
EX 2	✓	✓					87.98	91.44	91.58	87.88	79.80	89.90
EX 3	✓	✓	✓				88.56	92.05	92.63	88.89	81.80	90.91
EX 4	✓	✓	✓	✓			90.12	91.24	91.84	90.91	82.80	91.41
EX 5	✓	✓	✓	✓	✓		91.76	93.01	93.81	91.00	84.90	92.46
EX 6	✓	✓	✓	✓	✓	✓	92.03	94.53	94.85	91.09	86.00	93.00

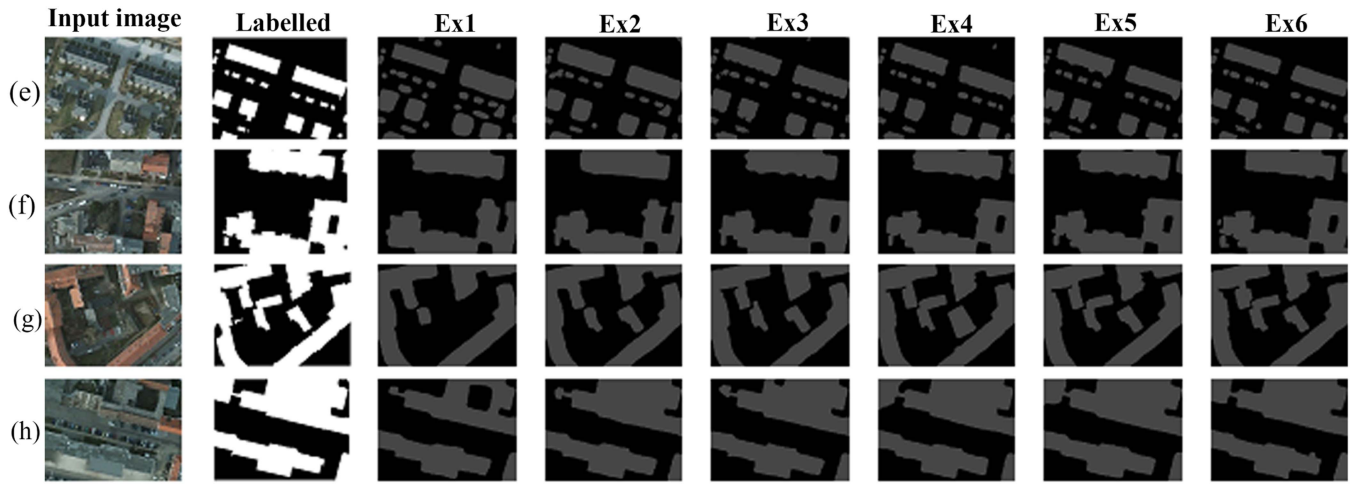

Fig. 10. Segmentation results of comparative on ISPRS Potsdam dataset.

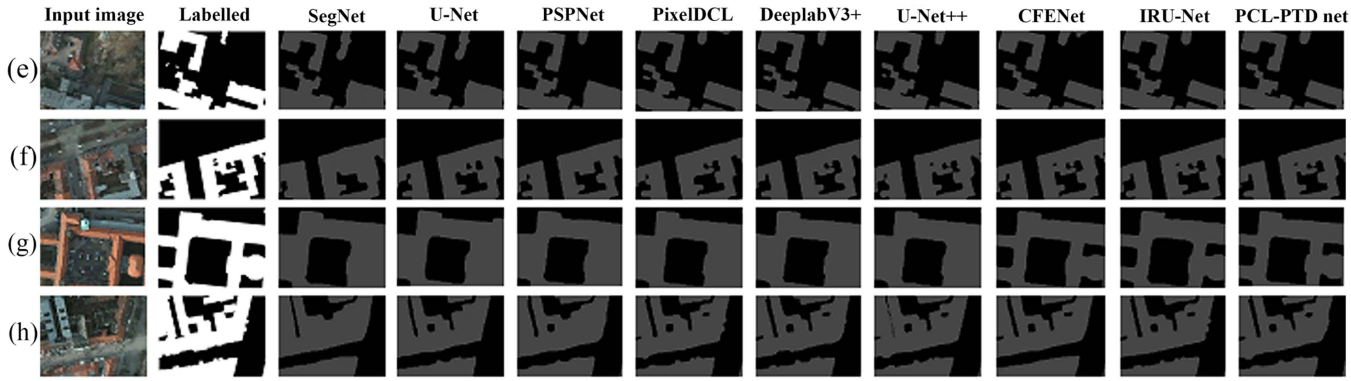

Fig. 11. Segmentation results of comparative networks on ISPRS Potsdam dataset.

