Understanding the Role of Receptive Field of Convolutional Neural Network for Cloud Detection in Landsat 8 OLI Imagery

Longkang Peng, Xuehong Chen, Jin Chen, Wenzhi Zhao, and Xin Cao

***Abstract*— Deep semantic segmentation networks perform better in cloud detection of satellite imagery than traditional methods due to their ability to extract high-level features over a large receptive field. However, a large receptive field often leads to loss of spatial details and blurring of boundaries. Therefore, it is crucial to understand the role of the receptive field on the segmentation results, which has rarely been investigated for cloud detection tasks. This study, for the first time, explored the relationship between the receptive field size and the performance of a cloud detection network. Six typical networks commonly used for cloud detection and nine modified UNet variants with different depths, dilated convolutions, and skip connections were evaluated based on the Landsat 8 Biome (L8 Biome) dataset. The theoretical receptive field (TRF) and the effective receptive field (ERF) were introduced to measure the receptive field sizes of different networks. The results revealed a negative correlation between the ERF size and cloud segmentation accuracies for different cloud distributions and a relatively weak negative correlation between the TRF size and segmentation accuracies. Furthermore, ERFs were considerably smaller than the corresponding TRFs for most networks, implying that large-scale contextual information was not learned after training. This result indicates the importance of using networks with a small receptive field for cloud detection of Landsat 8 OLI imagery. Moreover, as the boundary accuracies are significantly lower than the region accuracies, future efforts should be devoted to addressing inaccurate boundary localization rather than exploring the contextual information over a large receptive field.**

***Index Terms*— Cloud detection, convolutional neural network (CNN), Landsat 8, receptive field.**

I. Introduction

LANDSAT satellites publicly avail optical imagery datasets with long records and global coverage and have been widely used in various applications [1]–[5]. However, optical satellite imagery is likely to be affected by clouds, which covers more than 50% of the Earth's surface at any given moment [6]–[10]. Therefore, it is critical to accurately label cloud-covered areas before further processing and analysis of optical imagery.

Manuscript received September 28, 2021; revised January 26, 2022; accepted February 4, 2022. Date of publication February 8, 2022; date of current version March 23, 2022. This work was supported by the National Natural Science Foundation of China under Project 41871224. *(Corresponding author: Xuehong Chen.)*

The authors are with the State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Institute of Remote Sensing Science and Engineering, Beijing Normal University, Beijing 100875, China (e-mail: chenxuehong@bnu.edu.cn).

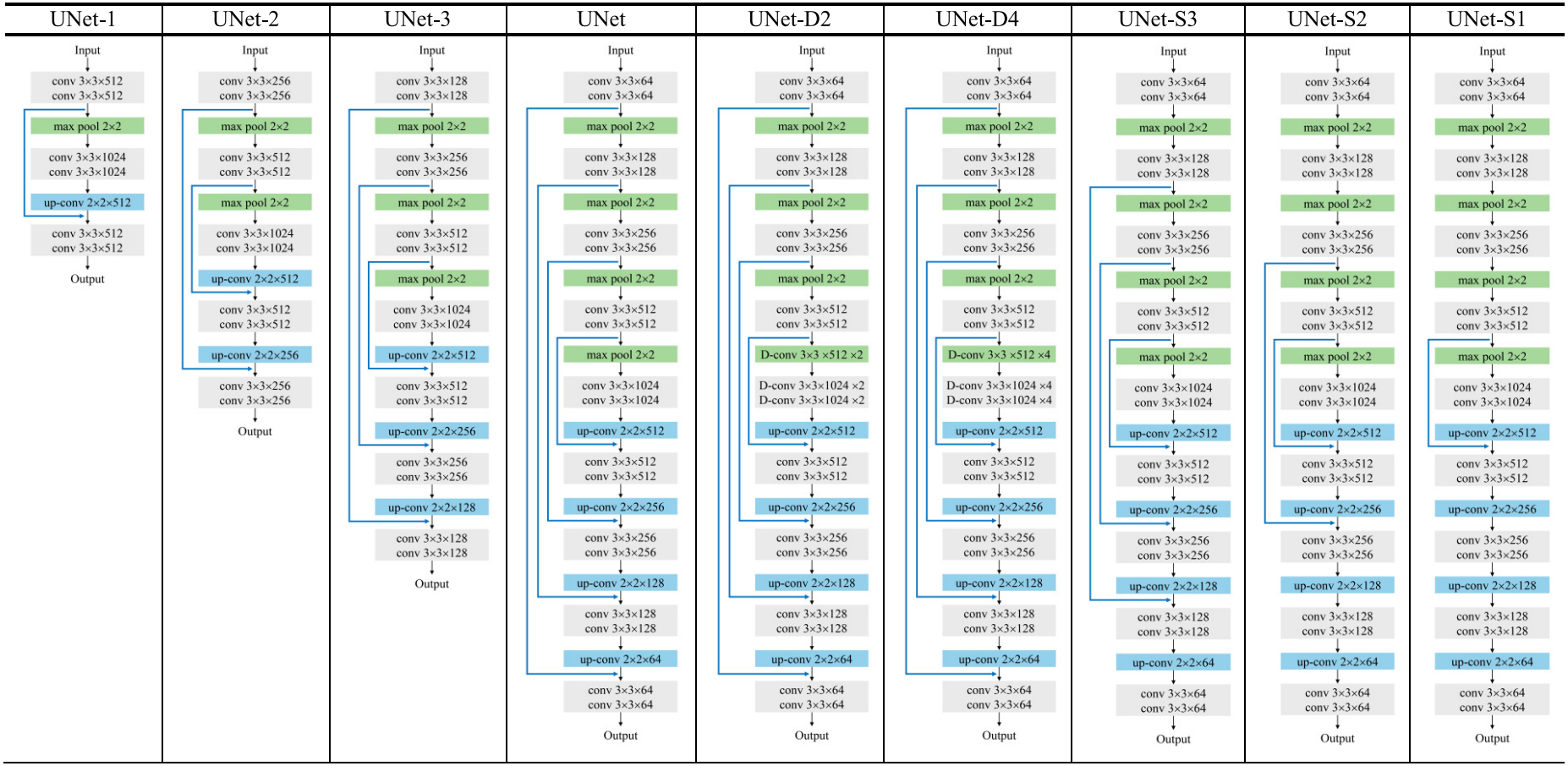
Digital Object Identifier 10.1109/TGRS.2022.3150083

Cloud detection or pixel-by-pixel cloud labeling for satellite imagery has received increasing attention over the last few decades [11], [12]. Cloud detection methods for the Landsat imagery are commonly classified into two categories: physical-rule- and machine-learning-based methods [11]. Physical-rule-based methods detect clouds by identifying their physical or empirical characteristics, such as "bright," "white," "cold," and "high" [13]–[19]. In contrast, machine learning techniques show great potential with respect to the image classification of remote sensing [20]–[22], [44]; therefore, they have been widely used for the classification of cloud pixels in Landsat imagery [3], [23]–[26]. Facilitated by deep convolutional neural networks (CNNs), networks that are capable of performing dense prediction tasks (i.e., semantic segmentation) have been recently applied and developed in cloud detection tasks and have demonstrated more optimized performance than traditional methods [3], [4], [25]–[31]. The superiority of deep CNNs to traditional methods is commonly attributed to their ability to automatically extract advanced features at a higher or more abstract level [32], [33]. A large receptive field covering the relevant image region is essential to learning high-level features for a CNN [34]. Generally, there are two main ways of increasing the receptive field. The first option is to deepen the network by stacking additional layers [34]. Therefore, CNN models have become deeper over the past decade and achieved improved performance in vision tasks [35]–[38]. For example, the DeepLab family, based on the ResNet backbone with significantly deeper layers than previous structures, yields more optimized segmentation results on several vision benchmarks than the shallower networks [39]–[41]. Another option is increasing the receptive field of each layer, for example, dilated convolutional kernel and subsampling operation, which can also increase the receptive field size of the entire network. Dilated convolution has been adopted in various networks, such as DeepLabv3 [40], DeepLabv3+ [41], and MSCFF [25], to efficiently explore the large-scale context features without increasing the computation cost and training time. PSPNet [42] and MF-CNN [26] enlarge the receptive field by introducing multiscale pooling, which consists of parallel multiscale subsampling operations. In summary, enlarging the receptive field is an important trend in the development of CNNs.

However, the final segmentation results yielded by a CNN with a large receptive field may lose the spatial details and cannot accurately delineate its borders [39], [43]–[45]. Efforts

TABLE I

CONFIGURATION OF NINE UNET VARIANTS. THE FEATURES FLOW DOWNWARD THROUGH A SERIES OF "CONV," "MAX POOL" AND "UP-CONV" OPERATIONS WITH SKIP CONNECTIONS MARKED BY THE BLUE ARROWS



"conv (m×m×n)": convolutional layer (with *n* m×m kernels) followed by Batch Normalization (BN) and Rectified Linear Units (ReLU).
"max pool (m×m)": max pooling layer (m×m kernel) used for down-sample the feature maps.
"up-conv (m×m×n)": deconvolutional layer (with *n* m×m kernels) used for up-sampling the feature maps.
"D-conv (m×m×n×r)": dilated convolutional layer (with *n* m×m kernels and rate of r) followed by Batch Normalization (BN) and Rectified Linear Units (ReLU).
Output layer: class confidence map obtained through a convolutional layer with two 3×3 kernels.

have been made to accurately localize object boundaries for segmentation networks. Exploiting features from intermediate layers is a common strategy for enhancing boundary localization. For example, UNet [46] and its variants specifically designed for cloud detection (e.g., TL-Net [28], MUNet [27], RS-Net [3], and Refined UNet [47]) adopt this strategy through a skip connection between two symmetrical paths to retain more spatial details. MF-CNN [26] and MSCFF [25], two networks specifically designed for the cloud detection of the Landsat imagery, also adopt a similar skip connection mechanism. Postprocessing of the segmentation results is another way of refining the segmentation results. For example, the conditional random field (CRF) [39] and multiwindow-guided filtering [48] were used to refine the coarse segmentation outputs of UNet, which has shown promising results in cloud detection tasks [47], [48]. Recently, formulating a loss function to focus more on the pixels near the boundary than those in the center of the cloud or clear region by penalizing errors in object boundaries was developed to optimize the boundary blur [49] and improve cloud detection results [50].

Although a lot of effort has been devoted to exploring rich information over a large receptive field and obtaining accurate boundary localization, balancing the tradeoff between these two aspects continues to be a challenge [44]. Therefore, understanding the effect of the receptive field size on the segmentation results is essential to selecting a suitable network or developing a novel network. Unfortunately, the role of the receptive fields of CNN networks for cloud detection has rarely been investigated. In this study, two types of receptive fields, the theoretical receptive field (TRF) and the effective receptive field (ERF), were introduced to quantify the receptive field size and explore their effects on the cloud detection accuracy of different networks for the Landsat 8 OLI imagery. TRF refers to the region where an output unit depends on the input, which is determined only by the network architecture [34]. Expanding the receptive field by improving the network architecture commonly results in a larger TRF. ERF refers to the effective area in the TRF, which can reveal the context effectively influencing semantic predictions [34], [51]. As the ERF is estimated after training, the ERF of the same network can vary with different training datasets. If the large-scale context is not significant for the classification or segmentation of a certain dataset, the ERF should be reduced after training, even if the designed network has a large TRF. Therefore, ERF is an effective tool for exploring whether the designed TRF plays a role in a certain task. Considering that clouds in satellite imagery have unique spectral features, herein, we reexamined the need for a large TRF on the cloud detection task of Landsat data and explored the relationship among TRF, ERF, and cloud detection accuracy. Six typical networks, namely, U-Net [46], DeepLabv3+ [41], TL-Net [28], MUNet [27], MF-CNN [26], and MSCFF [25], were selected for cloud detection of the Landsat 8 OLI imagery to explore the differences between TRF, ERF, and segmentation accuracy among various networks. Furthermore, considering that network depth and dilated convolution are two important ways of effectively enlarging the TRF, and skip connection is a commonly used structure that largely affects segmentation results, nine U-Net variants with different depths, dilated convolutions, and skip connections were designed and compared to examine the effect of the network architecture on TRF, ERF, and segmentation accuracy.

