http://communications.science.ankara.edu.tr

Commun.Fac.Sci.Univ.Ank.Ser. A2-A3
Volume 67, Number 1, Pages 27–42 (2025)
https://doi.org/10.33769/aupse.1498035
ISSN 1303–6009 E-ISSN 2618–6462

*Research Article; Received:June 08, 2024; Accepted: August 22, 2024*

# EFFECT OF DILATION RATE ON NESTED U-NET MODEL PERFORMANCE IN REMOTE SENSING

Irem ULKU[1]
[1]Department of Computer Engineering, Ankara University, Ankara, TÜRKİYE

**Abstract**. High spatial resolution remote sensing images contain substantial detailed multi-scale objects. Convolutional neural networks (CNNs) are not efficient enough for detecting these objects of varying sizes. Among the multitude of CNN approaches, the Nested U-Net (UNet++) model shows great potential to capture more complex details by progressively enriching high-resolution feature maps. However, there is more room for improving the Nested U-Net architecture by increasing its ability to detect multi-scale objects. The nested blocks used in this architecture rely on standard convolutional layers, which are of limited efficacy in capturing pixel information. Thus, larger receptive fields are required to extract multi-scale feature information. Although many approaches are available for increasing the receptive fields in the Nested U-Net model, these methods usually make the computational efforts very heavy. Therefore, this study uses dilated convolutions in the Nested U-Net architecture to broaden the receptive field without augmenting computational demand. To this extent, the paper performs experiments with different dilation rates in the convolution blocks to understand the benefits of employing dilated convolutions in Nested U-Net architecture. Experiments using two remote sensing image sets show that the Nested U-Net model with dilated convolutions performs well for images containing both visible and multispectral wavelengths. While being able to provide performance improvement, experimental results also demonstrate that only the optimal dilation rate scheme in the proposed approach is beneficial.

**Keywords.** Semantic segmentation, remote sensing, dilated convolution, multispectral images, Nested U-Net model.

## 1. Introduction

Compared with natural images obtained from the ground, high-resolution remote sensing images have a more complex detailed background with multi-scale objects. However, most of the CNNs only provide partial solutions that are insufficient for

irem.ulku@ankara.edu.tr-Corresponding author; 0000-0003-4998-607X; ror.org/02v9bqx10

2025 Ankara University
Communications Faculty of Sciences University of Ankara Series A2-A3: Physical Sciences and Engineering
The copyright of the works published in our journal belongs to the authors and these works are published as open access under the CC-BY 4.0 license.

objects of varying sizes. Since CNN models downsample feature maps through a series of pooling operations in each layer to acquire multi-scale contextual information, it leads to a loss of low-level local details for small targets [1]. In literature, many efforts are devoted to developing CNN-based models that can capture multi-scale details within remote sensing imagery [2–5].

The Nested U-Net [6] architecture can capture more complex details by progressively enriching high-resolution feature maps. The model is hypothesized based on adding nested convolution blocks with intense skip connections to achieve an easier optimization. Recent studies such as CSAN-UNet [7] and MSNUNet [8] focus on enhancing the performance of the Nested U-net through various mechanisms for multi-scale objects. There are also some studies in medical imaging, such as the LiM-Net [9], which utilizes the Nested U-net model and aims to extract multiscale fine-grained features. However, the Nested U-Net architecture demands further refinement to capture different object shapes [10].

An architecture inspired by densely connected convolutional networks, the Nested U-Net integrates U-Net models at various depths. Within each nested block, several convolutional layers extract semantic information. However, these standard convolutional layers can capture only a limited pixel information. Therefore, the network can utilize larger receptive fields to improve its ability to capture multi-scale feature information [11]. Previous methods utilize scale parameters [12], attention modules [13, 14], and transformer blocks [15] for increasing the receptive field of CNN networks in remote sensing. Although these methods yield high accuracies by increasing the receptive fields to capture different shapes and appearances, they also suffer from a high computational burden.

