IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 13, 2019, accepted December 29, 2019, date of publication January 8, 2020, date of current version January 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964760

Classification of Very High-Resolution Remote Sensing Imagery Using a Fully Convolutional Network With Global and Local Context Information Enhancements

HUANJUN HU[1], ZHENG LI[2], LIN LI[1,3], HUI YANG[1,4], AND HAIHONG ZHU[1]

[1]School of Resource and Environment Sciences, Wuhan University, Wuhan 430079, China
[2]Hubei Institute of Land Surveying and Mapping, Wuhan 430010, China
[3]RE-Institute of Smart Perception and Intelligent Computing, Wuhan University, Wuhan 430079, China
[4]Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China

Corresponding authors: Lin Li (lilin@whu.edu.cn) and Hui Yang (19050@ahu.edu.cn)

This work was supported in part by the National Natural Science Fund of China under Grant 41871298, Grant 41271443, Grant 41571395, Grant 41471328, and Grant 41971311, in part by the National Key Research and Development Program of China under Grant 2016YFB0501403, in part by the Scientific and Technological Leading Talent Fund of the National Administration of Surveying, Mapping and Geo-information (2014), in part by the National Natural Science Fund of Anhui Province, China, and in part by the Major Program of Science and Technology of Anhui Province under Grant 18030801111.

ABSTRACT Deep learning methods for semantic image segmentation can effectively extract geographical features from very high-resolution (VHR) remote sensing images. However, these methods experience over-segmentation in low-level features and a breakdown in the integrity of objects with fixed patch sizes due to the multi-scaled geographical features. In this study, a dual attention mechanism is introduced and embedded into densely connected convolutional networks (DenseNets) to form a dense-global-entropy network (DGEN) for the semantic segmentation of VHR remote sensing images. In the DGEN architecture, a global attention enhancement module is developed for context acquisition, and a local attention fusion module is designed for detail selection. This network presents the improved semantic segmentation performance of test ISPRS 2D datasets. The experimental results indicate an improvement in the overall accuracy (OA), F1, kappa coefficient and mean intersection over union (MIoU). Compared with the DeeplabV3+ and SegNet models, the OA improves by 2.79% and 1.19%; the mean F1 improves by 3.43% and 0.88%; the kappa coefficient improves by 4.04% and 1.82%; and the MIoU improves by 5.22% and 1.47%, respectively. The experiments showed that the dual attention mechanism presented in this study can improve segmentation and maintain object integrity during the encoding-decoding process.

INDEX TERMS Attention mechanism, DenseNet, semantic segmentation, very high-resolution remote sensing images.

I. INTRODUCTION

Very high-resolution (VHR) remote sensing images have become an important source of information for Earth surface monitoring. The semantic segmentation of VHR remote sensing images has been widely studied [1], [2] due to its outstanding performance in the classification of VHR remote sensing images. However, the finer spatial resolution in VHR leads to an increase in the intra-group variability between objects and a decrease in the inter-class variability between different classes, thus reducing the statistical separability of different land cover classifications in the spectral domain and impacting the accuracy of the classification [3], [4].

In addition, the limited spectral resolution of VHR sensors further increases the complexity of the semantic segmentation of VHR images [4]. Compared with low-resolution images, VHR images contain more ground details and have lower spectral resolution. Generally, VHR images have only red-green-blue (RGB) channels, and some contain near-infrared (NIR) channels [5]. In a VHR image, the information required for classification cannot be completely captured by the spectral intensity. The texture and spatial contexts become very

The associate editor coordinating the review of this manuscript and approving it for publication was Hiram Ponce.

important to the dense classification of each pixel. Therefore, many conventional methods focus on extracting features from the spatial neighborhood of pixels [1], [3], [6]–[18].

In general, object-oriented methods have been used for VHR image classification [1], [18] and mainly include two processes: segmentation and classification. During the segmentation process, multi-resolution (MR) [19], mean shift [20] and other image segmentation methods are applied for image segmentation; during the classification process, the computed target features (such as color, texture, and geometric features) are used as the inputs for supervised classification or unsupervised classification, or feature transformation rule sets are designated to achieve recognition and classification [21].

These kinds of models are constructed by features, which are artificially extracted, and are also limited to those features, since the various ground objects in the real data may not be classified using a specific set of features [22].

Deep learning provides another way to effectively identify features from the training set [22]. Deep learning enables the use of both supervised and unsupervised feature learning from very large raw image data sets [23]. In recent years, deep learning architectures have achieved great progress in the fields of computer vision, audio recognition, and natural language processing. Influenced by deep learning, many computational visual deep learning architectures have been used in remote sensing image analysis tasks [5], [24]–[27], and remarkable progress has been made.

Convolutional neural network (CNN) models based on deep learning theory are widely used in the semantic segmentation of VHR remote sensing images. The semantic segmentation models for VHR remote sensing images based on CNNs mainly include patch-based models [28]–[30] and fully convolutional network (FCN) [2], [27], [28], [31], [32]. Patch-based models were among the first deep learning semantic segmentation models developed for remote sensing images and predict each pixel by evaluating a region surrounding each pixel. This kind of model is faced with the problems of computing redundancy and the loss of edge information [2], [27].

Furthermore, FCN models have become the mainstream in the deep learning semantic segmentation of remote sensing images [31]. For example, Kampffmeyer *et al.* [28] proposed a CNN method to extract small targets from VHR remote sensing images. Marmanis *et al.* [5] used the U-net model for semantic image segmentation. Sherrah [27] proposed high-resolution aerial semantic image segmentation based on an FCN, which uses the dilated convolution of DeepLab [33] to enlarge the convolution receptive field. As upgrades, DeeplabV3 and DeeplabV3+ [34] are proposed with better performance based on DeepLab. Marmanis *et al.* [2] added boundary detection to the structure of SegNet [35] to improve the performance of the semantic segmentation of VHR remote sensing images. Audebert *et al.* [32] used deep convolution networks to fuse multimodal and multiscale remote sensing data for the semantic segmentation of VHR remote sensing images and effectively compensated for the lack of spectral information in the VHR data to achieve better semantic segmentation results. In addition, Xu *et al.* [36] also designed CNNs to extract buildings from VHR images and achieved good results. Yao *et al.* [37] proposed DCCN model to improve the classification by enhancing the object boudaries using coordinate convolution and achieved very good results.

In principle, an FCN model contains a downsampling process, an upsampling process and skip connections [38]. The downsampling process is composed of convolution layers and pooling layers and is employed to extract multilevel discriminative feature maps and generate low-resolution feature maps with improved discrimination. The upsampling process is composed of convolution layers and deconvolution layers. This process mainly restores the resolution of the feature maps generated by the downsampling process and generates segmentation maps with the same resolution as the original image. Downsampling can obtain a large context and generate discriminative feature maps by pooling layers to enlarge the receptive field of each pixel, but it also leads to losses of high-frequency details and spatial information of the object. Therefore, deconvolution or interpolation is applied to restore the spatial resolution of the downsampled feature map during upsampling, but deconvolution or interpolation cannot provide the precise localization of boundaries and high-frequency details, which significantly impacts the accuracy of semantic segmentation [38]. Skip connections help retain high-frequency details during the upsampling process and improve the accuracy of semantic segmentation by fusing the features generated during the early downsampling process with the highly discriminant feature maps from the deeper layer [32].

For the abundant detailed information regarding objects in VHR remote sensing images, such as roof materials and skylights, skip connections are used to recover many feature details, which will lead to the introduction of a large number of redundant low-level features. However, this process causes a decrease in the discrimination of high-level features [36].

For the widespread continuous objects in VHR remote sensing images, global context information can effectively improve the perception of the scene and improve the consistency of pixel classification [39]. In deep neural networks, the size of the receptive field can roughly indicate the degree of usage of contextual information. However, the literature [40] notes that the practical receptive field of a CNN is much smaller than the theoretical receptive field, especially in high-level layers. Additionally, to facilitate deep learning network training, remote sensing images are usually divided into fixed size patches (such as 256 × 256 size patches). This process will separate pixels from the same objects and lead to the loss of image context information, which requires further tiling and stitching. These conditions result in many networks that are unable to fully integrate important global context information.

To solve these problems, namely, (1) the mis-discrimination of the features of a skip connection and (2) context information losses of the FCN network when processing the VHR remote sensing images, in this paper, a novel semantic segmentation network for VHR remote sensing images called the dense-global-entropy network (DGEN) is proposed. In this network, the structure is based on a densely connected convolutional network (DenseNet) [41] to extract multilevel features due to the powerful capacities of feature extractions and information reuses [42]. Moreover, due to the capabilities of selecting the most informative components of the input image while suppressing the noise and background of deep-learning-based attention mechanisms [43], [44], which have attracted wide attention in the field of computational vision [39], [43], [45]–[47] and remote sensing scene classification [31], [48]–[52], the global attention enhancement module and the local attention fusion module (see Figure 2, 3 for details) are designed and embedded into the network to enhance the global context information and recover the mis-discrimination of objects. The global attention enhancement module is used to strengthen the global context information of highly discriminant features by enlarging the receptive fields in upsampling. The local attention fusion module assigns weights by information entropy to weaken the background information and obtain useful local detailed information.