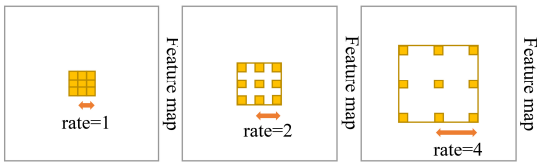


Fig. 1. Illustration of dilated convolution with a 3 × 3 kernels and different rates. Standard convolution is a special case for a rate of 1.

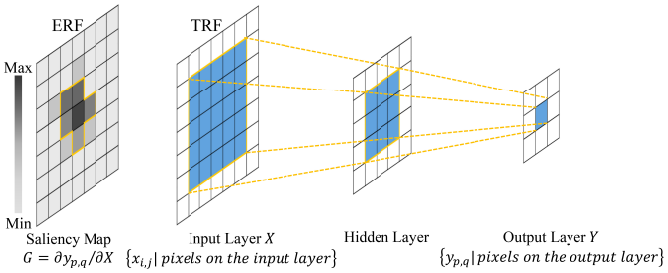


Fig. 2. Illustration of TRF and effective receptive field (ERF) for a simple model with three layers.

The remainder of this article is organized as follows. In Section II, we present the dataset and experimental settings, including the segmentation network structure, training, validation, and test details. The experimental results and analyses are presented in Section III. Further discussions are presented in Section IV, followed by the conclusions in Section V.

II. Dataset and Experimental Settings

A. Dataset and Preprocessing

Landsat 8 cloud cover assessment validation data, called Landsat 8 Biome (L8 Biome) (acquired from https://landsat.usgs.gov/landsat-8-cloud-cover-assessment-validation-data), were used in this study. The L8 Biome is one of the most widely used datasets for research on cloud detection [3], [25], [28], [52], [53]. The L8 Biome consists of 96 Landsat 8 scenes acquired over different biomes and the corresponding manually generated cloud masks, which flag pixels as five classes, including cloud, thin cloud, clear, cloud shadow, and fill for each scene [52].

In our experiments, the dataset was randomly divided into three parts: training (48 scenes), validation (24 scenes), and testing (24 scenes). The top of atmosphere (TOA) reflectance of eight bands of Landsat 8 OLI, including bands 2 (blue), 3 (green), 4 (red), 5 (NIR), 6 (SWIR1), 7(SWIR2), 9 (Cirrus), and 10 (TIR1), was used as network inputs. To focus the cloud detection performance, we merged classes "cloud" and "thin cloud" into "cloud," and classes "clear" and "cloud shadow" into "clear," similar to previous studies [41], [54]. Due to the GPU memory limitation, each scene was split into small 256 × 256 pixel patches; therefore, 26 558, 13 314, and 13 415 patches were generated for training, validation, and testing, respectively.

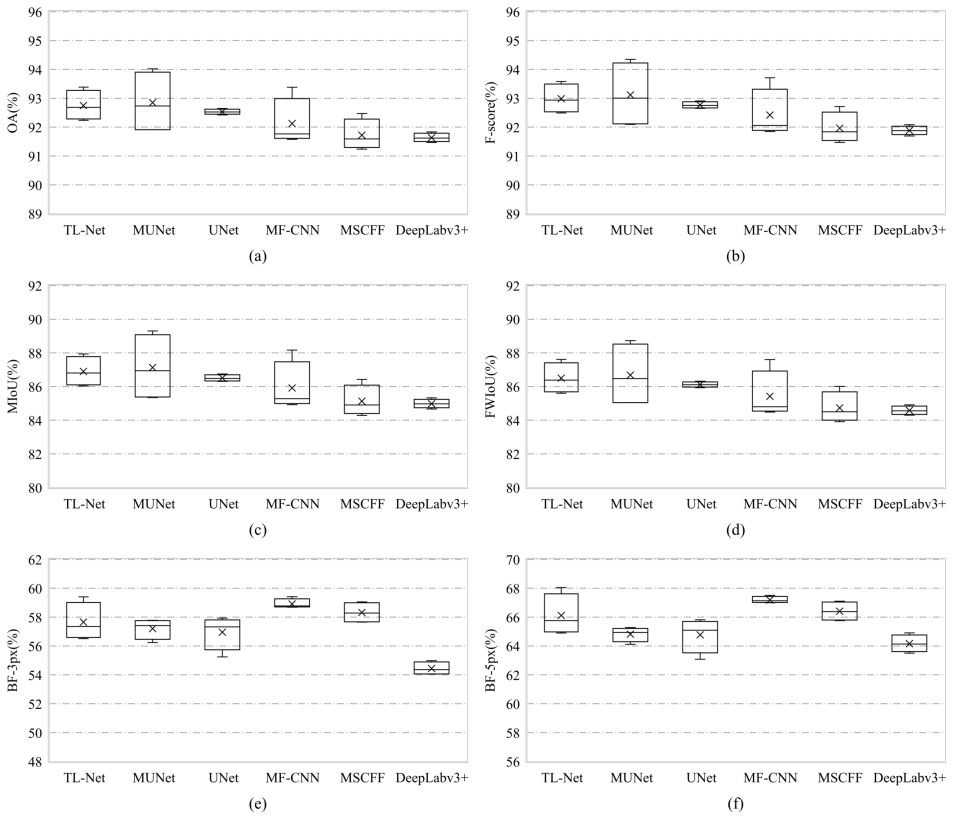


Fig. 3. Quantitative comparison results of six typical networks in terms of (a) OA, (b) F-score, (c) MIoU, (d) FWIoU, (e) BF-3px, and (f) BF-5px.

TABLE II
TRF SIZE, ERF SIZE, AND SIX EVALUATION METRICS OF SIX TYPICAL NETWORKS

Method	TRF	ERF	OA(%)	F-score(%)	MIoU(%)	FWIoU(%)	BF-3px(%)	BF-5px(%)
TL-Net	32	16.83±6.72	92.75±0.45	92.98±0.44	86.89±0.77	86.50±0.78	57.65±1.13	66.11±1.24
MUNet	140	17.38±14.93	92.85±0.95	93.11±1.00	87.13±1.75	86.68±1.66	57.21±0.61	64.82±0.43
UNet	140	11.74±8.36	92.53±0.08	92.76±0.09	86.51±0.16	86.12±0.14	56.86±1.02	64.77±1.03
MF-CNN	185	18.24±8.48	92.12±0.73	92.42±0.75	85.91±1.31	85.42±1.27	58.91±0.29	67.19±0.19
MSCFF	722	21.49±30.25	91.72±0.46	91.97±0.46	85.13±0.80	84.73±0.79	58.32±0.63	66.41±0.58
DeepLabv3+	1267	38.71±35.75	91.63±0.13	91.88±0.14	84.98±0.23	84.58±0.23	54.44±0.39	64.17±0.52

B. Segmentation Network Structure

1) Six Typical Cloud Detection Networks: The following six typical cloud detection networks were selected for our experiment: UNet [46], DeepLabv3+ [41], TL-Net [28], MUNet [27], MF-CNN [26], and MSCFF [25]. UNet and DeepLabv3+ were selected because they are widely used in various applications and are commonly used as benchmarks for cloud detection [8], [25], [55]–[57]. For DeepLabv3+, ResNet-101 [36] was adopted as the backbone architecture of the encoder. TL-Net, MUNet, MF-CNN, and MSCFF were selected because they were specifically designed for cloud detection of satellite imagery and yielded promising results [25]–[28]. TL-Net and MUNet were simplified from the UNet architecture by reducing the number of pooling steps and convolution kernels, respectively; therefore, their receptive field size is similar to that of UNet. MF-CNN and MSCFF were designed to extract multiscale global features by multiscale pooling filters and multiscale feature fusion, respectively, which can explore multiscale contextures in large receptive fields [25], [26].

2) Nine UNet Variants: Considering that a deepening network and dilated convolution are two important ways of enlarging the receptive field, two types of UNet variants with different depths and different dilated convolutions were built to explore the effect of these two efforts on enlarging the receptive field. Considering that skip connection is a commonly used structure that significantly affects segmentation results, UNet variants with different numbers of skip connections were also built to explore their effect on the receptive field and segmentation accuracy. The details of the three variants are listed in Table I.

The network depth was changed by varying the number of blocks in the UNet (conv → conv → max pool in the contracting path and up-conv → conv → conv in the expanding path) according to previous studies [28], [58]. In our experiment,

four UNet variants with depths ranging from 1 to 4 (i.e., UNet-1, UNet-2, UNet-3, and UNet) were built (see Table I).

UNet variants with different dilated convolutions were built by replacing the fourth pooling step and the two subsequent convolutional layers in UNet with three dilated convolutional layers with a kernel size of 3×3 and stride of 1 (see Table I). The position of this replacement is the same as that of dilated convolutions in DeepLabv3 [40]. The dilated convolution kernel inserts holes in the 3×3 filter to enlarge the receptive field without losing spatial resolution and increasing the number of parameters. As shown in Fig. 1, a larger dilation rate, corresponding to more holes in the filter, will lead to a larger TRF. Considering that the spatial resolution of the resulting feature maps of the dilated convolution layer is not reduced, the subsequent up-conv layer (the first up-conv layer) in the original UNet is not necessary and is, therefore, replaced by a 1×1 convolution to maintain the spatial resolution. In our experiment, two dilated U-Net variants with dilation rates of 2 and 4 were built and named UNet-D2 and UNet-D4, respectively.

UNet variants with different numbers of skip connections were built by gradually removing skip connections from shallow layers to deep layers. The differences in the number of skip connections did not change the TRF, whereas they probably changed the ERF. In our experiment, three UNet variants with four skip connections, namely, UNet, UNet-S3, UNet-S2, and UNet-S1, were built (see Table I).

Our code and trained models are shared online to provide support for the research in the field of cloud detection (https://github.com/LK-Peng/CNN-based-Cloud-Detection-Methods.git).

C. Training, Validation, and Test Details

All the networks were implemented in PyTorch (1.7.1) under the operating system Ubuntu 7.5.0 equipped with two NVIDIA GTX 1080Ti GPUs. The stochastic gradient descent (SGD) algorithm [59] with a momentum of 0.9 and a weight decay of 5×10^{-4} was used for training. The other settings were given as follows: fixed batch size of 24, initial learning rate of 0.01, and learning rate decay by a factor of 0.1 every 20 epochs. Subsequently, the optimal model was selected from 100 training epochs based on the accuracy of the validation set. Specifically, it had the highest average FWIoU in the windows of five consecutive epochs. Finally, the selected optimal model was evaluated using the test set. Four copies of the same network were trained from scratch for hundred epochs with different random seed initializations to consider the random effect of the weight initialization.

D. Receptive Field Size

The TRF and ERF sizes were computed to explore the effect of the receptive field. As shown in Fig. 2, the TRF refers to the region that is theoretically connected to the output unit through network operations, whereas the ERF refers to the region with an effective impact on the output unit. Therefore, the region of the ERF was smaller than that of the TRF (see Fig. 2).