Dilated convolution [16] expands the receptive field without increasing parameters or computation, enhances the resolution of output feature maps, and effectively acquires multi-scale features. There are already efforts to incorporate dilated convolutions to the Nested U-Net architecture to encompass objects at multiple scales. In one study, dilated convolutions are utilized only for the top layer, enabling the Nested U-Net model to capture more comprehensive feature information at full resolution [17]. Another study [18] proposes a method for extracting more contextual information in the Nested U-Net model by relying on the pyramid dilation technique. Influenced by the Nested U-Net, the ConDinet++ model [19] uses conditional dilated convolutions to obtain more contextual semantic information in cases of narrow and occluded roads in aerial images. A-DenseUNet [20] incorporates multiple dilated convolutions with various atrous rates in the Nested U-net network to attain a larger field of view and prevent spatial feature information loss. Another Nested U-Net-based model is the AEUNet++ with multi-task learning and attention mechanisms, which automatically extracts small and large buildings with precise boundaries from high spatial resolution imagery [21].

This study investigates the impact of replacing convolutional layers with dilated convolutional layers within the convolution blocks of the Nested U-Net architecture,

examining how the dilation factor influences the semantic segmentation model performance. The organization of this article is outlined in the following way: Section 2 furnishes detailed information on the Nested U-Net architecture, dilated convolution, and their integration. Section 3 describes the remote sensing image sets used, and Section 4 explains the experimental configuration and evaluation metrics. Section 5 elaborates on the findings obtained from the experiments. Section 6 delivers the study's conclusion.

## 2. Methodology

This part explains the proposed Nested U-Net model design with dilated convolutions. Section 2.1 explains the Nested U-Net architecture, while Section 2.2 discusses the dilated convolution operation.

### 2.1. **Nested U-Net Model.**

The Nested U-Net topology, as shown in Figure 1, is pyramid-shaped and comprises encoding and decoding parts linked by skip pathways. These dense skip connections enable the model to use the extracted contextual information more efficiently and create a global context. The output feature map of block $l$, denoted as $x_{l,j}$, is represented by Equation 1. Here, $l$ represents the $l^{th}$ path in the horizontal direction, while $j$ indicates the $j^{th}$ convolution block in the vertical path:

$$x_{l,j} = \begin{cases} \mathcal{H}(x_{l-1,j}), & \text{if } j = 0 \\ \mathcal{H}([x_{l,k}]_{k=0}^{j-1},\ U(x_{l+1,j-1})), & \text{if } j > 0 \end{cases} \tag{1}$$

where, $\mathcal{H}(.)$ represents the operations in the convolution block, $U(.)$ describes the upsampling and [.] symbol refers to the concatenation operation.

As shown in Equation 1 and illustrated in Figure 1, the input of the convolution block when $j = 0$ is defined as the feature map obtained from the $l - 1^{th}$ block. If $j > 0$, the convolution block's input comprises two distinct sections. One part represents the combination of the feature maps of all previous blocks in the same horizontal path, denoted as $[x_{l,k}]_{k=0}^{j-1}$. The other part is the output of the $j - 1^{th}$ horizontal path, expressed as $x_{l+1,j-1}$.

Figure 2 illustrates the detailed operations of the first skip pathway at the top horizontal level of the pyramid where $l = 0$. In this context, the connection between $x^{0,0}$ and $x^{0,4}$ consists of three convolution blocks and dense skip connections. Each $x^{0,l}$ convolution block operates a concatenation between the output maps from preceding blocks at the matching horizontal level with the output feature map of the associated block from the lower horizontal level after upsampling. For instance, the feature maps $x^{0,0}$, $x^{0,1}$, $x^{0,2}$, and $U(x_{1,2})$ are combined for $x^{0,3}$. The Nested U-Net design underlies a hypothesis of achieving semantically similar feature maps passed from the feature extractor to the relevant decoder.