The remainder of this paper is organized as follows. A detailed description of the proposed DGEN is given in Section II. The designment and metrics of the experiments are provided in Section III. The results and comparisons are given in Section IV. The discusstions are given in Section V. Finally, conclusion is drawn in Section VI.

II. METHODOLOGY

A. DGEN MODEL

In the DGEN, an encoding-decoding architecture is designed for the semantic segmentation of VHR remote sensing images based on DenseNet. The encoder part of the DGEN is mainly used to extract multilevel features from the input image. The encoder is built from the convolutional layers, the dense blocks, the transition layers, and the downsampling operation. The decoder part aims to enlarge the feature maps extracted by the encoder to produce a final segmentation map with the same resolution as the input image. The decoder is built from the convolutional layers, the nonlinear activation layers, and the upsampling operation. The decoder gradually recovers the details and spatial information of highly discriminant feature maps by using operations, such as the convolutional layers, the global attention enhancement modules and the local attention fusion modules. The overall structure of the DGEN proposed in this paper is shown in Figure 1.

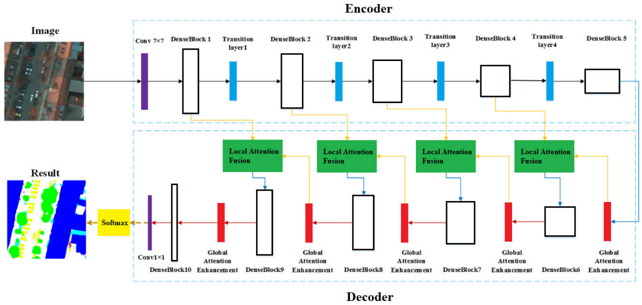

FIGURE 1. The overview architecture of the dense-global-entropy network (DGEN).

The DenseNet used in the DGEN encoder is different from the traditional network structure. In a traditional CNN, the convolutional network with L layers has L connections. Then, DenseNet has L (L + 1) / 2 direct connections. DenseNet uses dense connectivity to further improve the flow of information between layers. For each layer, the feature maps of all preceding layers are used as inputs, and its own feature maps are also used as inputs for all subsequent layers. In this way, each layer can access the loss gradient formed between the end and the beginning of the model during training. This connection pattern effectively improves the flow of information between layers, alleviates the problem of the disappearing gradient of the deep CNN, and makes the model easy to train. In addition, this connection pattern can also improve the reuse of features. However, the downsampling operation is very important for a deep CNN, and downsampling will lead to a change in the size of the feature map. The concatenation operation between feature maps of different sizes cannot be completed. Therefore, DenseNet divides the networks into multiple densely connected dense blocks (DenseBlocks). In each DenseBlock, all layers maintain dense connectivity; individual DenseBlocks are connected through the transition layers. In a single DenseBlock, the function F_l (.)(batch normalization -ReLU- convolution (3 × 3)) is used for the nonlinear transition between layers. The density connection can be defined as follows:

$$X_l = F_l\left([X_0, X_1, \ldots, X_{l-1}]\right) \tag{1}$$

where X_l is the output feature map of layer 1 and $[X_0, X_1, \ldots, X_{l-1}]$ is the concatenation of the feature maps of all preceding l layers. This concatenation DenseNet generates too many feature maps. To effectively control the parameters of the model, DenseNet defines a growth rate (K) to control the number of feature maps.

The transition layer is mainly used for convolution and pooling. Each transition layer is composed of a batch normalization, a ReLU, and an average pooling layer (2 × 2).

Generally, the decoder part of the encoder-decoder architecture uses a skip connection to recover the detailed information during the process of generating a final segmentation map with the same resolution as the input image. Unlike other encoder-decoder architectures, we use the global attention enhancement module to enhance the global context information of the highly discriminant features in the decoder. Subsequently, the deconvolution layers are used to expand the sizes of these feature maps, and the expanded feature maps and the low-level feature maps are used as inputs for the local attention fusion module. In the local attention fusion module, the highly discriminant feature graph, which incorporates the global context information as the local feature fusion, is used to guide the selection of low-level features to recover the high-frequency detailed information. Then, the features that fuse global context information, restore spatial information and identify local details are used as the inputs for the DenseBlock for further processing. Finally, we use the 1 × 1 convolution to map the feature maps into different classifications.

Compared to traditional FCN structures that directly use skip connection in upsampling to recover details, the proposed DGEN model (see details in Figure 1) uses the global attention enhancement module and local attention fusion module instead of upsampling processes. Therefore, the global attention enhancement module enhances the global information to further improve classification consistency. The enhanced global features are then used as bases of weighting for local information and are inputted to select the high-frequency low-level features to recover the detailed information and avoid information redundancy due to the skip connection.

B. THE GLOBAL ATTENTION ENHANCEMENT

Global average pooling is the simplest way to obtain global context information. The global average pooling operation has been used by SENet [47], the pyramid attention network [53] and other networks to extract the global features of an image. The global information enhancement module also uses the global average pooling layer to extract the global context information from the feature maps and enhance the global information of the feature maps. The architecture of the global information enhancement module is shown in Figure 2. In the global information enhancement module, first, we use a global pooling layer to extract the global context information from high-level feature maps. Then, the output of the global context information is activated by a sigmoid layer to be used as the weight of the feature maps. Finally, the weighted features are added to the feature maps to obtain the feature maps that integrate the global information.

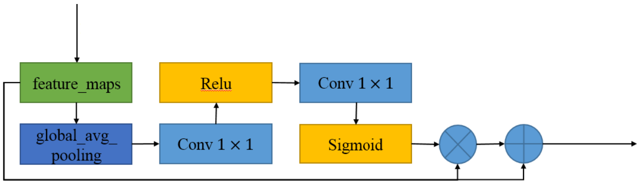

FIGURE 2. The architecture of the global attention enhancement module.

C. THE LOCAL ATTENTION FUSION

In deep CNNs, the feature maps acquired in the early layer with rich spatial location information are less discriminating than those acquired in the subsequent layers. Such feature maps are directly passed to the decoder by the skip connection and may increase the ambiguity of the final result. The attention mechanism is a tool used to assign resources to the most informative part of the input information. To prevent location information from weakening the classification information, we designed a local attention fusion module inspired by the attention mechanism. Information entropy is an index used to measure the level of information clutter. A higher value of information entropy indicates that the information is more uncertain and more chaotic. In contrast, a lower value of information entropy shows that the information is more defined and more stable. Therefore, we can use the information entropy of the pixels over highly discriminative features to guide the selection of low-level features. That is, when pixels have high entropy values, the information is unstable; that is, the more low-level features need to be fused in the highly discriminative feature maps to help restore the details and improve the accuracy of semantic segmentation. In contrast, less low-level information needs to be integrated. In the local attention fusion module, the high discriminative features map is first mapped to different classifications using a 1 × 1 convolution, and the output of the classification is activated by a SoftMax layer. The output of the activation is normalized to [0, 1]. Subsequently, the information entropy of each pixel is calculated by Equation 2, and this value is used as the weight of the low-level features. Then, the information entropy of each pixel is multiplied by low-level features to obtain the weighted low-level feature. Finally, weighted low-level feature maps are added to the highly discriminative feature maps as inputs for DenseBlock to gradually recover detailed object information. The architectures of the local information fusion module are shown in Figure 3.

$$\mathrm{H}\,(\mathrm{x}) = \mathrm{E}\left[-log_2\,(p_i\,(\mathrm{x}))\right] = -\sum\nolimits_{j=1}^{n} p_i(x) * log_2\,(p_i\,(\mathrm{x})) \tag{2}$$

where $\mathrm{E}\left[-log_2\,(p_i\,(\mathrm{x}))\right]$ is the information entropy of $p_i(x)$, and $p_i(x)$ is the probability of pixel x belonging to category j.

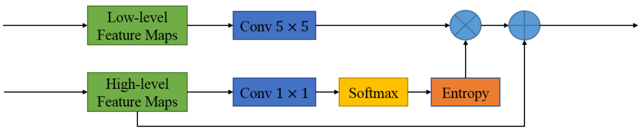

FIGURE 3. The architectures of the local attention fusion module.

D. IMPLEMENTATION

The DGEN contains 10 dense blocks in the encoder and decoder, including 52 convolution layers. The detailed architectures of DGEN are shown in Figure 4. Following the input of the network, we use a convolution layer with 16 convolution kernels of 7 × 7 to generate the initial feature map. In the encoder, the transition_down module is built from a 1 × 1 convolution layer, a drop_out layer and a 2 × 2 average pooling layer with stride=2. In the decoder, we use the 3 × 3 "transposed_convolution" layer with stride=2 to upsample.