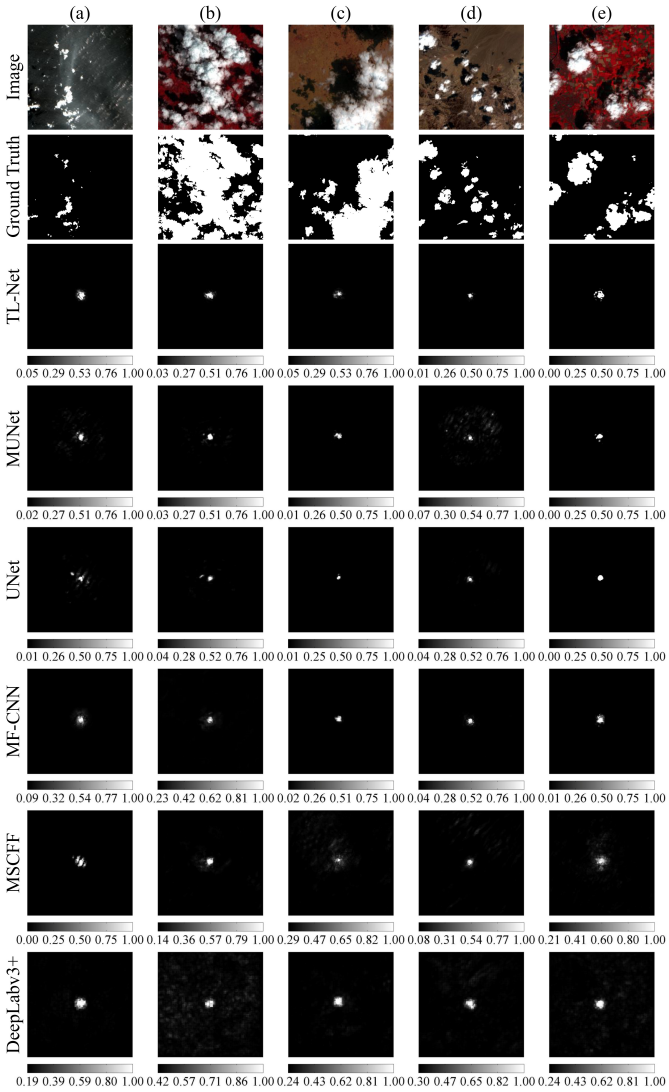


Fig. 4. Visual comparison of ERF for six typical networks in different locations. Receptive fields are visualized in the image space of 256×256 pixels. (a) From path 157 and row 45 (center pixel position: row 125 and column 131). (b) From path 16 and row 50 (center pixel position: row 126 and column 128). (c) From path 175 and row 51 (center pixel position: row 123 and column 124). (d) From path 159 and row 36 (center pixel position: row 128 and column 126). (e) From path 197 and row 24 (center pixel position: row 126 and column 125).

The detailed calculation methods for the areas of the TRF and ERF are described as follows.

1) TRF Size: The TRF size is determined only by the network architecture and can be calculated as follows [60]:

$$l_k = l_{k-1} + \left[(f_k - 1) \times \prod_{i=1}^{k-1} s_i \right], \quad \text{where } l_0 = 1 \tag{1}$$

where l_k is the TRF size of the network with k layers, f_k is the maximum kernel size of filters in the kth layer, and s_i is the stride of the kernels in the ith layer. Unfortunately, (1) is not valid for the upsampling layers because f_k cannot be correctly calculated for the deconvolutional filter and other upsampling methods. Therefore, in our experiment, the approximate TRF size was calculated based on the partial network structure

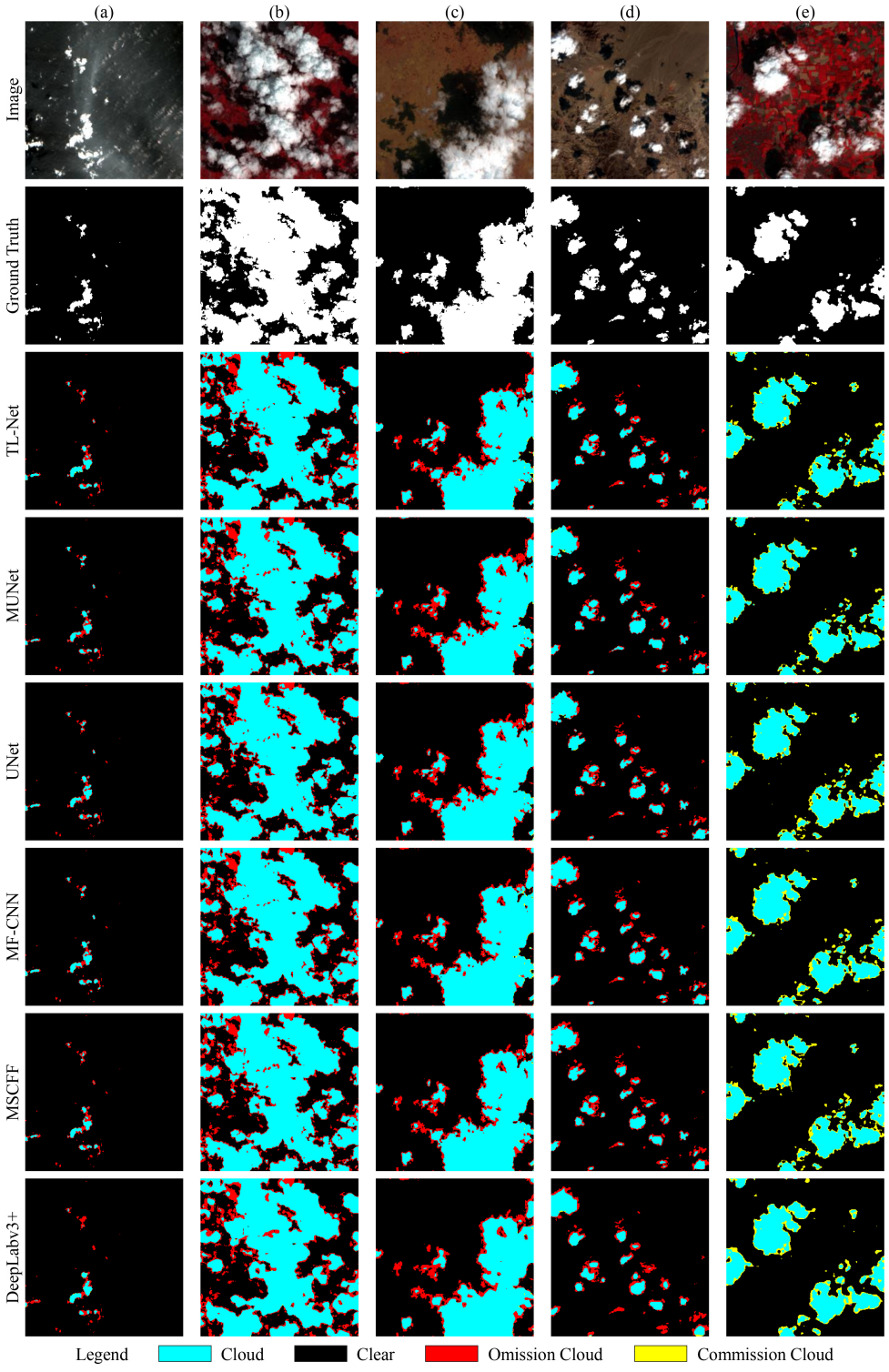


Fig. 5. Qualitative comparison of cloud detection results for six typical networks (under the random seed with optimal performance) in different locations. (a) From path 157 and row 45. (b) From path 16 and row 50. (c) From path 175 and row 51. (d) From path 159 and row 36. (e) From path 197 and row 24.

before the first upsampling layer. Such an approximation is reasonable because upsampling operations affect TRF marginally.

2) ERF Size: Because not all input pixels in a TRF contribute equally to an output unit's response, ERF was developed by Luo *et al.* [34] to represent the region containing any input pixels with a nonnegligible impact on that output unit. The ERF region commonly exhibits a Gaussian distribution and can be expressed using an impact gradient map (G). The gradient of each pixel ($g_{i,j}$) is the partial derivative of the output unit ($y_{p,q}$) with respect to the input pixel ($x_{i,j}$) on the

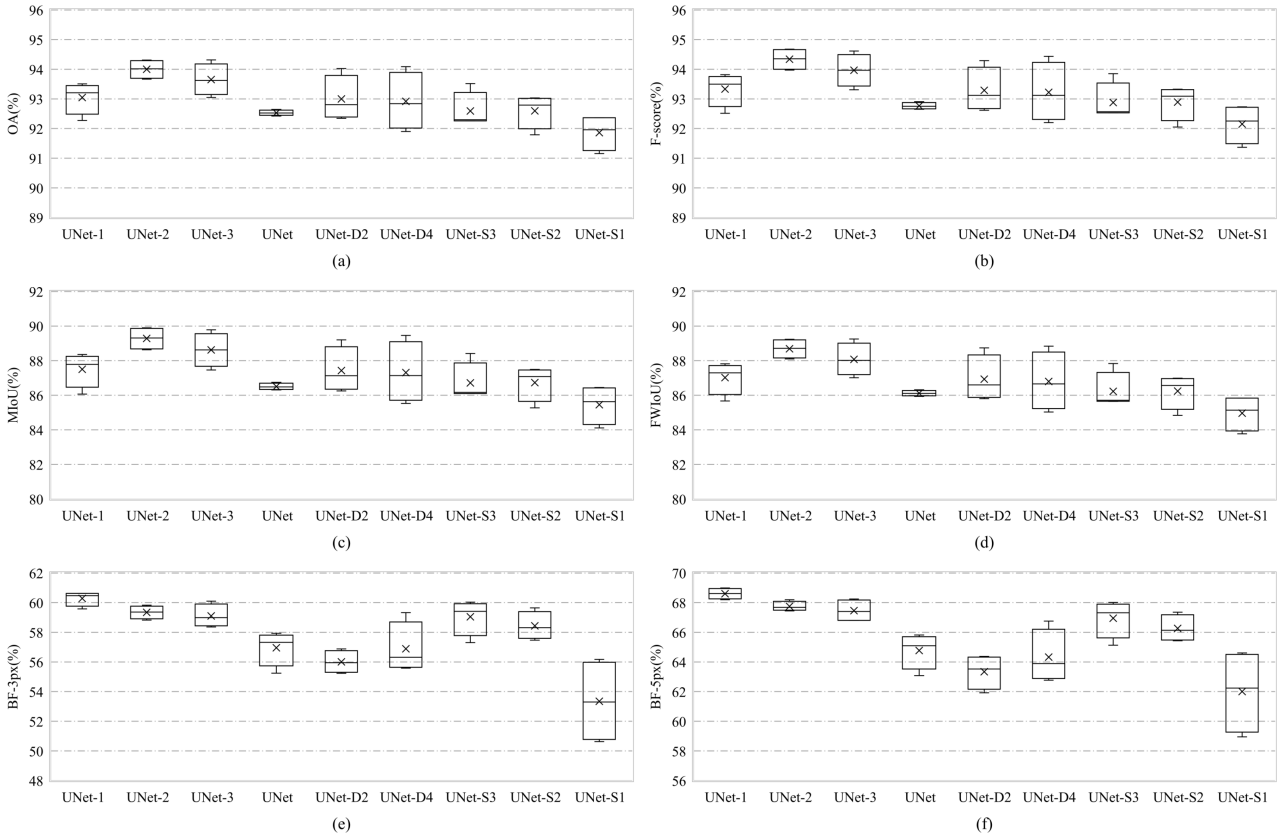


Fig. 6. Quantitative comparison results of six typical networks in terms of (a) OA, (b) F-score, (c) MIoU, (d) FWIoU, (e) BF-3px, and (f) BF-5px.