[EX4(f)]. The EX4 outperforms the EX3 by 91.41% (+0.5%) of OA and 82.8% (+1%) of mIoU. This experiment segments very well in unclear building features, but it shows some errors in the building shape, as shown in Fig. 10 [EX4(f)–(h)]. The EX5 improves 92.46% (+1.05%) of OA and 84.9% (2.1%) of mIoU over The EX4. The network can segment the building accurately, as shown in Fig. 10 [EX4(e)–(h)]. The EX6 overcomes the EX5 with 93.00% (+0.54%) of OA and 86.00% (+1.1%) of mIoU. The EX6 also shows an increase of per class IoU in building class with 92.03% over the EX1 (87.45%), the EX2 (87.98%), the EX3 (88.56%), the EX4 (90.12%), and the EX5 (91.76%), respectively. The proposed network can learn complex building features, shadow building, narrow gaps among buildings, and also segment the building in accurate shape as shown in Fig. 10 [EX6(e)–(h)].

To test the efficiency of PCL–PTDnet with baseline networks and adjustment networks, the accuracy results show in Table VII and segmented images show in Fig. 11. The standard encoder–decoder network (SegNet) achieves their accuracy result with 84.85% of OA and 69.7% of mIoU. The SegNet architecture can detect and extract the building over high resolution imagery as shown in Fig. 11 [SegNet(e)–(h)]. However, this network

TABLE VII
ACCURACY RESULT (%) OF PROPOSED NETWORK AND EIGHT COMPARATIVE METHODS ON ISPRS POTSDAM DATASET

Network	Per class IoU		Precision	Recall	mIoU	OA
	Building	Non building				
SegNet	85.22	85.37	84.16	85.86	69.70	84.85
U-Net	87.01	85.23	86.14	87.88	73.70	86.87
PSPnet	86.68	86.55	85.15	86.87	71.70	85.86
PixelDCL	87.54	85.75	86.14	87.88	73.70	86.87
DeeplabV3+	89.06	87.23	88.19	89.90	77.80	88.89
U-Net++	86.76	88.43	88.66	86.87	75.80	87.88
CFENet	90.23	87.63	88.24	90.91	78.80	89.40
IRU-Net	88.78	86.45	87.13	88.89	75.80	87.88
PCL-PTD net	90.78	88.63	89.11	90.91	79.80	89.90

shows worse segmentation in unclear building features, building shadow, and narrow gaps among buildings. The other baseline networks show better performance with 86.87% of OA and 73.7% of mIoU in the U-Net, 85.86% of OA and 71.7% of mIoU in the PSPnet, and 86.87% of OA and 73.7% of mIoU in the PixelDCL. These network architectures can overcome narrow gaps among buildings, as shown in Fig. 11 [U-Net(f) and (h), PSPnet(f) and (h), and PixelDCL(f) and (h)]. However, the U-Net and the PSPnet lacks to segment unclear building features as shown in Fig. 11 [U-Net(e) and PSPnet(e)]. For the improvement networks, the DeeplabV3+ and the U-Net++, these networks work well on building segmentation in narrow gaps among buildings and unclear building features with 88.89% of OA and 77.8% of mIoU in the DeeplabV3+, and 87.88% of OA and 75.8% of mIoU in the U-Net++, as shown in Fig. 11 [DeeplabV3+(e), (f), and (h), U-Net++(e), (f), and (h)]. Furthermore, the adjustment networks, the CFENet, and the IRU-Net were designed for building extraction. Their accuracy results are increasing up to 89.40% of OA and 78.8% of mIoU in the CFENet, and 87.88% of OA and 75.8% of mIoU in the IRU-Net. This network can segment shadow building, but it shows some errors on unclear building features, as shown in Fig. 11 [CFENet(e) and (g) and IRU-Net(e) and (g)]. Our proposed network architecture, the PCL–PTDnet, outperforms other network architectures with 89.90% of OA and 79.8% of mIoU. It shows the highest performance at 90.78% in per class IoU in building class than the SegNet (85.22%), the U-Net (87.01%), the PSPnet (86.68%), the U-Net++ (86.76%), the PixelDCL (87.54%), the DeeplabV3+ (89.06%), the IRU-Net (88.78%), the CFENet (90.23%), respectively. The segmentation results outperform unclear building shape, building gap, and shadow building as shown in Fig. 11 [PCL–PTDnet(e)–(h)].