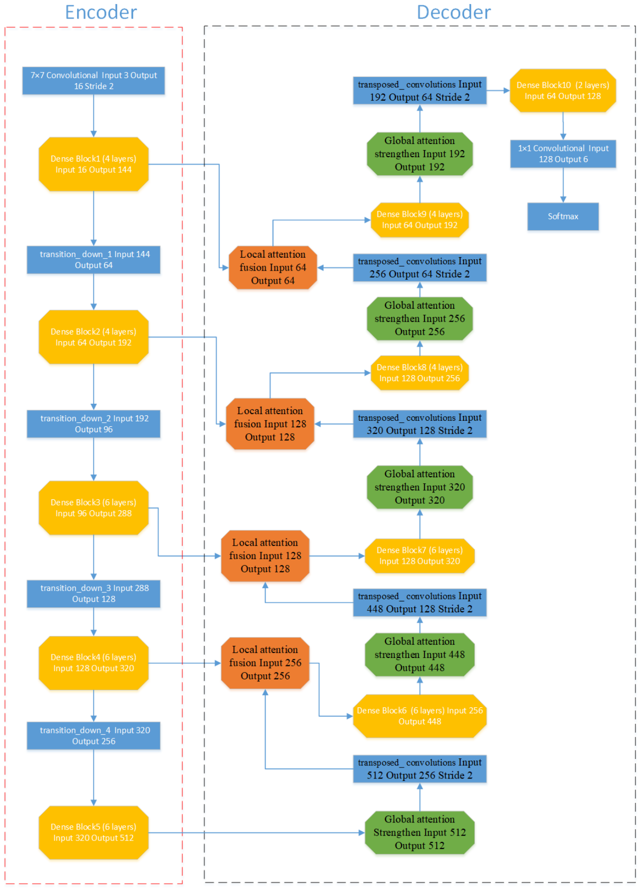
Encoder
Decoder
7×7 Convolutional Input 3 Output 16 Stride 2
Dense Block1 (4 layers) Input 16 Output 144
transition_ down_1 Input 144 Output 64
Dense Block2 (4 layers) Input 64 Output 192
transition_down_2 Input 192 Output 96
Dense Block3 (6 layers) Input 96 Output 288
transition_down_3 Input 288 Output 128
Dense Block4 (6 layers) Input 128 Output 320
transition_down_4 Input 320 Output 256
Dense Block5 (6 layers) Input 320 Output 512
Local attention fusion Input 64 Output 64
Local attention fusion Input 128 Output 128
Local attention fusion Input 128 Output 128
Local attention fusion Input 256 Output 256
transposed_ convolutions Input 192 Output 64 Stride 2
Global attention strengthen Input 192 Output 192
Dense Block9 (4 layers) Input 64 Output 192
transposed_ convolutions Input 256 Output 64 Stride 2
Global attention strengthen Input 256 Output 256
Dense Block8 (4 layers) Input 128 Output 256
transposed_ convolutions Input 320 Output 128 Stride 2
Global attention strengthen Input 320 Output 320
Dense Block7 (6 layers) Input 128 Output 320
transposed_ convolutions Input 448 Output 128 Stride 2
Global attention strengthen Input 448 Output 448
Dense Block6 (6 layers) Input 256 Output 448
transposed_ convolutions Input 512 Output 256 Stride 2
Global attention Strengthen Input 512 Output 512
Dense Block10 (2 layers) Input 64 Output 128
1×1 Convolutional Input 128 Output 6
Softmax

FIGURE 4. Detailed architecture of DGEN.

At the end of the network, we use a 1 × 1 convolution layer and a SoftMax classifier to output the final prediction. The growth rate of each dense block is set to 32.

III. EXPERIMENTS

A. DATASET

In this paper, we use VHR remote sensing images from the International Society for Photogrammetry and Remote Sensing (ISPRS) 2D Semantic Labeling Challenge. The ISPRS 2D datasets include data from Potsdam, Germany, and other cities. Potsdam is a typical historic city with large building blocks, narrow streets and dense settlement structures. Each dataset is manually classified into six of the most common land cover classes. The Potsdam dataset contains 38 images of 6000 × 6000, and 24 of these images have label data. To verify the validity of our model, we divide the labeled data into two parts: 80% for a training set and 20% for a test set. Of the 24 labeled images of Potsdam, 5 images with variety and balance on ground object type, which listed as 4_12, 5_11, 5_12, 6_12 and 7_11, are selected for testing set, and other 19 images are used in training set (see in Table 1). To facilitate training, we cut the labeled datasets into 256 × 256, 320 × 320 and 448 × 448 sections. Finally, 21527 training samples and 5326 test samples are generated.

TABLE 1. Image selection of ISPRS 2D semantic labeling data - Potsdam.

Traning set		Testing set
top_potsdam_2_10_label	top_potsdam_6_8_label	top_potsdam_4_12_label
top_potsdam_2_11_label	top_potsdam_6_9_label	top_potsdam_5_11_label
top_potsdam_2_12_label	top_potsdam_6_10_label	top_potsdam_5_12_label
top_potsdam_3_10_label	top_potsdam_6_11_label	top_potsdam_6_12_label
top_potsdam_3_11_label	top_potsdam_7_7_label	top_potsdam_7_11_label
top_potsdam_3_12_label	top_potsdam_7_8_label	
top_potsdam_4_10_label	top_potsdam_7_9_label	
top_potsdam_4_11_label	top_potsdam_7_10_label	
top_potsdam_5_10_label	top_potsdam_7_12_label	
top_potsdam_6_7_label		

B. TRAINING

Our experiments are based on TensorFlow. The "variance_scaling_initializer" is used to initialize our model, and the Adam optimizer with an initial learning rate of 0.001 is used to optimize the network when adjusting parameters, such as weights and biases. The rate of drop-out for all "drop_out" layers is set to 0.2. In addition, the number of feature maps generated by each convolution layer in dense blocks is set to 32. There are 150 epochs during training, and each epoch has 2000 iterations. The batch of training samples (batch) is set to 12. The graphics processing unit (GPU) used for training models is GTX 1080Ti.

C. METRICS

To test the performance of the proposed model, the pixel accuracy (PA), F1 score, kappa coefficient and mean intersection over union (MIoU) are chosen to evaluate the models.

1) PIXEL ACCURACY (PA)

The *PA* is a general statistic evaluation metric for accuracy. In this paper, this metric measures the precision of matched pixels, including the foreground and background. The equation is as follows:

$$PA = \frac{p_f + p_b}{N} \tag{3}$$

where P_f and P_b represent the positive numbers for the foreground and background at the pixel level, respectively, and N is the number of pixels in the test image.

2) F_1 SCORE

The F_1 score calculates through the precision (P) and recall (R) and is a powerful evaluation metric for the harmonic mean of P and R. The F_1 score can be calculated as follows:

$$F_1 = \frac{2*P*R}{P+R} = \frac{2*TP}{2*TP+FN+FP} \tag{4}$$

$$R = \frac{TP}{TP+FN} \tag{5}$$

$$P = \frac{TP}{TP+FP} \tag{6}$$

where R represents the proportion of matched pixels in the ground truth and P is the ratio of matched pixels in the prediction results. *TP, FP* and *FN* represent the number of true positives, false positives and false negatives, respectively.

Compared with *PA*, the F_1 score is more impacted by the smaller one in P and R. Under the same PA condition, a higher F_1 score requires the balance between precision and recall. The shortage of either precision or recall results in a more dramatic decrease in the F_1 score than in the *PA*.

3) KAPPA COEFFICIENT

The kappa coefficient is a common metric of a consistency check. This metric can also be used to evaluate the accuracy of semantic segmentation. Based on the confusion matrix, the kappa coefficient can be formulated as:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{7}$$

where p_0 is the proportion of units in which the judges agree and p_e is the proportion of units for which agreement is expected by chance.

Assume n pixels are expected to be classified into m classes, set the ground truth numbers of pixels for each class are $a_1, a_2, \cdots, a_m$; the classified numbers of pixels for each class are $b_1, b_2, \cdots, b_m$; the correct classified numbers of pixels of each class are $c_1, c_2, \cdots, c_m$. Thus:

$$p_0 = \frac{\sum_{i=1}^{m} c_i}{n} \tag{8}$$

$$p_e = \frac{\sum_{i=1}^{m} a_i \times b_i}{n \times n} \tag{9}$$

In the semantic segmentation context, p_0 and p_e can also be formulated as:

$$p_0 = \sum\nolimits_{i=1}^{m} (\frac{TP_i}{n}) \tag{10}$$

$$p_e = \sum\nolimits_{i=1}^{m} (\frac{TP_i + FP_i}{n} \cdot \frac{TP_i + FN_i}{n}) \tag{11}$$

where m is the number of expected classes; *TP, FP* and *FN* represent the number of true positives, false positives and false negatives, respectively.

Basically, the kappa coefficient changes between 0 and 1. A larger value indicates better consistency.

4) MEAN INTERSECTION OVER UNION (MIoU)

MIoU is a standard metric for segmentation (Garcia-Garcia et al., 2017). IoU is to ratio the intersection and the union of the ground truth set and the predicted segmentation set. The MIoU is the average value of the IoU of each class. The MIoU in semantic segmentation can be formulated as follows:

$$\text{MIoU} = \frac{1}{m} \sum (\frac{A_{pred} \cap A_{true}}{A_{pred} \cup A_{true}}) \tag{12}$$

where m is the number of expected classes; A_{pred} is the prediction set, and A_{true} is the set of corresponding ground truth.