TABLE III
TRF SIZE, ERF SIZE, AND SIX EVALUATION METRICS OF NINE UNET VARIANTS

Method	TRF	ERF	OA(%)	F-score(%)	MIoU(%)	FWIoU(%)	BF-3px(%)	BF-5px(%)
UNet-1	14	11.26±1.67	93.05±0.47	93.33±0.49	87.50±0.86	87.02±0.81	60.28±0.42	68.61±0.32
UNet-2	32	12.45±4.25	94.00±0.28	94.34±0.31	89.29±0.56	88.69±0.49	59.35±0.38	67.75±0.28
UNet-3	68	15.40±9.60	93.65±0.46	93.96±0.47	88.62±0.84	88.08±0.82	59.11±0.67	67.46±0.66
UNet	140	11.74±8.36	92.53±0.08	92.76±0.09	86.51±0.16	86.12±0.14	56.96±1.02	64.77±1.03
UNet-D2	164	11.42±7.57	93.00±0.66	93.29±0.64	87.43±1.14	86.93±1.15	56.01±0.67	63.34±1.00
UNet-D4	260	9.80±3.30	92.91±0.85	93.22±0.87	87.31±1.53	86.79±1.47	56.88±1.48	64.33±1.55
UNet-S3	140	26.07±25.52	92.59±0.54	92.88±0.56	86.71±0.98	86.23±0.93	59.04±1.04	66.95±1.10
UNet-S2	140	40.30±31.00	92.60±0.50	92.89±0.51	86.73±0.89	86.24±0.86	58.43±0.81	66.27±0.78
UNet-S1	140	65.63±35.33	91.86±0.52	92.15±0.57	85.45±0.99	84.97±0.90	53.35±2.45	62.01±2.46

input

$$g_{i,j} = \frac{\partial y_{p,q}}{\partial x_{i,j}} \tag{2}$$

which can be calculated using the backpropagation method [34]. Therefore, the larger absolute value of $g_{i,j}$ corresponds to a larger impact of pixel $x_{i,j}$ on the output ($y_{p,q}$). Furthermore, two-standard deviations were used as the criteria for calculating the ERF size. First, any pixel with an impact gradient value exceeding 1%–95.45% of the center point is considered as in the ERF [34]. The ERF size is then calculated as the square root of the number of pixels within the ERF. In our study, as the input was an eight-channel image, the absolute impact gradient values of each band were first summed and then used to calculate the ERF size. Considering that the ERF calculation is very time-consuming, we only calculated the ERF size on two randomly selected pixels located in the image center area (28 × 28) for each patch in the test set. Therefore, 26 830 pixels were selected to calculate the ERF size.

E. Evaluation Metrics

Two groups of quantitative measures were used to evaluate the performance of cloud detection networks. Four widely used

quantitative metrics, namely, overall accuracy (OA), F-score, mean intersection over union (MIoU), and frequency weighted intersection over union (FWIoU), were used to evaluate the accuracy of the network's cloud region prediction [50], [61]. These metrics are calculated as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\mathrm{MIoU} = \frac{1}{2} \times \left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \right) \tag{5}$$

$$\mathrm{FWIoU} = \frac{P}{P+N} \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} + \frac{N}{P+N} \times \frac{\mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

where TP and TN denote the correct prediction of cloud pixels and clear pixels, respectively, and FP and FN denote an incorrect outcome where the pixels are incorrectly identified as cloud pixels and clear pixels, respectively. Precision is defined as TP/(TP + FP), and Recall is defined as TP/(TP + FN). P and N denote the number of cloud pixels and clear pixels, respectively. As the classification error likely occurs in the boundary areas, the boundary F-score [62], [63] was used to quantitatively measure the boundary quality of the prediction results. This index computes the F-score along the boundary of the predicted mask, given a small slack in distance [63]. There is a slight difference in the calculation of Precision and Recall for the boundary F-score. When calculating Precision, TP denotes a correct prediction of boundary pixels in the dilation of the true boundary; when calculating Recall, TP denotes the true boundary pixels in the dilation of the predicted boundary. Similar to a previous study [63], we used thresholds of three and five pixels as the radius of dilation to calculate the F-score, i.e., BF-3px and BF-5px, in our experiments.

III. Experimental Results and Analysis

A. Comparison of Six Typical Networks

Table II reports the TRF sizes, ERF sizes, and the five evaluation metrics; and Fig. 3 shows the uncertainty of the evaluation metrics induced by random initiation seed. Large differences were observed in the TRF size, ranging from 32 to 1267, for the six networks. Generally, a network with a relatively large TRF size results in a relatively large ERF size after training on the L8 biome dataset. However, the ERF sizes of all six networks were below 39, thus significantly smaller than the corresponding TRF sizes of most networks. Several examples shown in Fig. 4 also demonstrate that only the pixels in the small central area play a role in predicting the cloud segmentation results. With respect to the accuracy, the performance of the network with relatively small ERF and TRF sizes is slightly more optimized. The OA, F-score, MIoU, and FWIoU of the optimal network (MUNet) were approximately 1%, 1%, 2%, and 2% higher than those of the inferior network (DeepLabv3+), respectively. However, such accuracy differences among the six networks are not significant based on one-way ANOVA ($p > 0.05$), considering the effect of random initialization seed [see Fig. 3(a)–(d)].

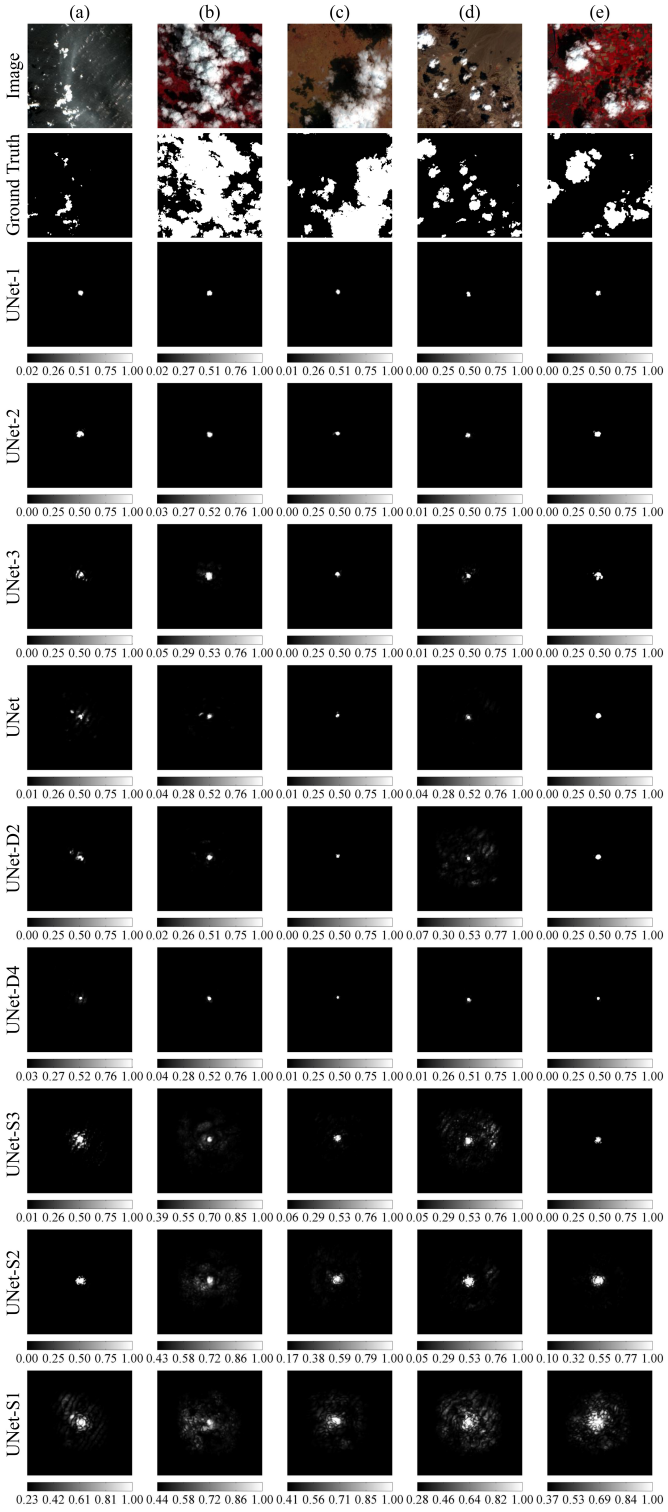


Fig. 7. Visual comparison of ERF for nine UNet variants in different locations. Receptive fields are visualized in the image space of 256 × 256. (a) From path 157 and row 45 (center pixel position: row 125 and column 131). (b) From path 16 and row 50 (center pixel position: row 126 and column 128). (c) From path 175 and row 51 (center pixel position: row 123 and column 124). (d) From path 159 and row 36 (center pixel position: row 128 and column 126). (e) From path 197 and row 24 (center pixel position: row 126 and column 125).

However, the six networks show significant differences in the boundary F-score of three and five pixel thresholds at the level of $p < 0.01$ based on one-way ANOVA, indicating that the main differences among the six networks occur in the cloud boundary. Generally, networks with relatively large TRF and ERF perform poorly in boundary localization. DeepLabv3+ with the largest TRF and ERF underperforms the optimal network (MF-CNN) by more than 4% and 3% in terms of BF-3px and BF-5px, respectively. This result is also demonstrated by the examples in Fig. 5, where all networks show similar prediction results, differing almost only at the cloud boundary. Visually, the boundary of DeepLabv3+ with the largest ERF and TRF was too smooth compared to the ground truth and the results of other networks. In summary, networks with relatively large TRFs or ERFs correspond to relatively low segmentation accuracy of cloud detection, especially in terms of boundary accuracy.

B. Comparison of Nine UNet Variants

Table III reports the TRF sizes, ERF sizes, and six evaluation metrics of nine UNet variants; and Fig. 6 shows the uncertainty of the evaluation metrics induced by random initiation seed. The TRF sizes of the nine networks ranged from 14 to 260 and were, thus, larger than the corresponding ERF sizes ranging from 10 to 66. However, a larger TRF size does not necessarily translate to a larger ERF size for the nine UNet variants in contrast to the result in Section III-A. This finding indicates that UNet variants can effectively learn small-scale spatial features, even for variants with large TRFs. In addition, variants with the same TRF sizes but fewer skip connections result in larger ERFs after training, indicating that skip connections facilitate the exploration of contextual information in small regions. These results can also be demonstrated by the examples shown in Fig. 7, where UNet variants with different depths and different dilated convolutions show similar ERF sizes, whereas UNet variants with fewer skip connections show larger ERF sizes. Regarding accuracy, significant differences were detected in both the region and boundary accuracies of the nine UNet variants based on one-way ANOVA ($p < 0.01$) (see Fig. 6). Generally, the networks with relatively large ERFs exhibited inferior performance in both boundary and nonboundary segmentation; particularly, UNet-S1, which had the largest ERF, underperformed the optimal network (UNet-1) by 7% and 6% in terms of BF-3px and BF-5px, respectively. The examples in Fig. 8 also show that the boundary of UNet-S1 with the largest ERF is too smooth compared to the ground truth and the results of other networks. This result might be attributed to the fact that skip connections effectively combine the high-level features in the expanding path and the low-level features in the contracting path, which can maintain boundary accuracy for networks with large TRFs. In summary, the nine UNet variants with different TRF sizes did not change the ERF size or segmentation accuracy; rather, UNet variants with relatively few skip connections resulted in an increase in ERF size and a decrease in segmentation accuracies, especially boundary accuracies.

C. Relationships Among TRF, ERF, and Accuracies

To elucidate the roles of the receptive field in cloud segmentation, the relationships among TRF size, ERF size, region accuracy, and boundary accuracy were analyzed for all 56 trained networks (four networks were trained for each type of network). As shown in Fig. 9(a) and (f), the TRF and ERF sizes are weakly correlated because ERF sizes are determined by many factors other than TRF. However, both TRF and ERF sizes are significantly negatively correlated with the two types of accuracies [see Fig. 9(b)–(e)], i.e., OA and BF-3px. This finding indicates that the TRF and ERF sizes of the network play important roles in determining segmentation quality. Such negative correlation between TRF size and segmentation accuracies disappeared if only UNet variants (TL-Net, MUNet, and nine UNet variants in Table I; and the subsequent "UNet variants" indicates these 11 networks) were considered [see Fig. 9(g)–(j)], indicating that ERF size is a better indicator of segmentation accuracy than TRF size. Moreover, the negative correlation between ERF size and boundary accuracy yields an R-squared value of 0.4657, showing greater strength than that between the ERF size and region accuracy. These data indicate that the ERF is a key factor affecting the determination of the cloud boundary. These results show that a larger ERF size adversely affects the cloud segmentation accuracy, especially the boundary accuracy.