D. Experiment 4: Quantitative and Qualitative Results on UAV Building Dataset

UAV building dataset is produced by UAV mapping with very high spatial resolution of 2–4 cm. It was collected over riverbank area in Chongqing City, China from 20 flights covering dense building area and countryside. There are 16 mappings examined by training set and 20% of training sample were randomly selected for validating set. The other four mappings were used for testing set, which represent the dense building and fairly sparse area. The annotated image comprises two classes: building and nonbuilding. Each mapping was clipped and split to 480 × 480 pixels, as shown in Fig. 12. The number of trainings, validating, and testing samples is shown in Table VIII.

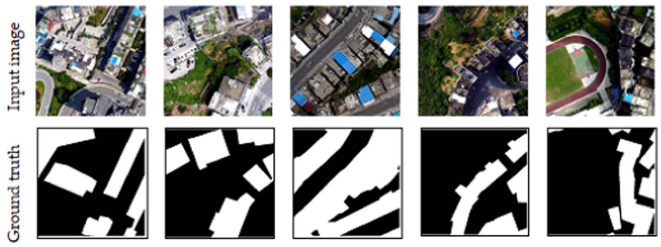

Fig. 12. Sample images from UAV building dataset: Upper is RGB images and lower is labeled images (white areas are buildings and black areas are nonbuilding).

TABLE VIII
NUMBER OF SAMPLES FOR TRAINING, VALIDATING, AND TESTING FROM UAV BUILDING DATASET

Dataset	Training	Validating	Testing
UAV building dataset	3000	600	200

This dataset is to examine the proposed network architecture. Table IX presents the quantitative accuracy, and Fig. 13 shows the qualitative segmentation results for building extraction. Following the baseline network (EX1), it gains 84.34% of OA and 68.7% of mIoU. This network can detect and extract the building, but it shows some extraction errors in narrow gaps among buildings [EX1(k) and (j)], unclear building features [EX1(k) and (j)], and building shadow [EX1(i) and (l)]. The EX2 presents better results than the EX1 with 84.85% (+0.51%) of OA and 69.7% (+1%) of mIoU. This network can differentiate the narrow gaps among buildings, as shown in Fig. 13 [EX2(j) and (k)], but it still lacks unclear building features, as shown in Fig. 13 [EX2(i) and (k)]. The EX3 increases 85.86% (+1.01%) of OA and 71.7% (+2%) of mIoU over the EX2. Its performance can segment narrow gaps among buildings accurately, as shown in Fig. 13 [EX3(i)–(k)]. The errors remain in building shadow. Moreover, the EX4 performs higher performance with 86.87% (+1.01%) of OA and 73.7% (+2%) of IoU than the EX3. It shows a good segmentation in building shadow area. However, unclear building features are worse in this network, as shown in Fig. 13 [EX4(j) and (k)]. The EX5 achieves in accuracy with 88.38% (+1.51%) of OA and 76.8% (+3.1%) of mIoU. The network can detect and extract complex building features, but it lacks the building shape, as shown in Fig. 13 [EX5(l)]. For our proposed network, the EX6 overcomes the EX5 with 89.39% (+1.01%) of OA and 78.8% (+2%) of mIoU. It shows a good performance of building segmentation in building shape, building shadow, and unclear building features, as shown in Fig. 13 [EX6(i)–(l)].

This experiment is also made by comparing the proposed network with others standard networks and adjustment

TABLE IX
ACCURACY RESULT (%) OF COMPARATIVE MODEL ON UAV BUILDING DATASET

Method	SegNet-based network	Parallel network	Cross-learning function	Residual unit function	Pixel deconvolution function	Pixel transferred deconvolution	Per class IoU		Precision	Recall	mIoU	OA
							Building	Nonbuilding				
EX 1	✓						82.35	85.67	85.42	82.83	68.70	84.34
EX 2	✓	✓					82.78	86.63	86.32	84.85	69.70	84.85
EX 3	✓	✓	✓				83.45	87.44	87.37	83.84	71.70	85.86
EX 4	✓	✓	✓	✓			84.45	88.23	88.42	84.85	73.70	86.87
EX 5	✓	✓	✓	✓	✓		85.77	90.44	90.43	85.86	76.80	88.38
EX 6	✓	✓	✓	✓	✓	✓	86.21	91.68	91.49	86.87	78.80	89.39