The MIoU is based on sets. The sets are regions of pixels of the same class on the image, which also means objects, instead of some mean value of the single pixels. The sets make the MIoU an object-oriented metric instead of pixel-based. The MIoU can reflect the overlap ratio of predicted objects and the corresponding object of ground truths.

IV. RESULTS AND COMPARISONS

A. RESULTS

After 150 epochs of iterations, our DGEN achieves state-of-the-art results on the datasets (see Table 2). All results are listed based on infrared, green and blue (IrGB) images without any preprocessing or post-processing.

TABLE 2. The PA for each class, the OA, F1 score, Kappa coefficient (a) for each class, and the MIoU (b) for the classification of all validation datasets by the DGEN model.

Datasets	Buildings	Imp surf	Tree	Low_veg	Car	**OA**	***F1***	**Kappa**
All	**95.40%**	**91.23%**	**79.06%**	**78.46%**	**88.89%**	**89.31%**	**86.61%**	**84.47%**
Image1	95.10%	92.16%	79.58%	86.20%	88.63%	89.61%	88.33%	85.68%
Image2	95.18%	89.85%	83.33%	79.10%	89.04%	88.40%	87.30%	83.85%
Image3	95.51%	89.44%	76.48%	73.11%	87.35%	89.05%	84.38%	83.09%
Image4	96.00%	92.02%	82.12%	76.56%	91.33%	90.44%	87.61%	85.93%
Image5	95.19%	92.70%	73.79%	77.33%	88.08%	89.06%	85.42%	83.82%

(A)

	***IoU* (%)**					**Mean *IoU* (%)**
	Buildings	Imp surf	Tree	Low veg	Car	
All	**91.2**	**83.91**	**65.51**	**64.77**	**80.02**	**77.08**
Image1	90.66	85.47	66.09	75.74	79.59	79.51
Image2	90.8	81.57	71.43	65.42	80.24	77.89
Image3	91.4	80.9	61.92	57.62	77.54	73.88
Image4	92.31	85.23	69.66	62.02	84.04	78.65
Image5	90.83	86.39	58.47	63.05	78.7	75.49

(B)

After 150 epochs of iterations, our DGEN achieves state-of-the-art results on the datasets (see Table 2). All results are listed based on infrared, green and blue (IrGB) images without any preprocessing or post-processing.

Our model achieves high scores for all five validation images under the 4 metrics mentioned above. Specifically, the overall accuracy (OA) of a pixel reaches 89.31%, which is a good general PA; the mean F1 score reaches 86.61%, which indicates that both precision and recall are satisfactory; the kappa coefficient reaches 74.47%, which shows good consistency; and the MIoU reaches 77.08%. These data show that the proposed DGEN performs well for land cover classification from multiple perspectives.

B. PERFORMANCE OF THE DGEN

1) STATISTICAL COMPARISON

To demonstrate the performance of the DGEN, we also implemented U-net [5], DeeplabV3+ [34] and SegNet [35] models as comparisons to quantify the improvement. The results of this comparison are listed in Table 3. For this dataset, the results of the U-net lack performance and are 20%-35% lower than all the selected metrics. Therefore, the U-net will not be discussed in the following detailed analysis of this part.

TABLE 3. The comparison of the OA, F1 score, kappa coefficient (A) and (B) for the classification of all validation datasets by the DeeplabV3+, SegNet and DGEN models.

Method	Building	Imp surf	Tree	Low_veg	Car	**OA**	**Mean *F1***	**Kappa**
U-net	62.10%	72.05%	36.30%	57.66%	43.95%	61.28%	54.41%	45.73%
DeeplabV3+	92.39%	88.20%	75.70%	77.11%	82.48%	86.52%	83.18%	80.43%
SegNet	93.04%	89.99%	78.98%	**79.77%**	86.88%	88.12%	85.73%	82.65%
DGEN	**95.40%**	**91.23%**	**79.06%**	78.46%	**88.89%**	**89.31%**	**86.61%**	**84.47%**

(A)

	***IoU* (%)**					**Mean *IoU* (%)**
	Building	Imp surf	Tree	Low_veg	Car	
U-net	45.17	56.46	22.43	41.08	28.19	38.67
DeeplabV3+	85.95	79.01	60.99	63.05	70.28	71.86
SegNet	87.12	81.83	**65.59**	**66.69**	76.82	75.61
DGEN	**91.2**	**83.91**	65.51	64.77	**80.02**	**77.08**

(B)

According to Table 3, DeeplabV3+ shows scores of 86.52% for OA, 83.18% for F1, 80.43% for the kappa coefficient and 71.86% for the MIoU. The DGEN by all selected metrics outperforms the DeeplabV3+ model. The overall PA of DGEN improved 2.79% compared with the DeeplabV3+ model. The mean F1 score improved by 3.43%, which means that both precision and recall of the DGEN are better balanced and have less shortage. The kappa coefficient improved by 4.04%, which means a better classification consistency. The MIoU improved by 5.22%, which is much more than the improvement in the OA. This result means that the improvement in the object-oriented point of view is even more than the pixel-based point of view.

The results of SegNet show an 88.12% OA, an 85.73% F1, an 82.65% kappa coefficient and a 75.61% MIoU. The SegNet model exhibits better performance than DeeplabV3+ in this experiment. However, according to Table 3, our DGEN model still generally outperforms SegNet in every metric that we used. Compared to the results of SegNet, the overall PA of the DGEN improved by 1.19%. The mean F1 score improved by 0.88%, which means that both the precision and recall of the DGEN are slightly better balanced and have less shortage. The kappa coefficient improved by 1.82%, which means a better classification consistency. The MIoU improved by 1.47%, which means an improvement in the object-oriented point of view. Overall, although with an approximate performance for trees and a shortage for low vegetation, the proposed DGEN proved better performance with respect to the OA, F1, kappa coefficient and mean IoU than SegNet.

The overall performance improvements of the DGEN from DeeplabV3+ and SegNet by the selected metrics are shown in Figure 5.

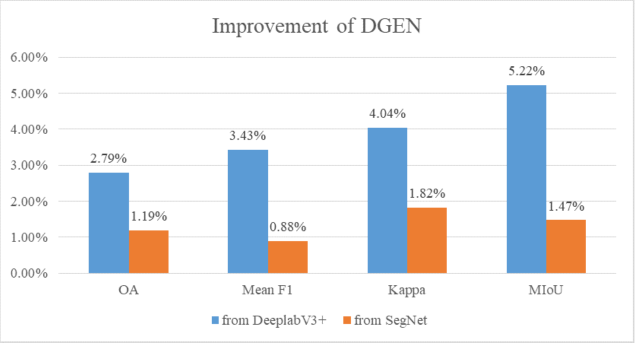

FIGURE 5. Improvement in the DGEN from DeeplabV3+ and SegNet by the OA, mean F1, kappa coefficient and mean IoU models.

2) VISUAL INSPECTION

A visual inspection and comparison of the classification results by DeeplabV3+, SegNet and proposed DGEN are presented in Figure 6. The columns of pictures are the 5 original images, their DeeplabV3+ results, the SegNet results, the proposed DGEN results and the corresponding ground truth results.

All 3 models have their gains and losses. Generally, the DGEN model performs the best, the SegNet follows and DeeplabV3+ takes the third place, which supports the statistical results in Table 3.

Specifically, the 3 models all have some misclassified objects (see Figure 7-8). The visible misclassified areas are basically the large-flat roofs of buildings. All 3 models are occasionally misclassified as roofs impervious by over-segmentation details.

However, despite the closeness of the statistical results by metrics, the object integrity and boundary quality of the DGEN is visibly improved.

In Figure 9-10, we can say that DGEN performs best with respect to object integrity. However, it is worth noting that

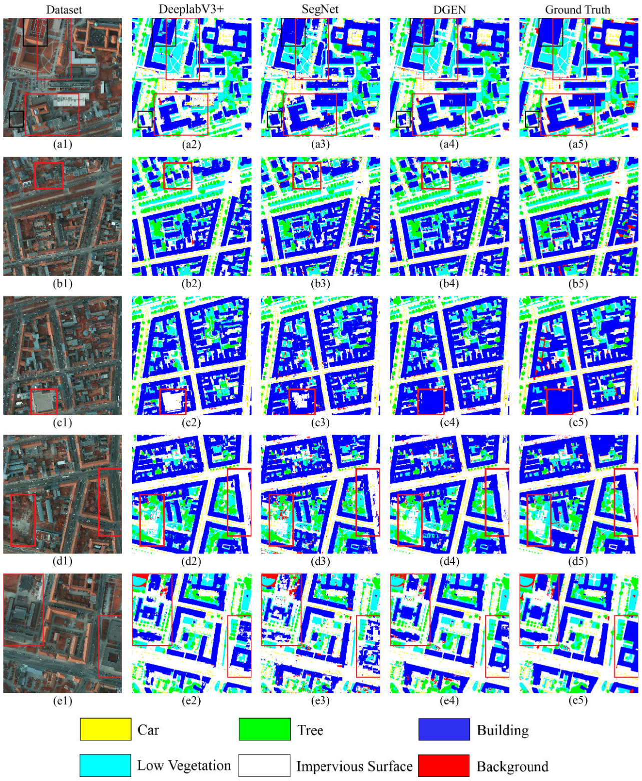

FIGURE 6. The results of the classification from DeeplabV3+, SegNet and proposed DGEN.