IV. Discussion

A. Receptive Field Size and Segmentation Accuracies in the Scenes With Different Cloud Distributions

The TRF size is determined only by the network architecture, whereas the ERF size is related to the network parameters and input scenes. Specifically, the ERF size may differ in scenes with different cloud distributions, i.e., cloud thickness, cloud object size, and cloud amount. The network performance might also vary in scenes with different cloud distributions due to the different spectral and spatial features of different cloud types. Therefore, to comprehensively elucidate the effects of the receptive field on network performance, the relationships between TRF/ERF and accuracy were investigated for UNet variants in scenes with different cloud distributions.

The results showed that the ERF sizes and segmentation accuracies could largely differ for different cloud distributions (see Figs. 10–12). For scenes with different relative percentages of thick clouds, both the region and boundary accuracies are substantially lower in the scenes with more thin clouds because the spectra of thin clouds are more difficult to distinguish from the land surface spectra. In addition, ERF sizes are particularly small for the scenes with a mixture of thick and thin clouds (i.e., the "MidFracThickCloud" group in Fig. 10), probably because the fragmented mixture of clouds might lead to a high density of cloud edges; therefore, the spatial contextual features could be exhibited in small regions, corresponding to a small ERF. Regarding the scenes with different cloud object sizes, the corresponding ERF sizes were similar, whereas the segmentation accuracies showed obvious differences. Specifically, the region accuracies were lower in the scenes with small cloud object sizes [see Fig. 11(b-1)] than

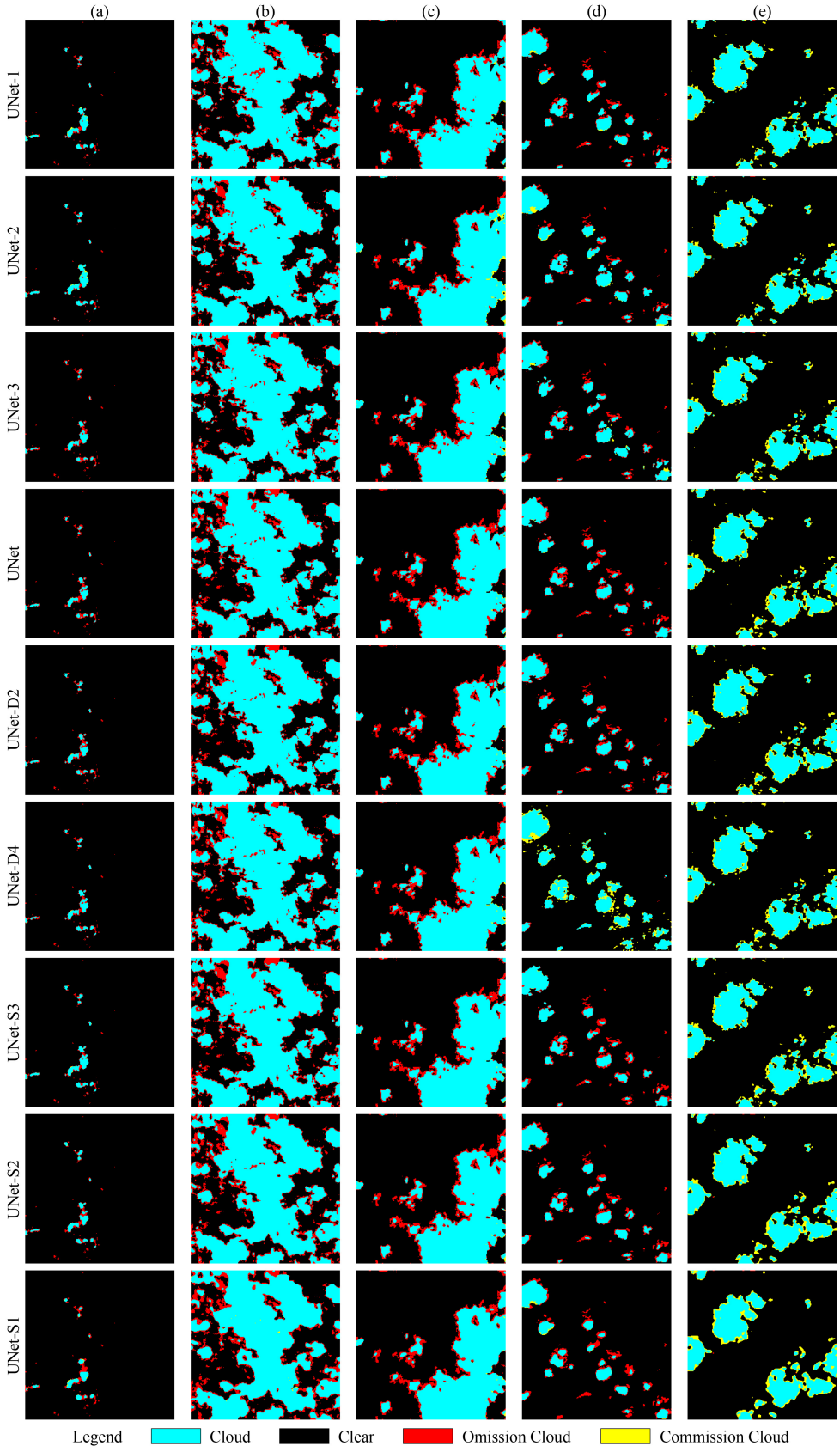


Fig. 8. Qualitative comparison of cloud detection results for nine UNet variants (under the best performing random seed) in different locations. (a) From path 157 and row 45. (b) From path 16 and row 50. (c) From path 175 and row 51. (d) From path 159 and row 36. (e) From path 197 and row 24. The satellite images and the corresponding ground truths are the same as Fig. 5.

those in the scenes with large cloud object sizes [see Fig. 11(b-3)]; in contrast, the boundary accuracies are lower in the scenes with large cloud object sizes [see Fig. 11(d-3)], probably because the detection of small objects (small cloud objects and small clear holes inside large cloud objects) is a relatively challenging task for CNN [64]. For scenes with different cloud percentages, the region and boundary accuracies were significantly lower in the group of scenes with low cloud percentages (see Fig. 11) because there are relatively more small cloud objects and thin clouds in this case. Similarly, the ERF sizes were particularly small for the group of scenes with a middle cloud percentage because of the high density of cloud edges. Despite the differences in the ERF sizes and accuracies for different cloud distributions, significantly negative correlations were observed between ERF size and both the region and boundary accuracies (see Figs. 10–12), except that all UNet variants achieved extremely high region accuracies (OA > 95%) for scenes with thick clouds [see Fig. 10(b-3)] and scenes with few clouds [see Fig. 12(b-1)]. Compared to the ERF size, the TRF size is weakly or insignificantly correlated with the cloud segmentation accuracies. In summary, the relationships between TRF/ERF size and accuracy generally remain unchanged for different cloud distributions, consistent with the results in Section III-C.

B. Receptive Field Size and Segmentation Accuracies in the Scenes With Different Spectral Configurations

Landsat 8 OLI imagery contains abundant spectral information with 11 spectral bands, which potentially provides adequate spectral features for cloud detection. However, the lack of many spectral bands in certain satellite sensors might change the role of the receptive field in cloud detection. Therefore, to understand the effects of the receptive field on network performance in the case of insufficient spectral information, cloud detection experiments with UNet variants were conducted on simulated datasets by removing some bands in Landsat 8 OLI imagery. Two commonly designed spectral configurations for remote sensing, the red/green/blue/NIR (RGBN) bands and RGBN/SWIR-1/SWIR-2 (RGBNS) bands, were considered. When fewer spectral bands were used in the networks, the ERF sizes increased, indicating that spatial contextual information plays a relatively more important role in cloud segmentation (see Fig. 13). Moreover, the negative correlation between TRF/ERF and region accuracy becomes very weak when the RGBN bands are used [see Fig. 13(b-1)–(e-1)]. The TRF size is positively correlated with the boundary accuracy in this case [see Fig. 13(d-1)]. The relationship between ERF and boundary accuracy becomes a quadratic function with an optimal ERF size of 33 for RGBNS data and 70 for RGBN data. These data suggest that cloud detection on imagery with fewer spectral bands would require a larger receptive field. These results are largely different from those presented in Section III-C, indicating that the negative effects of TRF and ERF on cloud detection accuracy are inapplicable to other satellite imagery. Furthermore, the other remote sensors would differ from Landsat 8 OLI in terms of spatial resolution and radiometric resolution. Therefore, the role of the receptive field in the cloud detection of other satellite imagery should be reinvestigated using more datasets.

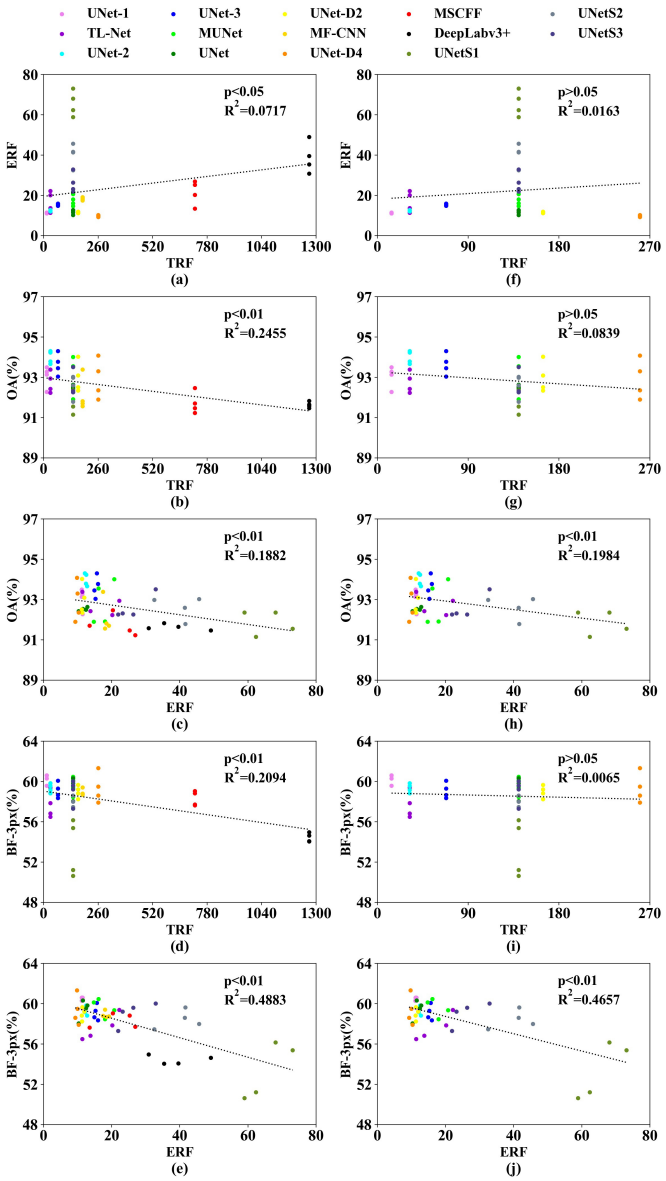


Fig. 9. Relationships among TRF, ERF, and accuracies for (a)–(e) all models and (f)–(j) UNet variants.