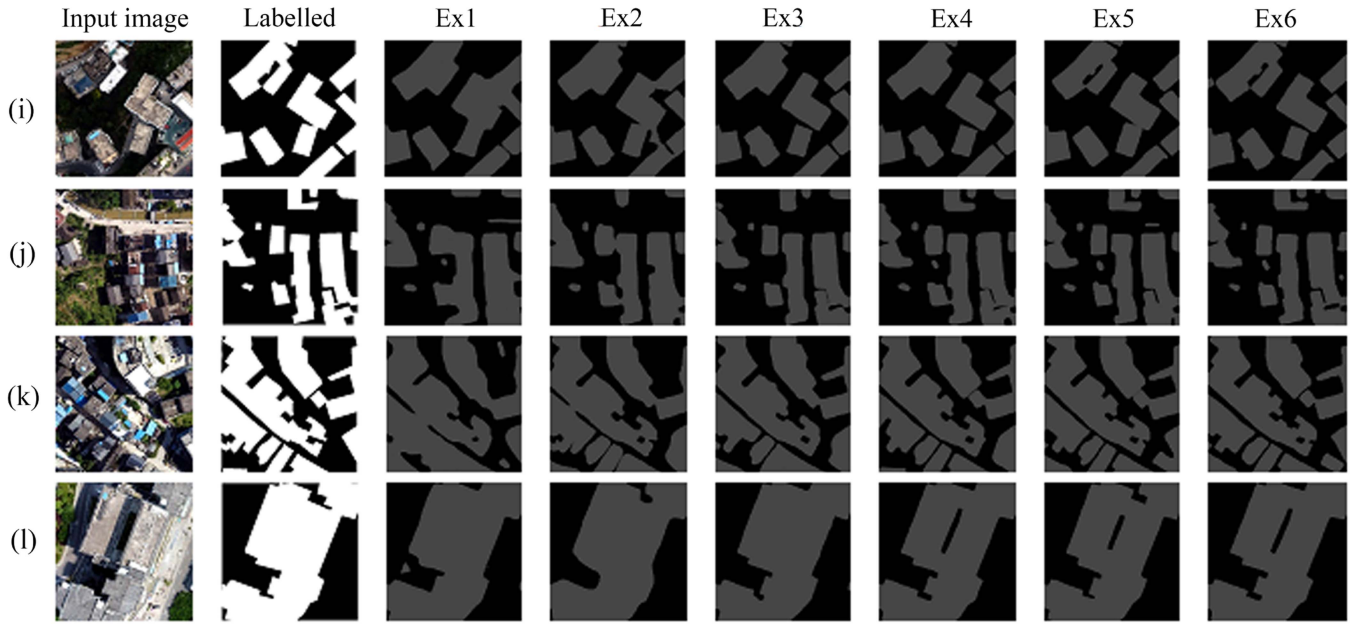

Fig. 13. Segmentation results of comparative model on UAV building dataset.

TABLE X
ACCURACY RESULT (%) OF PROPOSED NETWORK AND EIGHT COMPARATIVE METHODS ON UAV BUILDING DATASET

Network	Per class IoU		Precision	Recall	mIoU	OA
	Building	Non-building				
SegNet	82.35	85.67	85.42	82.83	68.70	84.34
U-Net	84.01	86.89	86.60	84.85	71.70	85.86
PSPnet	83.26	87.56	87.37	83.84	70.90	85.65
PixelDCL	85.48	84.86	85.00	85.86	73.70	85.35
DeeplabV3+	84.87	87.56	87.50	84.85	72.70	86.36
U-Net++	84.23	86.75	86.60	84.85	71.70	85.86
CFENet	84.35	85.43	85.72	84.85	70.70	85.35
IRU-Net	83.96	87.02	87.37	83.83	72.40	85.86
PCL-PTD net	86.12	85.37	86.00	86.87	72.70	86.36