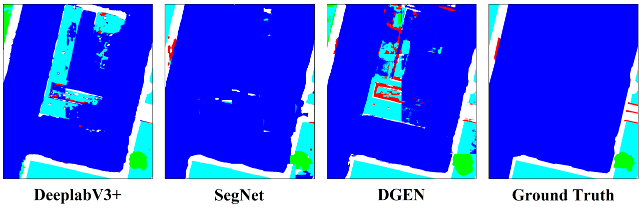

FIGURE 7. Misclassified in the top-left black box area of Figure 5 (a).

although SegNet performs very similarly to DGEN by statistical metrics (less than 2%), the object integrities of SegNet are hardly applicable. The buildings in Figures 8-9 of SegNet are nearly completely broken. The shapes and boundaries of these buildings are basically unrecognizable.

Moreover, the dissociated spots are also visible on SegNet (see Figure 5 (a, d, e)). Figure 11-12 shows the high-density regions of the dissociated spots in the SegNet results.

Furthermore, the DGEN model also significantly improves the boundary quality. As shown in Figure 13-15, the rough boundaries of DeeplabV3+ and/or SegNet are restored by DGEN.

In summary, our DGEN model is proved improvements by both statistical metrics and visual inspection. In addition to the improvements in the classification accuracies, the object integrities and boundary qualities are restored and guaranteed by our novel DGEN model. Notably, the statistical results do not always represent the actual classification performance. For example, in this experiment, the results of SegNet show high scores but low integrities. Therefore, the visual inspection of the images and results is still necessary.

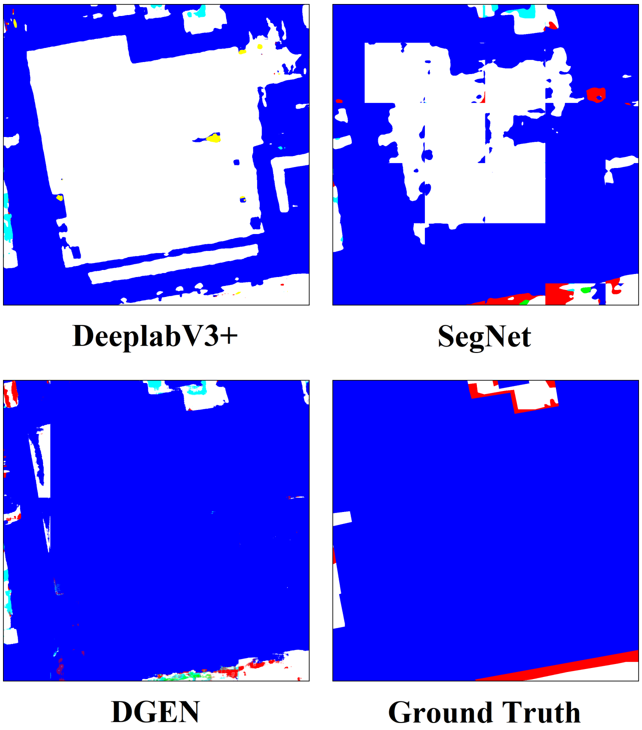

FIGURE 8. Misclassified in the bottom red box area of Figure 5 (c).

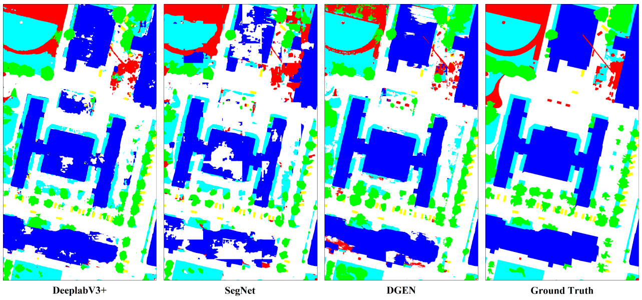

FIGURE 9. Comparison of the object integrity in the left-top red box area Figure 5 (e).

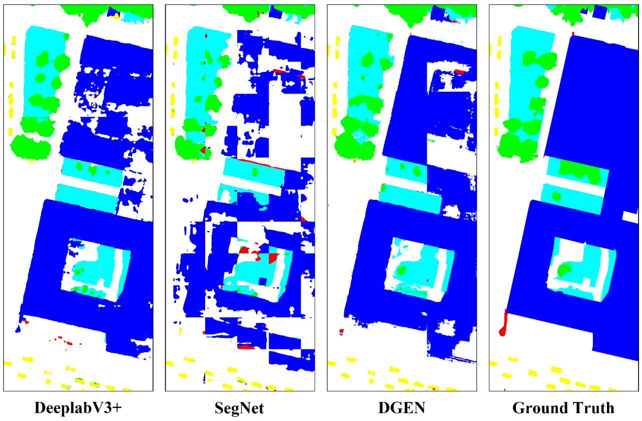

FIGURE 10. Comparison of the object integrity in the right red box area of Figure 5 (e).

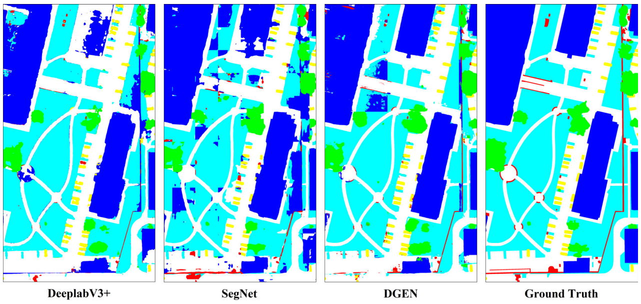

FIGURE 11. Comparison of the dissociated spots in the top-mid red box of Figure 5 (a).

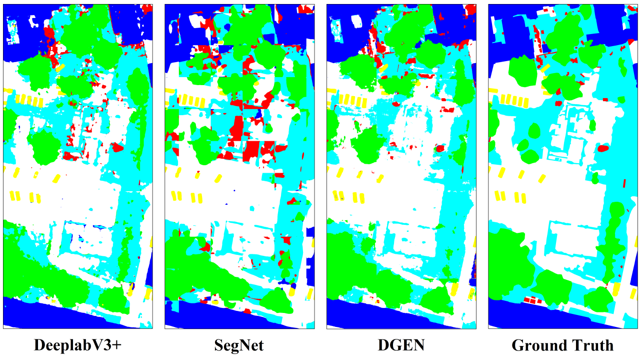

FIGURE 12. Comparison of the dissociated spots in the bottom-left red box of Figure 5 (d).

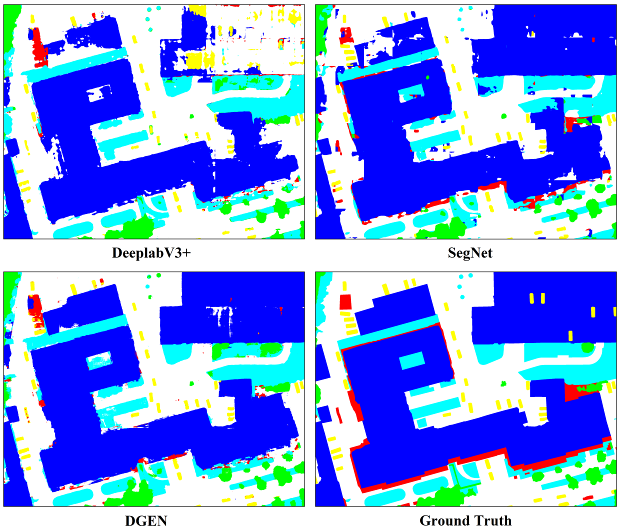

FIGURE 13. Comparison of the boundary quality in the bottom red box of Figure 5 (a).

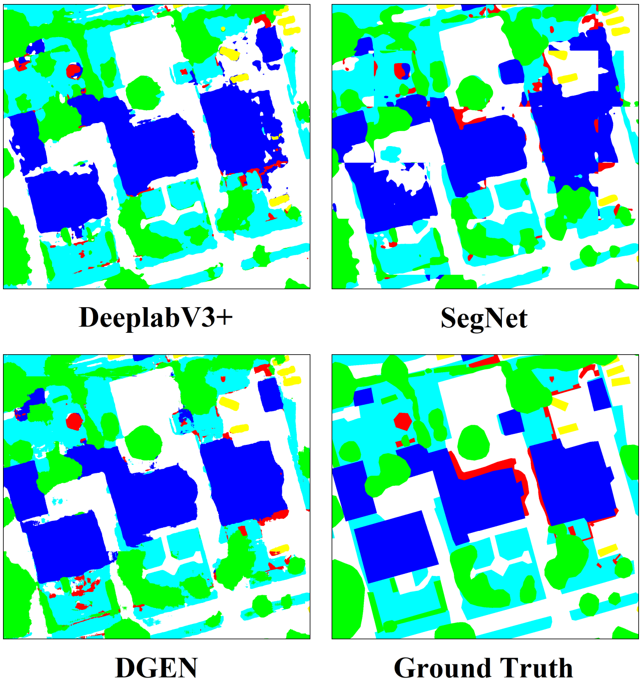

FIGURE 14. Comparison of the boundary quality in the top red box of Figure 5 (b).