C. Implications for the Selection of Cloud Detection Networks

Due to the tradeoff between accurate boundary localization and rich contextual information extraction for different CNNs, comprehending the effects of the TRF and ERF sizes on the cloud segmentation result is essential to selecting a suitable network. The experiments in this study show that networks with relatively small receptive fields yield more optimized region and boundary accuracies for cloud detection in Landsat 8 OLI imagery. Moreover, the ERF sizes were significantly smaller than the corresponding TRF sizes for all of the networks, particularly for MSCFF and DeepLabv3+ with relatively large TRFs, implying that large-scale features

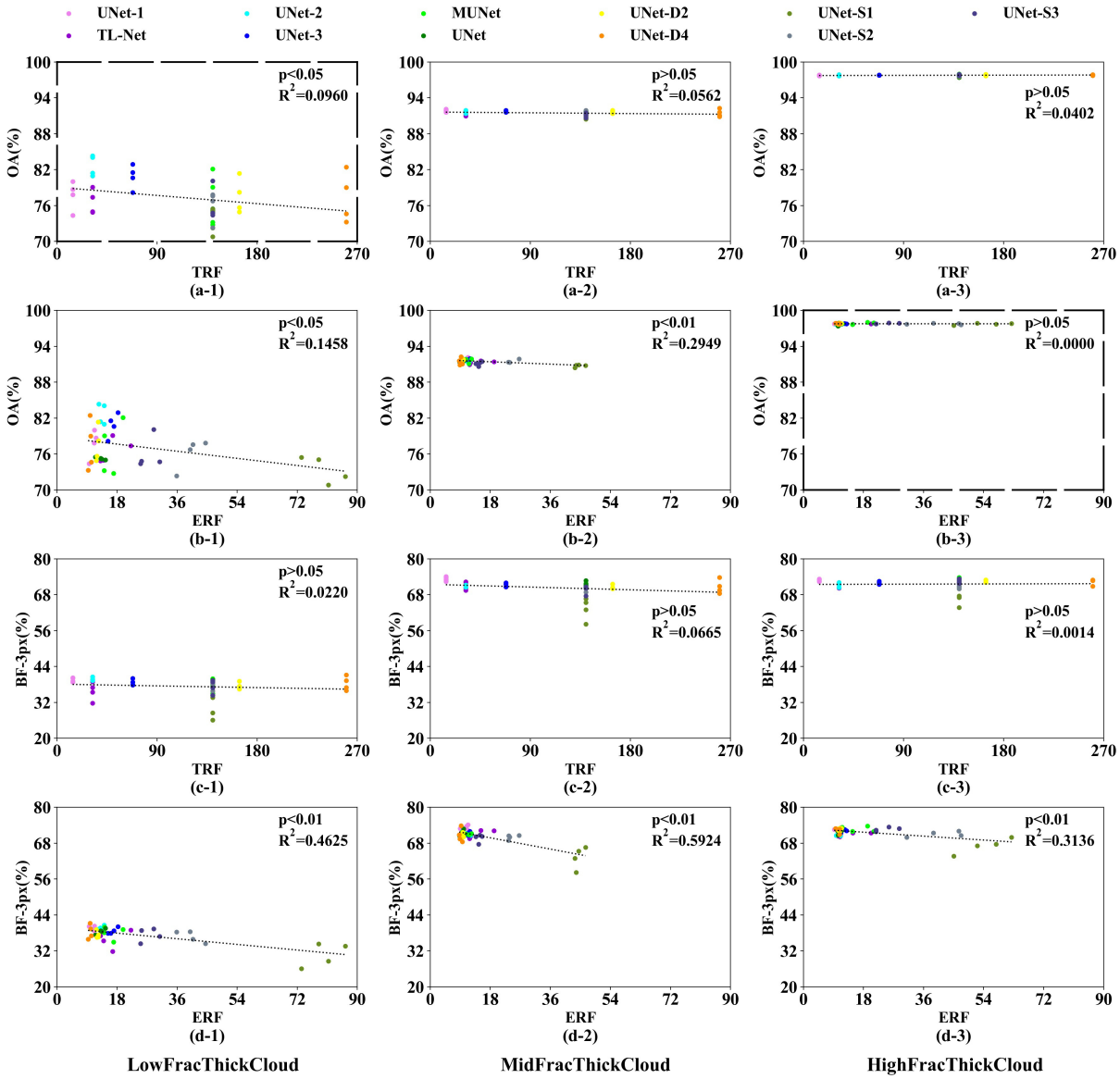


Fig. 10. Relationships between (a) TRF and OA, (b) ERF and OA, (c) TRF and BF-3px, and (d) ERF and BF-3px of UNet variants in scenes with different relative percentages of thick clouds (percentages of thick clouds to all clouds). "LowFracThickCloud" denotes the results for scenes with the relative percentage of thick cloud below 35%, "MidFracThickCloud" denotes the results for scenes with the relative percentage of thick cloud ranging between 35% and 65%, and "HighFracThickCloud" denotes the results for scenes with the relative percentage of thick cloud over 65%. Subplots with dashed and bolded borders indicate that the correlation between the x-axis and the y-axis has changed from significant to insignificant or vice versa.

had not been learned after training. In other words, large-scale features were gradually disregarded during training, and only small-scale features were retained. Therefore, networks with small TRFs are adequate for detecting clouds for Landsat 8 OLI imagery due to their abundant spectral information. The spatial contextual information in the high-level features cannot benefit the cloud detection task; rather, it negatively affects the boundary localization and further reduces the region accuracy.

In addition to the gain in cloud detection accuracy, networks with relatively small TRFs generally contain relatively few parameters and require relatively few computational resources.

These networks generally have lower giga-floating point operations (GFLOPs), resulting in shorter processing time and higher implementation efficiency (see Table IV). The running time in Table IV is based on a platform with two NVIDIA GTX 1080 Ti GPUs. GFLOPs are related to network complexity and the input patch size. Due to the GPU memory limitation, it is necessary to divide the entire remote sensing image into small patches for computing on the GPU. However, the size of each patch cannot be smaller than the ERF size; otherwise, the lack of sufficient spatial information would influence the prediction results. Therefore, using networks with a small TRF could further reduce the computation cost by reducing the input patch size, providing the potential for deployment on more GPU devices with less memory. Therefore, we recommend using TL-Net, which yields both high accuracy and efficiency for cloud detection in Landsat

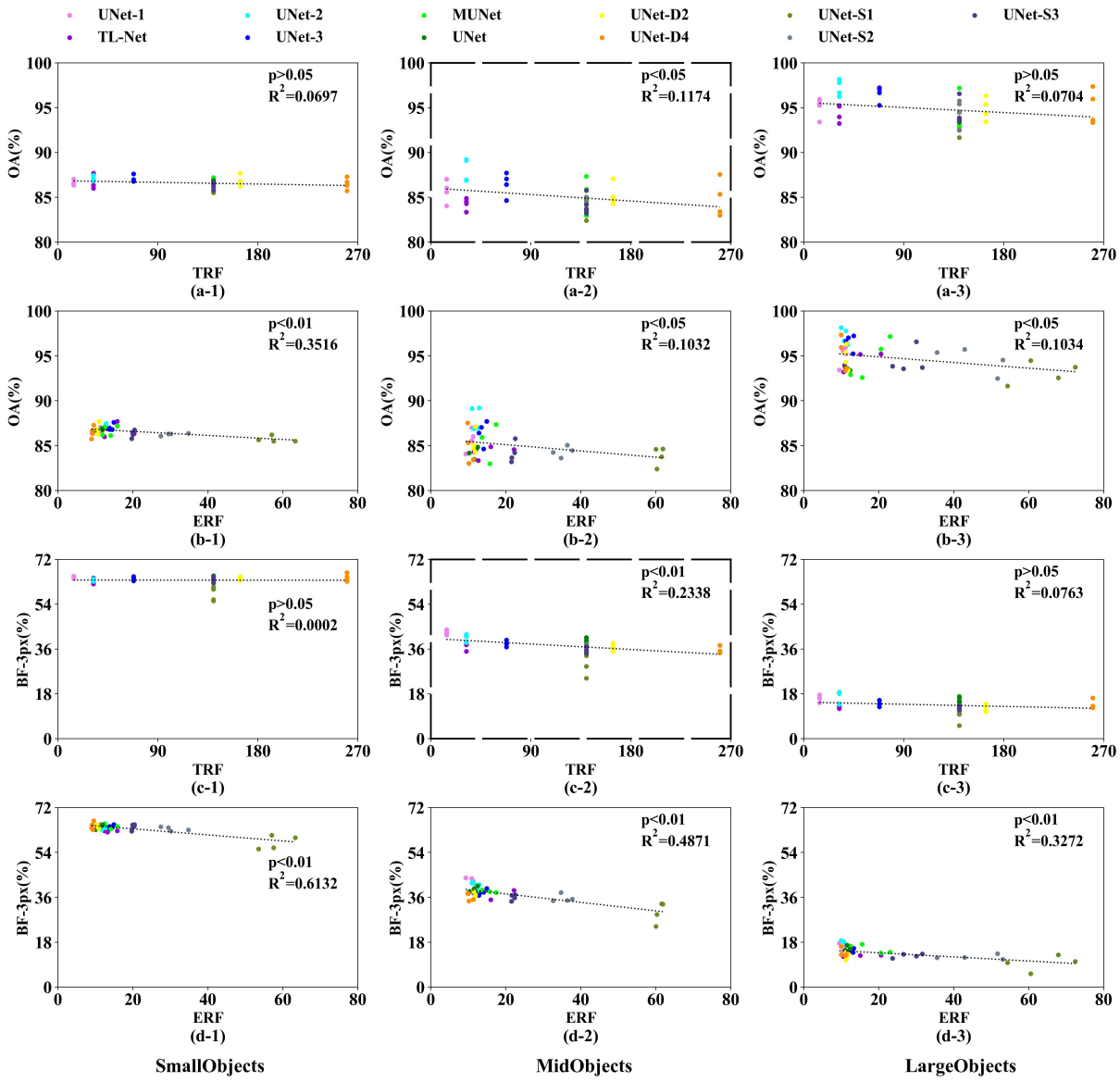


Fig. 11. Relationships between (a) TRF and OA, (b) ERF and OA, (c) TRF and BF-3px, and (d) ERF and BF-3px of UNet variants in scenes with different cloud object sizes. "SmallObjects" denotes the results for scenes with an average cloud object size below 4096 pixels but exceeding zero pixels, "MidObjects" denotes the results for scenes with an average cloud object size ranging between 4096 and 36 864 pixels, and "LargeObjects" denotes the results for scenes with an average cloud object exceeding 36 864 pixels. Subplots with dashed and bolded borders indicate that the correlation between the x-axis and the y-axis has changed from significant to insignificant or vice versa.

8 OLI imagery. However, this suggestion is only applicable to Landsat 8 OLI imagery according to the different results shown in Section IV-B.