networks. The accuracy results are shown in Table X and Fig. 14. The lowest accuracy is the SegNet with 84.34% of OA and 68.7% of mIoU. It shows errors in unclear building features and shadow building, as shown in Fig. 14 [SegNet(i)–(l)]. Other standard networks gain better performance with 85.86% of OA and 71.7% of mIoU in the U-Net, 85.65% of OA and 70.9 of mIoU in the PSPnet, and 85% of OA and 73.7% of mIoU in the PixelDCL. These networks can learn complex building features, but it is worse segmentation in narrow gaps among buildings. For improvement network, these networks show the high performance with 86.36% of OA and 72.7% of mIoU in the DeeplabV3+, and 85.86% of OA and 71.7% of mIoU in the U-Net++, respectively. The results show better segmentation over narrow gaps among buildings, as shown in Fig. 14 [DeeplabV3+(i), (j), and (l) and U-Net++(j) and (l)]. However, the segmented result lacks shadow building as shown in Fig. 14 [DeeplabV3+(k) and U-Net++(i) and (k)]. Furthermore, the adjustment network architectures for building extraction perform 85.35% of OA and 70.7% of mIoU in the CFENet, and 85.86% of OA and 72.4% of mIoU in the IRU-Net, which overcome other standard network architectures. But the IRU-Net has limitation for extracting shadow building, as shown in Fig. 14 [IRU-Net(k)]. The PCL–PTDnet achieves the highest OA of 86.36% and mIoU of 72.7%. For per class IoU in building class, the PCL–PTDnet gains 86.12%, which also overcomes other network architectures with 82.35% of the SegNet, 83.26% of the PSPnet, 83.96% of the IRU-Net, 84.01% of the U-Net, 84.23% of the U-Net++, 84.35% of the CFENet, 84.87% of the DeeplabV3+, and 85.48% of the PixelDCL, accordingly. The performance also shows a good segmentation over shadow building, narrow gaps among buildings, building

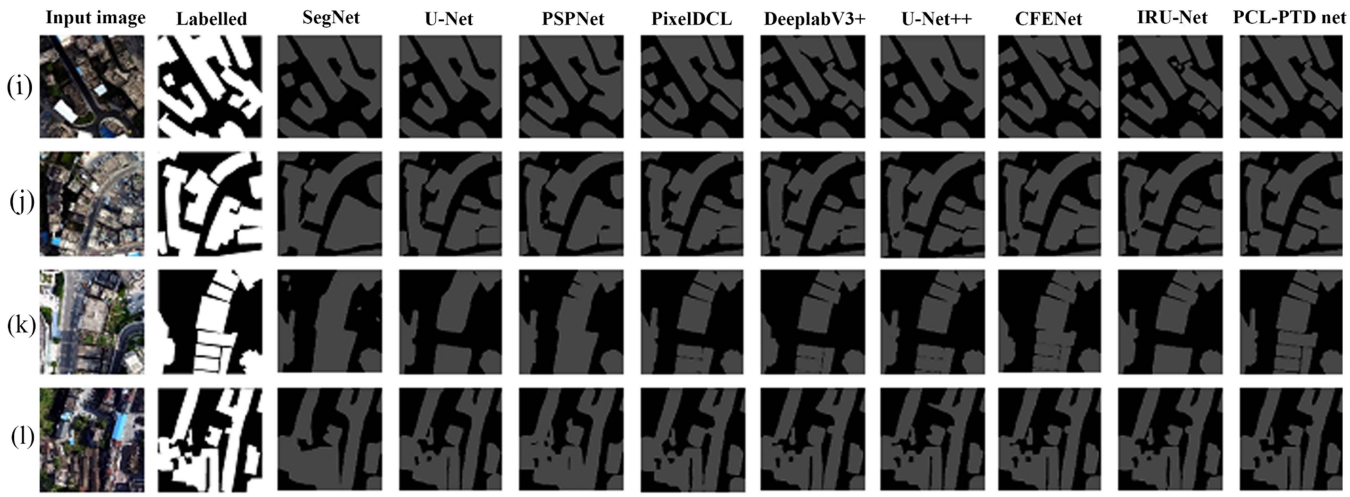

Fig. 14. Segmentation results of comparative networks on UAV building dataset.

shape, and unclear building features, as shown in Fig. 14 [our proposed(i)–(l)].