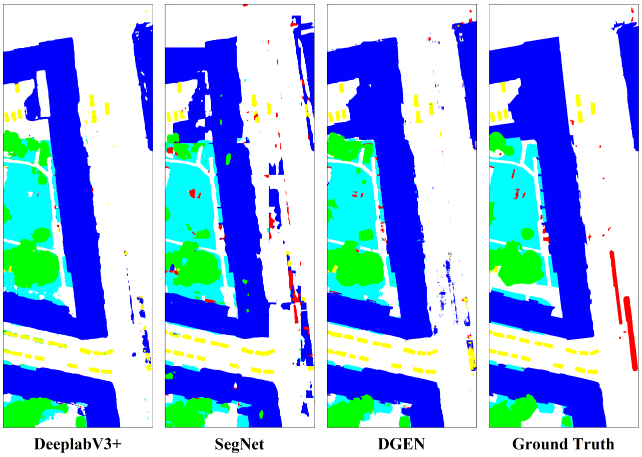

FIGURE 15. Comparison of the boundary quality in the right red box of Figure 5 (d).

V. DISCUSSIONS

A. CONTRIBUTION OF THE DUAL-ATTENTION MECHANISM

To quantize the improvement in the dual attention mechanisms, we also implement the model without the attention modules. Both the global attention enhancement module and local attention fusion module in the model are blocked and skipped. This process changes the model back to our own fully convolutional DenseNet (FC-DenseNet). It needs to be clarified that this model here is only an FC-DenseNet-typed network but is not the same as any published FC-DenseNet model. Table 4 and Figure 16 shows the comparison of this FC-DenseNet without a global attention enhancement module and the local attention fusion module.

Compared with the no-attentional model, the average OA of the DGEN model improved by 0.29%, the mean F1 score improved by 0.48%, the kappa coefficient improved by 0.45%, and the MIoU improved by 0.54%.

Although these changes are not as significant as the DGEN compared to DeeplabV3+ or SegNet (improves approximately 2-6%), the improvements are almost in all subitems.

TABLE 4. The comparison of the OA, F1, kappa coefficient (a) and MIoU Scores (b) for classification by the model with or without the "global information enhancement module" and "local information fusion module".

Method	Building	Imp surf	Tree	Low_veg	Car	**OA**	**Mean *F1***	**Kappa**
No Attention	95.22%	90.89%	78.58%	78.19%	87.76%	89.02%	86.13%	84.02%
DGEN	**95.40%**	**91.23%**	**79.06%**	**78.46%**	**88.89%**	**89.31%**	**86.61%**	**84.47%**

(A)

	IoU (%)					Mean *IoU*
	Buildings	Imp surf	Tree	Low_veg	Car	(%)
No Attention	90.88	**84.26**	64.9	64.43	78.21	76.54
DGEN	**91.2**	83.91	**65.51**	**64.77**	**80.02**	**77.08**

(B)

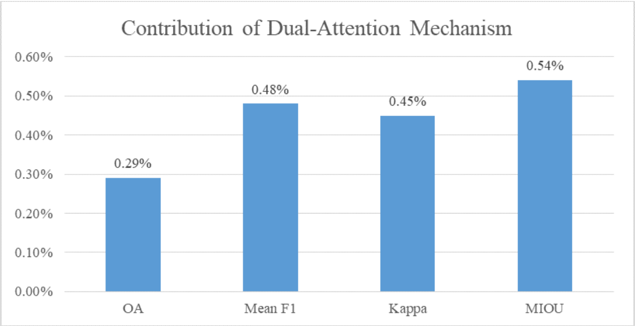

FIGURE 16. Improvement contribution of the dual attention mechanism.

All the statistical scores by any metric are increased. The PA of every single class is improved at the same time. This finding indicates that by the contribution of the dual attention mechanism, the performance of the network improves insignificantly but robustly.

Moreover, visual inspection shows abundant evidence of visible improvements in the object integrities and boundary qualities. According to the visual comparisons shown in Figure 17-21, several misclassifications are corrected, the integrities of a large number of objects are restored and the qualities of several boundaries are improved.

B. GENERALITY OF THE DGEN

To verify the generality of our DGEN model, we re-implement the comparison experiment of the U-net, DeeplabV3+, SegNet and DGEN models with another dataset. This dataset is a group of images of Vaihingen, Germany, which also includes the ISPRS 2D Semantic Labeling Challenge. The Vaihingen dataset contains 33 images with 9 cm resolution. 16 images of them have label data. In this experiment, images of area 11, 28 and 30 are selected for testing set. The other 13 images are training set. Unlike the downtown Potsdam data, the landscape of Vaihingen is a mixture of urban areas and countrysides, which hypothetically can provide different results from previous experiments.

The results of the OA, mean F1, Kappa coefficient and mean IoU scores are listed in Table 6.

According to Table 6, the results of the DGEN with different datasets also outperform the other models. The overall improvements in the DGEN are plotted in Figure 22.

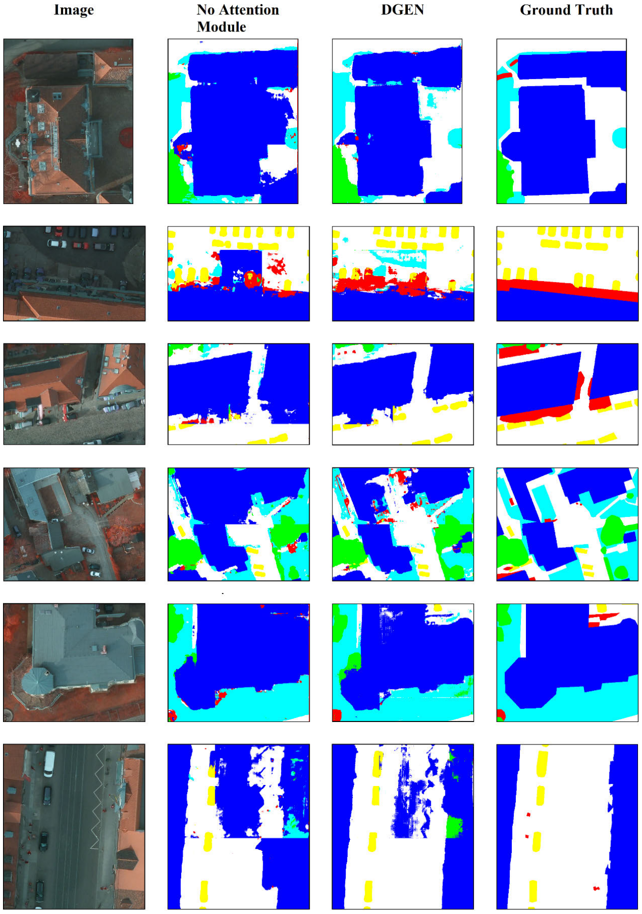

FIGURE 17. Comparison of the object validity and integrity with/without the dual attention mechanism.

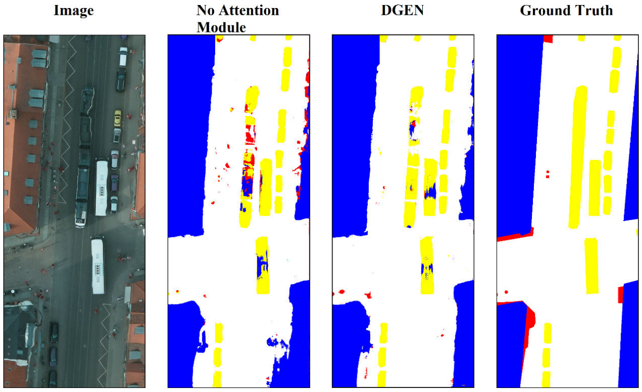

FIGURE 18. Comparison of the boundary quality with/without the dual attention mechanism (1).

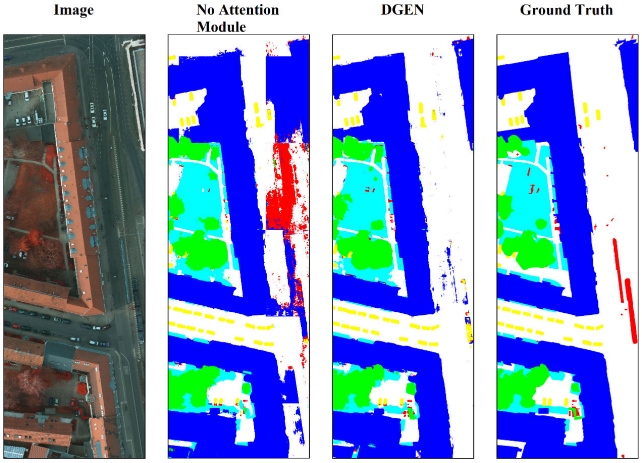

FIGURE 19. Comparison of the boundary quality with/without the dual attention mechanism (2).