D. Future Perspectives

Although the recommended TL-Net with a small receptive field size achieved high region accuracy (F-score higher than 90%), the cloud boundary accuracy (BF-5px less than 70%) was considerably lower than the region accuracy. Visually, the boundaries generated by networks with large receptive field sizes tend to be relatively smooth. We used the fractal dimension index (FRAC) [65]–[67] to quantitatively compare the complexity difference between the cloud boundary predicted by the networks and the reference cloud boundary. As shown in Fig. 14, both TRF and ERF sizes are significantly negatively correlated with the FRAC of the prediction results of the

TABLE IV
PROCESSING TIME AND GFLOPS PER BATCH AND THE NUMBER OF PARAMETERS (PARAMS) OF SIX TYPICAL NETWORKS

Method	Train time (ms)	GFLOPs (Train)	Test time (ms)	GFLOPs (Test)	Params
TL-Net	199.0±3.8	46.90	70.9±5.2	46.90	0.097M
MUNet	237.1±7.1	331.93	83.4±4.2	331.93	7.768M
UNet	736.4±8.6	1317.76	233.9±1.6	1317.76	31.047M
MF-CNN	480.2±8.3	761.96	125.0±2.2	761.96	14.785M
MSCFF	334.7±30.2	3369.13	1305.7±30.3	3369.13	51.883M
DeepLabv3+	676.6±10.7	537.42	349.5±4.2	1566.64	59.355M

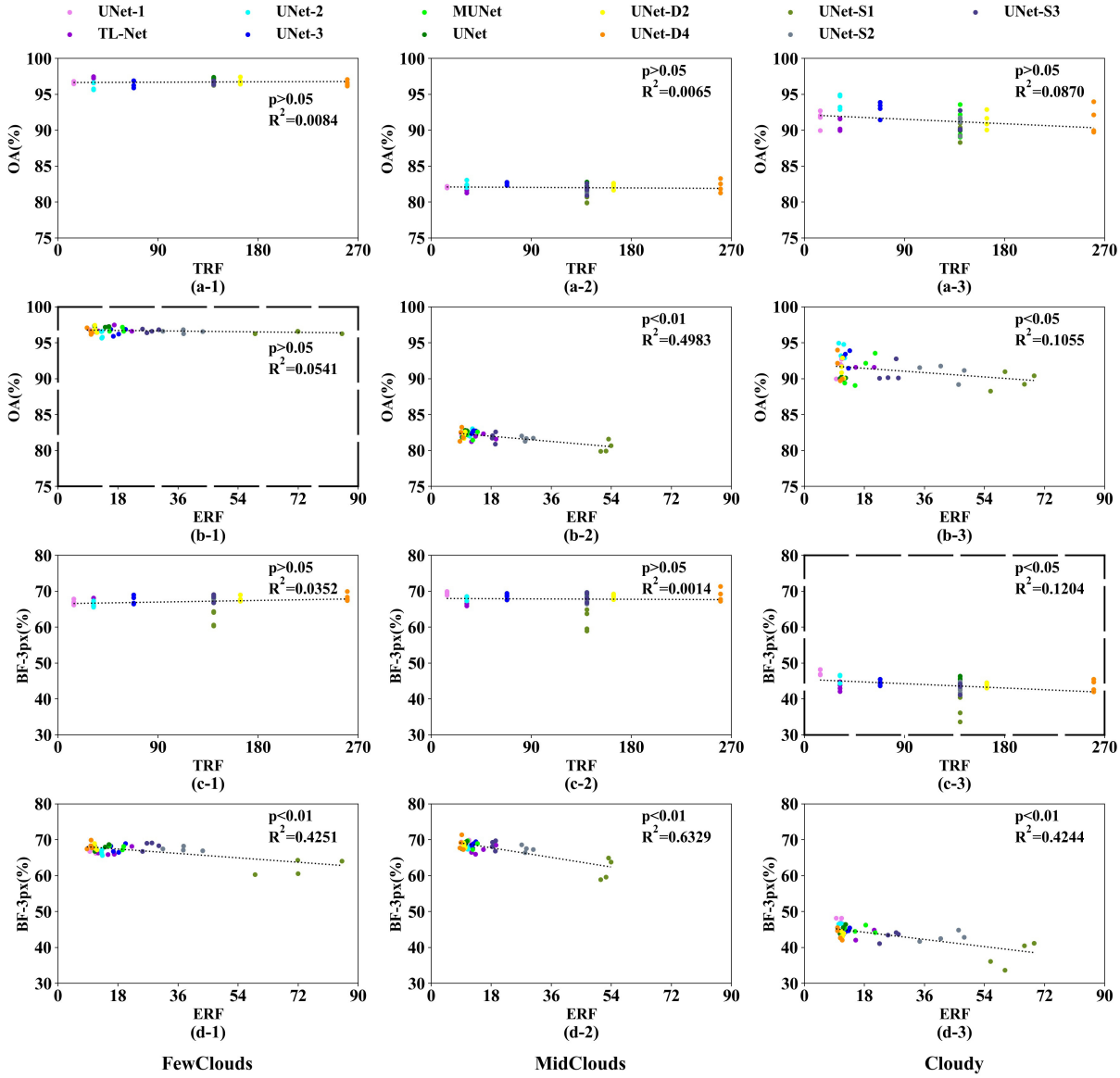


Fig. 12. Relationships between (a) TRF and OA, (b) ERF and OA, (c) TRF and BF-3px, and (d) ERF and BF-3px of UNet variants in scenes with different cloud percentages. "FewClouds" denotes the results for scenes with the cloud percentage less than 35%, "MidClouds" denotes the results for scenes with the cloud percentage ranging between 35% and 65%, and "Cloudy" denotes the results for scenes with a cloud percentage over 65%. Subplots with dashed and bolded borders indicate that the correlation between the x-axis and the y-axis has changed from significant to insignificant or vice versa.

cloud detection networks. In particular, ERF sizes exhibit a remarkable correlation with FRAC for all models or UNet variants with an R square higher than 0.6. In addition, the FRAC of the prediction results of all the models is lower than that of the reference cloud boundary (see Fig. 14). These results indicate that the predicted boundaries of all networks are too smooth, even for networks with small receptive field sizes. Therefore, we suggest that future research on cloud detection networks should focus on enhancing the prediction accuracy on the boundary rather than enlarging the receptive field.

The ERF tool employed in this study illustrates the importance of each input pixel and reveals the spatial scale of the learned features; it may help unveil the "black box" of neural networks and guide network selection and development. Although the ERF size explained approximately 20% variance of OA and 47% variance of BF-3px in the experiments (see Fig. 9), it is far from clearly explaining the behavior of the neural network. Apart from the ERF tool, different techniques have been developed to illustrate how deep neural networks work. For example, the Taylor decomposition [68], [69], layerwise relevance propagation [69], dropout as a Bayesian approximation [70], and DeepLIFT [71] have been proposed to determine the importance of the input pixel to the output unit. Zeiler and Fergus [72] proposed a Deconvnet method to visualize the features learned by deep models. It was then modified by Springenberg *et al.* [73] to further improve the feature visualization. Li *et al.* [74] adapted an attention mechanism to reveal the subtle relations between geological depositions and seismic spectral responses. Wickstrøm *et al.* [75]

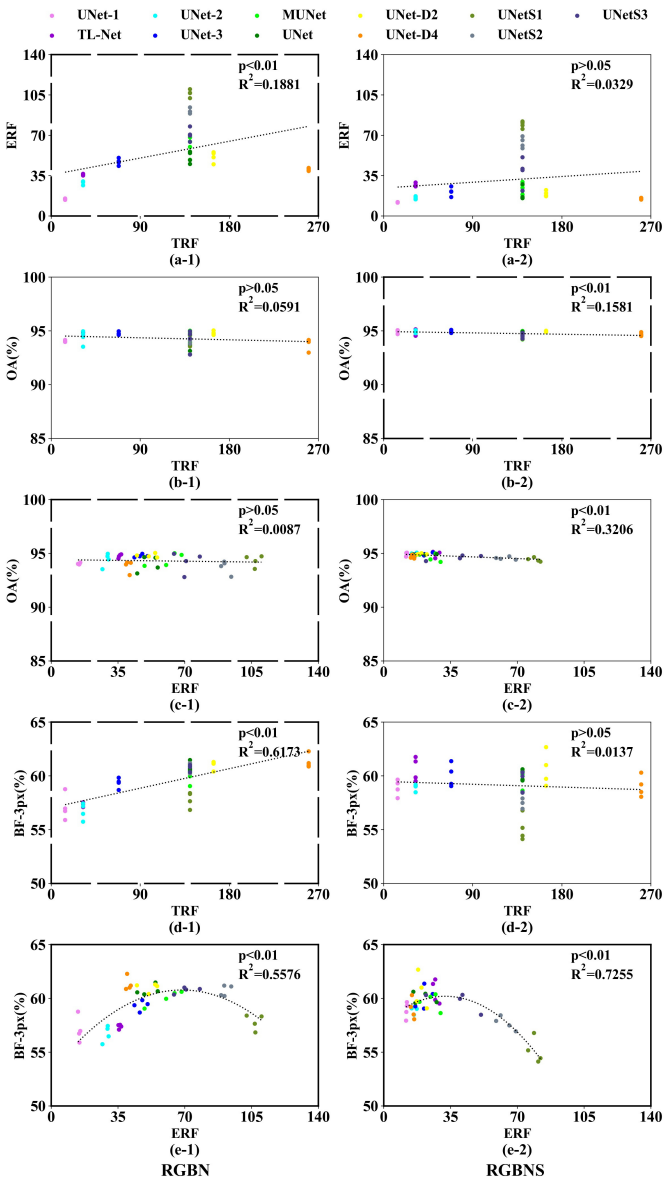


Fig. 13. Relationships among TRF, ERF, and accuracies for UNet variants in scenes with spectral configurations of (a-1)–(e-1) "RGBN" and (a-2)–(e-2) "RGBNS." "RGBN" denotes the results for models using the RGBN bands, and "RGBNS" denotes the results for models using the RGBN/SWIR1/SWIR2 bands. Subplots with dashed and bold borders indicate that the correlation between the x-axis and the y-axis has changed from significant to insignificant or vice versa, and subplots with bolded borders indicate that the scatters are fitted with a quadratic polynomial.

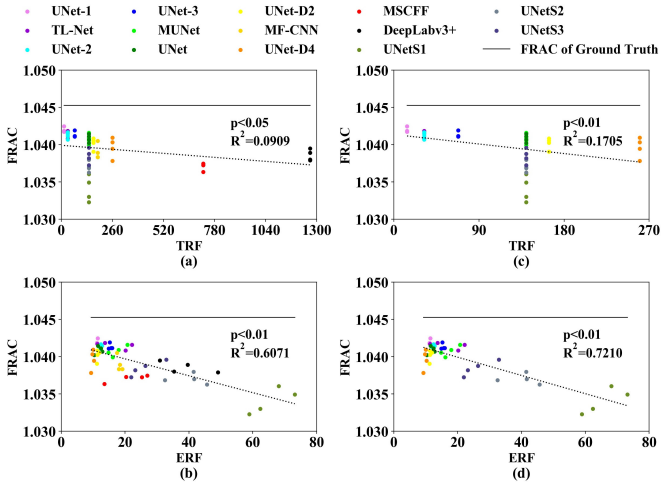


Fig. 14. Relationships between receptive field and predicted cloud boundary quality for all models under each seed. FRAC change with TRF or ERF for (a) and (b) all models and (c) and (d) UNet variants.

V. Conclusion

In this study, we explored the role of the receptive field of six typical networks and nine UNet variants for cloud detection in Landsat 8 OLI imagery. The results show that networks with relatively large receptive field sizes, especially the ERF size, show relatively low segmentation accuracies in scenes with different cloud distributions. Moreover, the ERF sizes of all networks were significantly smaller than the corresponding TRF sizes after training with the L8 Biome dataset, indicating that large-scale contextual information cannot benefit the cloud detection of Landsat 8 OLI imagery. Hence, we recommend using networks with small TRFs (e.g., TL-Net) that can achieve both high accuracy and efficiency for cloud detection in Landsat 8 OLI imagery. Moreover, all networks generated cloud boundaries that were too smooth with substantially lower boundary accuracies than region accuracies. Therefore, boundary localization is the main challenge in cloud detection tasks; thus, we recommend that additional effort should be made to improve the boundary localization rather than enhancing the contextual information exploration over a large receptive field. Finally, the aforementioned conclusions are only applicable for cloud detection of Landsat 8 OLI imagery because these findings vary for data with different spectral bands.

proposed a novel method for estimating the uncertainty in the importance of input features. Such techniques help users determine whether the model detects features that are actually associated with the object or exploits artifacts in the data [75] and, to a certain extent, addresses the lack of transparency and elucidates the underlying mechanics of deep models. However, these methods have yet to be utilized in cloud detection based on deep learning. Therefore, additional attention should be paid to comprehending or visualizing how deep neural networks work on cloud detection to help improve cloud detection networks in the future.