IV. Discussion

The proposed PCL–PTDnet shows its advantage in detecting and extracting buildings in dense building areas with shadows and narrow gaps among buildings. In the whole adjustment process, it can be seen that the baseline SegNet network, which is with an end-to-end network architecture, shows its good performance in learning building features and extracting buildings. The model takes advantage of deep convolution layer and lower resolution feature maps to learn multidimensional building features, and uses pooling indices of the corresponding encoder to upsample feature maps in accurate building features. This network has worse segmentation results in building shapes. It is because of its simple deconvolution method. The model is weak in dealing with dense building segmentation. While the parallel SegNet network with different receptive fields can enhance learning ability to detect and extract complex building features. The parallel SegNet network improves the number of learning filters. The multiple receptive fields help the model to detect and extract the buildings in multidimensional object features. This model shows better performance in learning and extracting building features in dense building zones. Later, the combination of parallel SegNet network and cross-learning function shows its performance to detect and extract the dense building area and narrow gaps among buildings accurately. This model shares feature maps and convolutes building features in different sizes of convolutional filter. The supervised learning model improves learning ability to differentiate between building features and other object features. The integration of parallel SegNet network, cross-learning function, and residual block improves segmentation accuracy. This network enhances encoder capacity in learning multiple dimensional building features and solves degradation problems when the feature maps are convoluted through deep convolution layers. The model can detect and extract unclear building features, gaps among buildings, and building in shadow areas. Furthermore, by adding a pixel deconvolution function in the decoder part, this adjustment network is designed to solve checkerboard problems. It is because of no direct relationship in adjacent pixels to perform upsampling feature maps. This supervised learning model improves the segmentation results of building features in urban areas and dense building zones. The dense and tall buildings with shadow have been largely solved. The proposed adjustment network (PCL–PTDnet), which comprises a parallel network, cross learning, residual unit, and pixel transferred deconvolution, shows better performance in quantitative accuracy and qualitative segmentation results. The supervised learning model can detect and extract complex building features, narrow gaps among buildings, building shadow, and building shape accurately. The proposed algorithms benefit each other synergistically to yield improved building segmentation performance.

Experiments on the three datasets (Inria aerial dataset, ISPRS Potsdam dataset, and UAV building dataset) have shown that the proposed architecture has competitive performance with the six baseline networks (SegNet, U-net, PSPnet, PixelDCL, DeeplabV3+, and U-Net++ network) and two adjustment networks (CFENet and IRU-Net network). The quantitative and qualitative results illustrate that all networks perform relatively well in building extraction. But our proposed network architecture achieved the highest OA and mIoU value. The segmentation results overcome building extraction in case of dense building area, shadow building, and unclear texture features. For Inria aerial dataset, which presents very dense and high building structures and cement textures in roofs and grounds, this article conducted many experiments, including quantitative and qualitative analysis. The adjustment network shows the improvement of building extraction when applies the proposed functions. It yielded better segmentation over complexity of building features in small buildings covered partly by tree branches, dense building area, narrow gaps among buildings, and buildings in shadow area. Compared with the baseline networks and adjustment networks, the segmented results demonstrate that the PCL–PTDnet has improved the reliability of building extraction

to the comparative algorithms. Furthermore, the ISPRS Potsdam dataset, which represents the dense settlement structures in Postdam city, is also applied for building extraction. The performance of our proposed network shows that PCL–PTDnet also achieves good competitive results in building extraction over very high-resolution aerial photograph. The segmented images can handle the problems of inadequate building extraction in shadow area and differentiate the texture features between building and ground. It is beneficial in enhancing building extraction results, particularly in the building shape and unclear building features. In addition, the UAV building dataset derived from UAV shows dense building and fairly sparse area with various building styles over the very high spatial resolution imagery. This dataset also evaluates the PCL–PTDnet. For this adjustment network, the increasing accuracy of comparative model can confirm the improvements of our adjustment network architecture. Its performance shows high accuracy which is the same as Inria aerial dataset and ISPRS Potsdam dataset. In comparison with other networks, our proposed network still gains robust segmented results in detecting and extracting the buildings in shadow area, unclear building features, and adjacent houses in dense building zone. Whereas the performance of other networks was affected by the various building patterns, complex structures, narrow gaps among buildings, and unique building styles. This verifies the effectiveness of our proposed network against the complex building features over very high-resolution imagery. In conclusion, in quantitative and qualitative results of three challenging datasets, it can prove that the proposed network architecture, PCL–PTDnet, can detect and extract the buildings in complex surrounding environment, shadow area, dense building area, and unclear building features more accurate than other tested architectures.