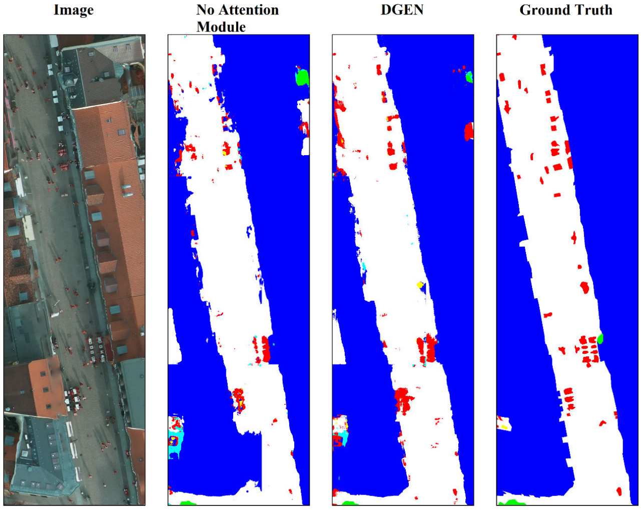

FIGURE 20. Comparison of the boundary quality with/without the dual attention mechanism (3).

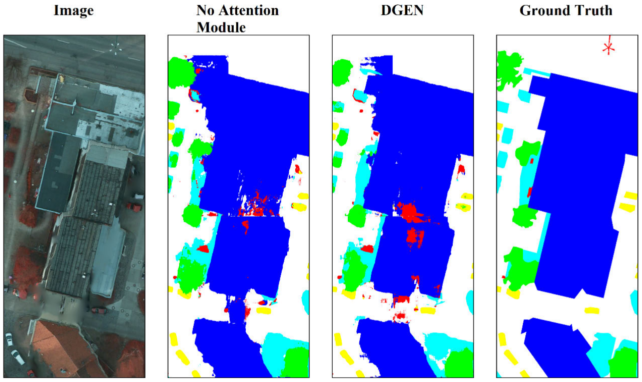

FIGURE 21. Comparison of the boundary quality with/without the dual attention mechanism (4).

TABLE 5. Image selection Of ISPRS 2D semantic labeling data - vaihingen.

Traning set		Testing set
top_mosaic_09cm_area1	top_mosaic_09cm_area21	top_mosaic_09cm_area11
top_mosaic_09cm_area3	top_mosaic_09cm_area23	top_mosaic_09cm_area28
top_mosaic_09cm_area5	top_mosaic_09cm_area26	top_mosaic_09cm_area30
top_mosaic_09cm_area7	top_mosaic_09cm_area32	
top_mosaic_09cm_area13	top_mosaic_09cm_area34	
top_mosaic_09cm_area15	top_mosaic_09cm_area37	
top_mosaic_09cm_area17		

TABLE 6. The comparison of the OA, F1, kappa coefficient (a) and Miou scores (b) for classification on the vaihingen dataset by U-Net, DeeplabV3+, SegNet and DGEN.

Method	Building	Imp surf	Tree	Low_veg	Car	**OA**	**Mean *F1***	**Kappa**
U-net	86.24%	80.97%	80.36%	62.14%	60.98%	78.36%	74.14%	71.18%
DeeplabV3+	81.25%	78.47%	74.08%	41.49%	37.53%	72.38%	62.56%	63.01%
SegNet	88.10%	**84.61%**	81.38%	67.33%	60.76%	81.03%	76.44%	74.72%
DGEN	**89.63%**	84.36%	**82.75%**	**70.40%**	**64.67%**	**82.03%**	**78.36%**	**76.14%**

(A)

	IoU (%)					**Mean *IoU***
	Building	Imp surf	Tree	Low_veg	Car	(%)
U-net	75.87	68.09	67.68	45.15	43.89	60.13
Deeplab V3+	68.47	64.76	59.28	26.31	23.25	48.41
SegNet	78.75	**73.38**	69.09	50.86	43.66	63.15
DGEN	**81.25**	73.02	**70.83**	**54.39**	**48.05**	**65.51**

(B)

Compared with U-net, our DGEN by all selected metrics outperforms in this experiment. The overall PA of the DGEN improved 3.67% over U-net. The mean F1 score improved by 4.22%, which means that both the precision and recall of the DGEN are better balanced and have less shortage. The kappa coefficient improved by 4.96%, which indicates a significant enhancement in the consistency of the classification. The MIoU improved by 5.38%, which means a dramatic improvement in the object-oriented point of view.

Compared with SegNet, DGEN outperforms generally with only the impervious surface class surpassed. The overall PA of the DGEN improved 1.00% from SegNet. The mean F1 score improved by 1.92%, which indicates a better balance between the precision and recall. The kappa coefficient improved by 1.42%. The MIoU improved by 2.36%.

Notably, in the Vaihingen dataset experiment, the results of the U-net become competitive, but the results of DeeplabV3+

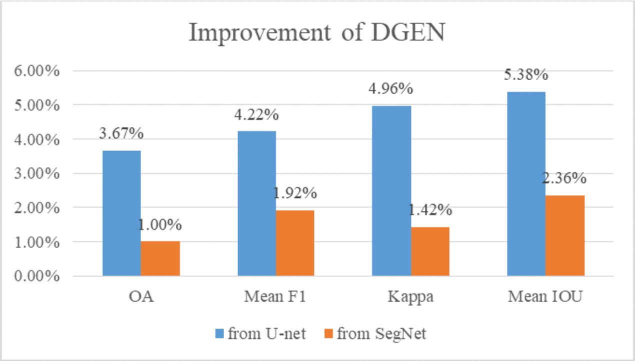

FIGURE 22. Improvement in the DGEN from U-net and SegNet with respect to the selected metrics.

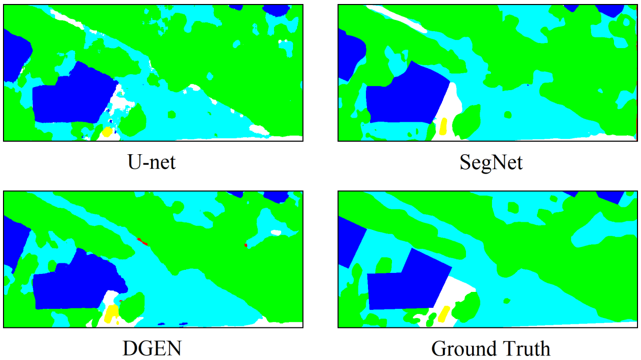

FIGURE 23. Comparison of the boundary qualities of trees.

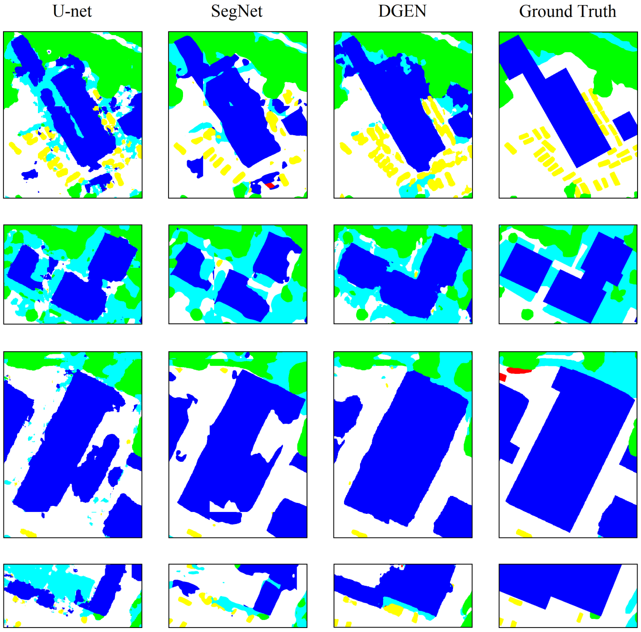

FIGURE 24. Comparison of the object integrities.

are poorer by more than 10% with regard to the selected metrics. Combining the results of U-net in Table 3, the stability and generality of U-net and DeeplabV3+ are not as satisfactory as SegNet and our DGEN.

Moreover, visual inspection also shows improvements in the object integrities and boundary qualities. According to the visual comparisons shown in Figure 23-24, the integrities of several objects are restored, and the quality of the boundaries is improved.

In summary, the result of the Vaihingen dataset experiment also shows the leading performance among current state-of-art models. The highest scores of all selected metrics (OA, mean F1, kappa coefficient and MIoU) indicate the statistically best results, and visual comparisons show an improvement in the object integrity and boundary quality. These results are consistent with the results of the Potsdam dataset experiment, which proves the generality of our DGEN model.

VI. CONCLUSION

By embedding an attention mechanism into a densely connected convolutional network (DenseNet), this study presents a dense-global-entropy network (DGEN) for the semantic segmentation of VHR remote sensing images. Tests on ISPRS 2D datasets verified the better performance improvement of the DGEN compared with U-net, DeeplabV3+ and SegNet. In the comparison experiments of both datasets, the proposed DGEN model shows good generality stability (while U-net and DeeplabV3+ can only be competitive in one dataset) and the highest score in all selected metrics, including the OA, F1, kappa coefficient and mean IoU. Furthermore, the two intensely existing shortages of recent deep-learning-based semantic segmentation models, namely, over-segmentation in low-level features and reduced object integrity, are improved by the dual attention mechanism.