Acknowledgment

The authors would like to thank Junxiong Zhou, Yang Chen, Ms. Yuru Zhao, Licong Liu, Yuean Qiu, and Shuaijun Liu for their beneficial advices on the writing.

References

[1] D. P. Roy *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sens. Environ.*, vol. 145, pp. 154–172, Apr. 2014.

[2] M. A. Wulder *et al.*, "The global Landsat archive: Status, consolidation, and direction," *Remote Sens. Environ.*, vol. 185, pp. 271–283, Nov. 2016.

[3] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, May 2019.

[4] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, pp. 307–316, May 2019.

[5] M. Xu, X. Jia, M. Pickering, and A. J. Plaza, "Cloud removal based on sparse representation via multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2998–3006, May 2016.

[6] R. M. Downs and F. A. Day, *National Geographic Almanac of Geography*. Washington, DC, USA: Nat. Geographic Soc., 2005.

[7] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 7, pp. 3826–3852, Jul. 2013.

[8] C.-C. Liu *et al.*, "Clouds classification from sentinel-2 imagery with deep residual learning and semantic image segmentation," *Remote Sens.*, vol. 11, no. 2, p. 119, Jan. 2019.

[9] L. Sun *et al.*, "Satellite data cloud detection using deep learning supported by hyperspectral data," *Int. J. Remote Sens.*, vol. 41, no. 4, pp. 1349–1371, Feb. 2020.

[10] Y. Guo, X. Cao, B. Liu, and M. Gao, "Cloud detection for satellite imagery using attention-based U-Net convolutional neural network," *Symmetry*, vol. 12, no. 6, p. 1056, Jun. 2020.

[11] Z. Zhu, S. Qiu, B. He, and C. Deng, "Cloud and cloud shadow detection for landsat images: The fundamental basis for analyzing Landsat time series," in *Remote Sensing Time Series Image Processing*. Boca Raton, FL, USA: CRC Press, 2018, pp. 25–46.

[12] S. Mahajan and B. Fataniya, "Cloud detection methodologies: Variants and development—A review," *Complex Intell. Syst.*, vol. 6, no. 2, pp. 251–261, 2020.

[13] W. B. Rossow and L. C. Garder, "Cloud detection using satellite measurements of infrared and visible radiances for ISCCP," *J. Climate*, vol. 6, no. 12, pp. 2341–2369, Dec. 1993.

[14] R. R. Irish, J. L. Barker, S. N. Goward, and T. Arvidson, "Characterization of the landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 10, pp. 1179–1188, 2006.

[15] D. P. Roy *et al.*, "Web-enabled landsat data (WELD): Landsat ETM+ composited mosaics of the conterminous United States," *Remote Sens. Environ.*, vol. 114, no. 1, pp. 35–49, Jan. 2010.

[16] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in landsat imagery," *Remote Sens. Environ.*, vol. 118, pp. 83–94, Mar. 2012.

[17] M. Xu, X. Jia, M. Pickering, and S. Jia, "Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 215–225, Mar. 2019.

[18] X. Zhu and E. H. Helmer, "An automatic method for screening clouds and cloud shadows in optical satellite image time series in cloudy regions," *Remote Sens. Environ.*, vol. 214, pp. 135–153, Sep. 2018.

[19] S. Chen, X. Chen, J. Chen, P. Jia, X. Cao, and C. Liu, "An iterative haze optimized transformation for automatic cloud/haze detection of Landsat imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2682–2694, May 2016.

[20] J. Peng *et al.*, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geosci. Remote Sens. Mag.*, early access, Jun. 10, 2021, doi: 10.1109/MGRS.2021.3075491.

[21] W. Sun, K. Ren, X. Meng, C. Xiao, G. Yang, and J. Peng, "A band divide-and-conquer multispectral and hyperspectral image fusion method," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2020.3046321.

[22] Y. Li, R. Fu, X. Meng, W. Jin, and F. Shao, "A SAR-to-optical image translation method based on conditional generation adversarial network (cGAN)," *IEEE Access*, vol. 8, pp. 60338–60343, 2020, doi: 10.1109/ACCESS.2020.2977103.

[23] J. Wei *et al.*, "Cloud detection for landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches," *Remote Sens. Environ.*, vol. 248, Oct. 2020, Art. no. 112005.

[24] M. J. Hughes and D. J. Hayes, "Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing," *Remote Sens.*, vol. 6, no. 6, pp. 4907–4926, May 2014.

[25] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, Apr. 2019.

[26] S. Zhenfeng *et al.*, "Cloud detection in remote sensing images based on multiscale features-convolutional neural network," *IEEE Trans. Geosci. Remote Sens.,* vol. 57, no. 6, pp. 4062–4076, Jun. 2019.

[27] M. Wieland, Y. Li, and S. Martinis, "Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network," *Remote Sens. Environ.*, vol. 230, Sep. 2019, Art. no. 111203.

[28] G. Mateo-García, V. Laparra, D. López-Puigdollers, and L. Gómez-Chova, "Transferring deep learning models for cloud detection between landsat-8 and proba-V," *ISPRS J. Photogramm. Remote Sens.*, vol. 160, pp. 1–17, Feb. 2020.

[29] Y. Shendryk *et al.*, "Deep learning for multi-modal classification of cloud, shadow and land cover scenes in planetscope and sentinel-2 imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 157, pp. 124–136, Nov. 2019.

[30] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDnet: CNN-based cloud detection for remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 6195–6211, Aug. 2019.

[31] J. Guo, J. Yang, H. Yue, H. Tan, C. Hou, and K. Li, "CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 700–713, Jan. 2020.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, Feb. 2015.

[33] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, Mar. 2015.

[34] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4905–4913.

[35] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

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

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[38] D. Zhou *et al.*, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[39] 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.

[40] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[41] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

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

[43] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.

[44] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[45] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, Sep. 2020.

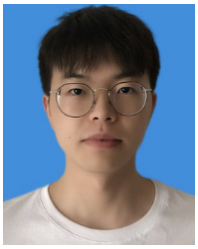
[46] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[47] L. Jiao, L. Huo, C. Hu, and P. Tang, "Refined UNet: UNet-based refinement network for cloud and shadow precise segmentation," *Remote Sens.*, vol. 12, no. 12, p. 2001, Jun. 2020.

[48] X. Wu and Z. Shi, "Utilizing multilevel features for cloud detection on satellite imagery," *Remote Sens.*, vol. 10, no. 10, p. 1853, 2018.

[49] Y. Chen, A. Dapogny, and M. Cord, "SEMEDA: Enhancing segmentation precision with semantic edge aware loss," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107557.

[50] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
[51] M. Oršić and S. Šegvić, "Efficient semantic segmentation with pyramidal fusion," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107611.
[52] S. Foga *et al.*, "Cloud detection algorithm comparison and validation for operational landsat data products," *Remote Sens. Environ.*, vol. 194, pp. 379–390, Jun. 2017.
[53] S. Qiu, B. He, Z. Zhu, Z. Liao, and X. Quan, "Improving Fmask cloud and cloud shadow detection in mountainous area for landsats 4–8 images," *Remote Sens. Environ.*, vol. 199, pp. 107–119, Sep. 2017.
[54] L. Peng, L. Liu, X. Chen, J. Chen, X. Cao, and Y. Qiu, "Generalization ability of cloud detection network for satellite imagery based on DeepLabv3+," *Nat. Remote Sens. Bull.*, vol. 25, no. 5, pp. 1169–1186, 2021.
[55] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, "Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN)," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111446.
[56] M. Luotamo, S. Metsamaki, and A. Klami, "Multiscale cloud detection in remote sensing images using a dual convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4972–4983, Jun. 2021.
[57] X. Wu, Z. Shi, and Z. Zou, "A geographic information-driven method and a new large scale dataset for remote sensing cloud/snow detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 87–104, Apr. 2021.
[58] S. Shah, P. Ghosh, L. S. Davis, and T. Goldstein, "Stacked U-nets: A no-frills approach to natural image segmentation," 2018, *arXiv:1804.10343*.
[59] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 437–478.
[60] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, "The receptive field as a regularizer in deep convolutional neural networks for acoustic scene classification," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.
[61] 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.
[62] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 724–732.
[63] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5229–5238.
[64] 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./Jul. 2016, pp. 1–9.
[65] B. B. Mandelbrot, "Fractals," in *Form Chance Dimensional*. New York, NY, USA: Freeman, 1977.
[66] B. B. Mandelbrot and B. B. Mandelbrot, *The Fractal Geometry of Nature*, vol. 1. New York, NY, USA: WH Freeman, 1982.
[67] K. McGarigal, S. A. Cushman, and E. Ene, "FRAGSTATS v4: Spatial pattern analysis program for categorical and continuous maps," Comput. Softw. Program Prod. Authors, Univ. Massachusetts Amherst, Amherst, MA, USA, 2012. [Online]. Available: https://www_Umass_Edulandecoresearchfragstatsfragstats_Html
[68] S. Bazen and X. Joutard, "The Taylor decomposition: A unified generalization of the Oaxaca method to nonlinear models," Aix-Marseille Univ., Marseille, France, Tech. Rep. 2013-32, 2013.
[69] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.
[70] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 2016, pp. 1050–1059.
[71] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[72] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[73] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.
[74] F. Li, H. Zhou, Z. Wang, and X. Wu, "ADDCNN: An attention-based deep dilated convolutional neural network for seismic facies analysis with interpretable spatial–spectral maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1733–1744, Feb. 2021.
[75] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101619.



Longkang Peng received the B.E. degree in spatial informatics and digitalized technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2019. He is currently pursuing the M.S. degree with Beijing Normal University, Beijing, China.

His research interests include remote sensing classification and segmentation.

Xuehong Chen received the B.S. degree in physics and the M.S. degree in civil engineering from Beijing Normal University, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in Earth and environmental science from Nagoya University, Nagoya, Japan, in 2012.

He is currently an Associate Professor with the Faculty of Geographical Science, Institute of Remote Sensing Science and Engineering, Beijing Normal University. His research interests include preprocessing, data fusion of satellite images, and land cover mapping by remote sensing.

Jin Chen received the B.A. and M.S. degrees in geography from Beijing Normal University, Beijing, China, in 1989 and 1992, respectively, and the Ph.D. degree in civil engineering from Kyushu University, Fukuoka, Japan, in 2000.

He was a Post-Doctoral Researcher with the University of California at Berkeley, Berkeley, CA, USA, from 2000 to 2001, and the National Institute for Environmental Studies, Tsukuba, Japan, from 2001 to 2004. He is currently a Professor with the Faculty of Geographical Science, Institute of Remote Sensing Science and Engineering, Beijing Normal University. His research interests include remote sensing modeling and vegetation parameter retrieval through the inversion of remote sensing models.

Wenzhi Zhao received the Ph.D. degree from Peking University, Beijing, China, in 2018.

In 2017, he was a Visiting Scholar with the University of Colorado, Boulder, Boulder, CO, USA. He is currently a Lecturer with Beijing Normal University, Beijing. His research interests include remote sensing big data, spatial–temporal data mining, and machine learning, especially deep networks and their applications in remote sensing.

Xin Cao received the B.S. and M.S. degrees in geography from Beijing Normal University, Beijing, China, in 2002 and 2005, respectively, and the Ph.D. degree in environmental engineering from Nagoya University, Nagoya, Japan, in 2008.

From 2008 to 2009, he was a Researcher with Nagoya University. He is currently an Associate Professor with the Faculty of Geographical Science, Institute of Remote Sensing Science and Engineering, Beijing Normal University. His research interests include urban remote sensing and remote sensing modeling in grassland.