The limitation of the proposed network was that the number of model parameters is relatively large. It caused the model in computational cost and time-consuming. The complexity of network architecture is added, the large number of model parameters will be increased. This model may be difficult to segment the building shape or boundary that has similar texture feature, especially building roofs and ground. This problem could be solved by integration of DSM. Furthermore, the huge number of data samples with complex building features and styles may lead to better segmentation results.

V. Conclusion

This article demonstrates that the proposed PCL–PTDnet is a good supervised learning model in detecting and extracting building features from very high spatial resolution imageries in urban areas with dense building and shadows. Performance comparisons with other baseline networks (SegNet, U-net, PSPnet, PixelDCL, DeeplabV3+, and U-Net++) and adjustment networks (CFENet and IRU-Net) also confirm that the proposed network architecture has obvious advantage in term of extraction accuracy, and the supervised learning model can differentiate buildings under shadow and extract buildings in dense area well. The segmentation results also show an accurate building segmentation with less error on unclear building features. However, it has limitations in segmenting building shape and precise border. In the following article, we will consider to add a DSM and to integrate some postprocessing methods.

References

[1] A. Pasquinelli, G. Agugiaro, L. C. Tagliabue, M. Scaioni, and F. Guzzetti, "Exploiting the potential of integrated public building data: Energy performance assessment of the building stock in a case study in Northern Italy," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 1, 2019, Art. no. 27.

[2] Y. Zhong, A. Ma, Y. S. Ong, Z. Zhu, and L. Zhang, "Computational intelligence in optical remote sensing image processing," *Appl. Soft Comput.*, vol. 64, pp. 75–93, 2018.

[3] X. Jin and C. H. Davis, "Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 14, pp. 1–11, 2005.

[4] R. Khatami, G. Mountrakis, and S. V. Stehman, "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research," *Remote Sens. Environ.*, vol. 177, pp. 89–100, 2016.

[5] A. Mishra, T. Vu, A. V. Veettil, and D. Entekhabi, "Drought monitoring with soil moisture active passive (SMAP) measurements," *J. Hydrol.*, vol. 552, pp. 620–632, Sep. 2017, doi: 10.1016/j.jhydrol.2017.07.033.

[6] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.,* vol. 169, 2021, Art. no. 114417.

[7] Y. Qin, Y. Wu, B. Li, S. Gao, M. Liu, and Y. Zhan, "Semantic segmentation of building roof in dense urban environment with deep convolutional neural network: A case study using GF2 VHR imagery in China," *Sensors*, vol. 19, no. 5, 2019, Art. no. 1164.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[9] 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.

[10] 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, 2020.

[11] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[12] H. Gao, H. Yuan, Z. Wang, and S. Ji, "Pixel deconvolutional networks," 2017, *arXiv:1705.06820*.

[13] S. Hao, Y. Zhou, Y. Zhang, and Y. Guo, "Contextual attention refinement network for real-time semantic segmentation," *IEEE Access,* vol. 8, pp. 55230–55240, 2020.

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[15] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[16] 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.

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

[19] 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.

[20] 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.

[21] L. Luo, P. Li, and X. Yan, "Deep learning-based building extraction from remote sensing images: A comprehensive review," *Energies*, vol. 14, no. 23, 2021, Art. no. 7982.

[22] 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.
[23] Y. Ding, M. Wu, Y. Xu, and S. Duan, "P-LinkNet: LinkNet with spatial pyramid pooling for high-resolution satellite imagery," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 35–40, 2020.
[24] F. Chen, N. Wang, B. Yu, and L. Wang, "Res2-Unet, a new deep architecture for building detection from high spatial resolution images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1494–1501, 2022.
[25] Y. Lei, J. Yu, S. Chan, W. Wu, and X. Liu, "SNLRUX++ for building extraction from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 409–421, 2022.
[26] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,* vol. 14, pp. 10548–10559, 2021.
[27] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U-Net," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 71–85, 2022.
[28] M. A. A. Sheikh, T. Maity, and A. Kole, "IRU-Net: An efficient end-to-end network for automatic building extraction from remote sensing images," *IEEE Access*, vol. 10, pp. 37811–37828, 2022.
[29] J. Chen, D. Zhang, Y. Wu, Y. Chen, and X. Yan, "A context feature enhancement network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2276.
[30] G. Wu et al., "A boundary regulated network for accurate roof segmentation and outline extraction," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1195.
[31] G. Yang, Q. Zhang, and G. Zhang, "EANet: Edge-aware network for the extraction of buildings from aerial images," *Remote Sens.*, vol. 12, no. 13, 2020, Art. no. 2161.
[32] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.
[33] W. Liu et al., "Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2912.
[34] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.
[35] ISPRS, "2D Potsdam Semantic Labelling Dataset," 2018. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