The DGEN can achieve high accuracy on large-scale real remote sensing data without any supplemental data and without preprocessing or post-processing. The results produced in this study are comparable to the results of state-of-the-art models, and the experiments in this study indicate that by any selected metrics, the proposed DGEN model outperforms the other models.

Although the proposed the DGEN model has been validated with better performance by experiments, there are still some misclassifications by the boundaries of objects and between objects with similar textures. The main reason for these misclassifications may be the lack of the effective repair of category pixel localization in the decoders and unbalanced semantic classes in datasets. Our future studies will focus on exploring the architecture of the DGEN and improving network performance.

AUTHOR CONTRIBUTIONS

Lin Li, Hui Yang and Huanjun Hu designed the experiments; Haihong Zhu contributed analysis tools; Zheng Li and Hui Yang performed the experiments; and Huanjun Hu and Hui Yang wrote the paper. All authors have read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENT

The authors would like to thank the ISPRS for making the Potsdam datasets available and organizing the semantic labeling challenge. They would also like to thank Ms. X. Zhou and X. Yao (Anhui University) for providing data preprocessing.

REFERENCES

[1] X. Hu, C. V. Tao, and B. Prenzel, "Automatic segmentation of high-resolution satellite imagery by integrating texture, intensity, and color features," *Photogramm. Eng. Remote Sens.*, vol. 71, no. 12, pp. 1399–1406, Dec. 2005.

[2] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[3] E. Binaghi, I. Gallo, and M. Pepe, "A neural adaptive model for feature extraction and recognition in high resolution remote sensing imagery," *Int. J. Remote Sens.*, vol. 24, no. 20, pp. 3947–3959, Jan. 2003.

[4] A. Carleer, O. Debeir, and E. Wolff, "Comparison of very high spatial resolution satellite image segmentations," *Proc. SPIE*, vol. 5238, pp. 532–542, Feb. 2004.

[5] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNNs," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 473–480, Jun. 2016.

[6] M. Herold, X. Liu, and K. C. Clarke, "Spatial metrics and image texture for mapping urban land use," *Photogramm. Eng. Remote Sens.*, vol. 69, no. 9, pp. 991–1001, Sep. 2003.

[7] Z. Hu, Z. Wu, Q. Zhang, Q. Fan, and J. Xu, "A spatially-constrained color–texture model for hierarchical VHR image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 120–124, Jan. 2013.

[8] A. Izadipour, B. Akbari, and B. Mojaradi, "A feature selection approach for segmentation of very high-resolution satellite images," *Photogramm. Eng. Remote Sens.*, vol. 82, no. 3, pp. 213–222, Mar. 2016.

[9] A. Izadipour, B. Akbari, and B. Mojaradi, "A new feature selection method for segmentation of VHR satellite image," in *Proc. Int. Conf. Commun., Signal Process., Appl.*, Feb. 2015, pp. 1–5.

[10] Z. Lv, P. Zhang, and J. Atli Benediktsson, "Automatic object-oriented, spectral-spatial feature extraction driven by Tobler's first law of geography for very high resolution aerial imagery classification," *Remote Sens.*, vol. 9, no. 3, p. 285, Mar. 2017.

[11] H. Mahi, H. Isabaten, and C. Serief, "Zernike moments and SVM for shape classification in very high resolution satellite images," *Int. Arab J. Inf. Technol.*, vol. 11, pp. 43–51, Jan. 2014.

[12] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[13] P. Tokarczyk, J. D. Wegner, S. Walk, and K. Schindler, "Features, color spaces, and boosting: New insights on semantic classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 280–295, Jan. 2015.

[14] L. Bruzzone and L. Carlin, "A multilevel context-based system for classification of very high spatial resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2587–2600, Sep. 2006.

[15] X. Chen, T. Fang, H. Huo, and D. Li, "Graph-based feature selection for object-oriented classification in VHR airborne imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 353–365, Jan. 2011.

[16] M. Pesaresi and J. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, 2001.

[17] D. Tuia, F. Pacifici, M. Kanevski, and W. Emery, "Classification of very high spatial resolution imagery using mathematical morphology and support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3866–3879, Nov. 2009.

[18] W. Su, J. Li, Y. Chen, Z. Liu, J. Zhang, T. M. Low, I. Suppiah, and S. A. M. Hashim, "Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery," *Int. J. Remote Sens.*, vol. 29, no. 11, pp. 3105–3117, Jun. 2008.

[19] J. Tian and D. Chen, "Optimization in multi-scale segmentation of high-resolution satellite images for artificial feature recognition," *Int. J. Remote Sens.*, vol. 28, no. 20, pp. 4625–4644, Oct. 2007.

[20] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, 1995.

[21] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sens.*, vol. 9, no. 5, p. 498, May 2017.

[22] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[23] G. E. Hinton, "Learning multiple layers of representation," *Trends Cogn. Sci.*, vol. 11, no. 10, pp. 428–434, Oct. 2007.

[24] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Computer Vision*, vol. 10111, S. H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2017, pp. 180–196.

[25] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[26] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[27] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*. [Online]. Available: https://arxiv.org/abs/1606.02585

[28] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.

[29] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sens.*, vol. 8, no. 4, p. 329, Apr. 2016.

[30] Y. F. Zhong, F. Fe, and L. P. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, Apr. 2016, Art. no. 025006.

[31] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 446, May 2017.

[32] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[33] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *CoRR*, vol. abs/1412.7062, pp. 1–14, Dec. 2014.

[34] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.

[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[36] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, Jan. 2018.

[37] X. Yao, H. Yang, Y. Wu, P. Wu, B. Wang, X. Zhou, and S. Wang, "Land use classification of the deep convolutional neural network method reducing the loss of spatial features," *Sensors*, vol. 19, no. 12, p. 2792, Jun. 2019.

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

[39] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," presented at the Brit. Mach. Vis. Conf., 2018.

[40] B. L. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," presented at the Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015.

[41] G. L. Huang, Z. V. D. Maaten, L. Weinberger, and Q. Kilian, "Densely connected convolutional network," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.

[42] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers Tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1175–1183.

[43] Y. Yang, Z. Zhong, T. Shen, and Z. Lin, "Convolutional neural networks with alternately updated clique," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2413–2422.

[44] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[45] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6450–6458, Jul. 2017.

[46] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017.

[47] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[48] R. Xu, Y. Tao, Z. Lu, and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification," *Remote Sens.*, vol. 10, no. 10, p. 1602, Oct. 2018.

[49] S. Liu, Q. Wang, and X. Li, "Attention based network for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4740–4743.

[50] Z. Xiong, Y. Yuan, and Q. Wang, "AI-NET: Attention inception neural networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2647–2650.

[51] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[52] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11, p. 1768, Nov. 2018.

[53] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *CoRR*, vol. abs/1805.10180, 2018.

HUANJUN HU received the B.S. degree in surveying from Wuhan University, China, in 2010, and the M.Sc. degree in geomatics from The Hong Kong Polytechnic University, in 2012. He is currently pursuing the Ph.D. degree in cartography and geographical information engineering with Wuhan University. His research interests include remote sensing image classification based on deep learning, pedestrian evacuation dynamic modeling, and geographic information modeling.

ZHENG LI received the B.S. degree in geographic information system from the Wuhan University of Technology, China, in 2009, and the M.S. degree in geographical information engineering from Wuhan University, China, in 2012. She is currently with the Hubei Institute of Land Surveying and Mapping. Her research interests include geographical information mining, remote sensing image progressing, and artificial intelligence.

LIN LI received the Ph.D. degree from Wuhan University, China, in 1997. He was a Professor with the School of Resource and Environmental Science, Wuhan University. He was with Joseph Fourier University, France, and The University of Tokyo, Japan, for over four years. His current research interests include 3-D modeling and visualization, geographical ontology, 3-D cadastre, the integration of ubiquitous location information, and feature extraction from point cloud data.

HUI YANG received the B.S. degree in geography from Anqing Normal University, China, in 2010, the M.S. degree in surveying from Wuhan University, China, in 2012, and the Ph.D. degree in cartography and geographical information engineering from the School of Resource and Environment Sciences, Wuhan University, in 2019. He is currently a Lecturer with the Institutes of Physical Science and Information Technology, Anhui University, China. His research interests include coastline map gerneralization, remote sensing image semantic segmentation, artificial intelligence, and geographic information mining.

HAIHONG ZHU received the B.S. and M.S. degrees in cartography from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1986 and 1996, respectively, and the Ph.D. degree in cartography from Wuhan University, China, in 2013. She is currently a Professor and the Ph.D. Advisor with the School of Resource and Environmental Sciences, Wuhan University. Her research interests include navigation digital map, map designs, geographical ontology, and the 3-D modeling and visualization of geographical information.