Wuttichai Boonpook received the B.A. degree in geography from Thammasat University, Bangkok, Thailand, in 2009, the B.B.A degree in international business from Ramkhamhaeng University, Bangkok, in 2009, the M.Sc. degree in spatial information system in engineering from Chulalongkorn University, Bangkok, in 2013, and the D.Eng. degree in traffic and transportation engineering from Beihang University, Beijing, China, in 2019.

He is currently a Lecturer with the Department of Geography, Faculty of Social Science, Srinakharinwirot University, Bangkok. His research interests include GIS and remote sensing applications, deep learning semantic segmentation for remote sensing data, and UAV remote sensing.

Yumin Tan received the B.Sc. degree in surveying from the Guilin University of Technology, Guangxi, China, in 1998, the M.Sc. degree in geodesy from the Tsinghua University, Beijing, China, in 2001, and the Ph.D. degree in GIS from the Chinese Academy of Science, Beijing, China, in 2004.

She is currently a Professor of transportation engineering, and the Dean of the Department of Civil Engineering. Her research interests include intelligent remote sensing data analysis, UAV remote sensing, and trust evaluation of remote sensing products.

Kritanai Torsri received the B.Sc. degree in mathematics from Thaksin University, Songkhla, Thailand, in 2002, fully funded by the Ministry of University Affairs, the M.Sc. degree in environmental technology from the Joint Graduate School of Energy and Environment, King Mongkut's University of Technology Thonburi, Bangkok, Thailand, in 2012, and the Ph.D. degree in meteorology from the Institute of Atmospheric Physics, Chinese Academy of Science (CAS), Beijing, China, in 2022, granted by the CAS and the World Academy of Sciences (CAS-TWAS) President's Fellowship Programme.

He is currently a Researcher (Expert Level) with the Climate and Weather Section, Hydro-Informatics Innovation Division, Hydro-Informatics Institute, Ministry of Higher Education, Science, Research and Innovation (MHESRI), Bangkok. His research interests focus on hydro-climate variability, extremes, and their predictions using dynamical models and deep learning methods.

Patcharin Kamsing received the B.Eng. degree in telecommunication engineering from the King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, in 2009, and the M.Eng. degree in space technology and its applications with major in satellite communication (SATCOM) from Beihang University, Beijing, China, in 2013, and the Ph.D. degree in cartology and geographic information sciences from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2016.

She is currently a Lecturer with the International Academy of Aviation Industry, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. Her research interests include space technology, such as satellite communication, remote sensing, and space object tracking systems.

Peerapong Torteeka received the B.E. degree in mechatronics engineering from the Mahanakorn University of Technology (MUT), Bangkok, Thailand, in 2007, the 1st M.E. degree in control and instrument engineering from the MUT, in 2011, and the 2nd M.E. degree in global navigation satellite system engineering from the Beihang University, Beijing, China, in 2014, and the Ph.D. degree in celestial mechanic and astrometry engineering from the National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China, in 2018.

In 2019, he joined the Research Group, National Astronomical Research Institute of Thailand (Public Organization), as a Research engineer and a Project Manager with the Thai Space Consortium Pathfinder Satellite. His research interests include guidance, navigation and control system and space system engineering. In recent years, his research interests include processing techniques of sensor fusion, nonlinear parameter estimation, and control system engineering.

Attawut Nardkulpat received the B.Sc. degree in geography, and the M.Sc. degree in geoinformatics from the Burapha University, Chon Buri, Thailand, in 2013 and 2017, respectively. He is currently working toward the Ph.D. degree in geoinformatics with the Department of Geography, Faculty of Social Sciences, Srinakharinwirot University, Bangkok, Thailand.

He is currently a Research Officer with the Faculty of Geoinformatics, Burapha University, Chon Buri, Thailand. His research interests include GIS, remote sensing, deep learning, machine learning, and natural language processing.