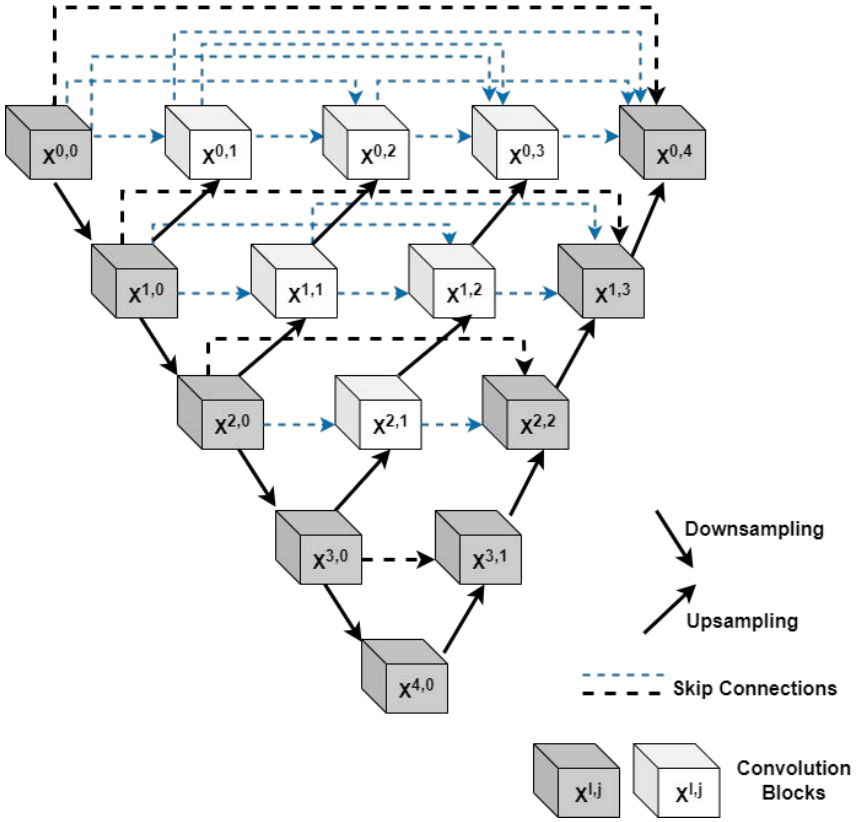



FIGURE 1. The overall framework for Nested U-Net architecture.

2.2. **Dilated Convolution.** As shown in Figure 3, the $2D$ dilated convolution operation within the $j^{th}$ convolution block in any $l^{th}$ horizontal path extracts features at intervals specified by the dilation rate $d$ for each spatial location $(h, w)$ and is defined as follows:

$$y^{(h,w)} = \sum_{i=1}^{H} \sum_{k=1}^{W} x^{(h+d\times i, w+d\times k)} \omega\left(i,k\right), \tag{2}$$

where $y^{(h,w)}$ and $x^{(h,w)}$ represent the output and input of the dilated convolution operation at position $(h, w)$, respectively. $\omega\left(i,k\right)$ denotes the convolutional filter with indices $i$ and $j$. If $d = 1$, the dilated convolution becomes a standard convolution. For a $k \times k$ convolution kernel, the effective kernel size of the dilated convolution operation becomes $kd \times kd$, where $kd$ is defined as follows:

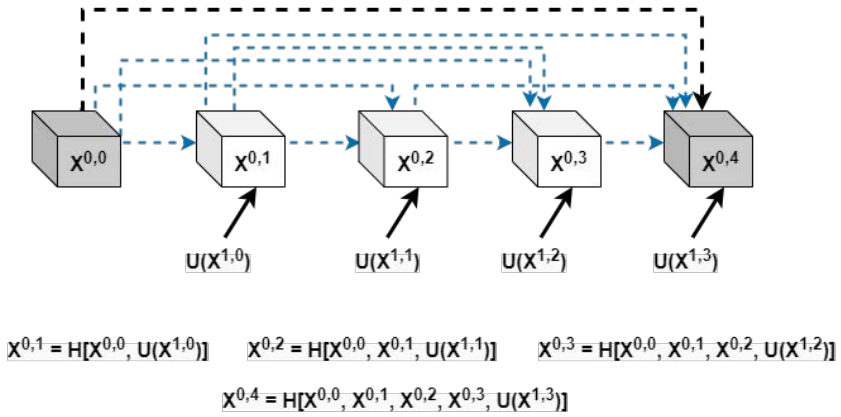



FIGURE 2. The first skip pathway of Nested U-Net architecture.

$$kd = k + (k-1)\,.\,(d-1)\,. \tag{3}$$

Here, there are still only $k \times k$ kernel parameters, which allows changing the receptive field by selecting different dilation rates with the same number of parameters. This approach preserves the spatial resolution. As shown in Figure 3, when a $3 \times 3$ convolutional kernel is used with $d = 2$, the feature maps are sampled as if using a $5 \times 5$ standard convolutional kernel, and when $d = 3$, the feature maps are sampled as if using a $7 \times 7$ standard convolutional kernel.

Figure 4 shows the structure of convolution blocks used in the Nested U-Net architecture. When the input tensor of the convolution block is $x^{l,j}$, then the output $y^{l,j}$ is obtained by passing the input through a standard $3 \times 3$ convolution followed by the feature-wise normalization (abbraviated as BN) and the rectified linear unit (abbraviated as ReLU) activation layers twice, as follows:

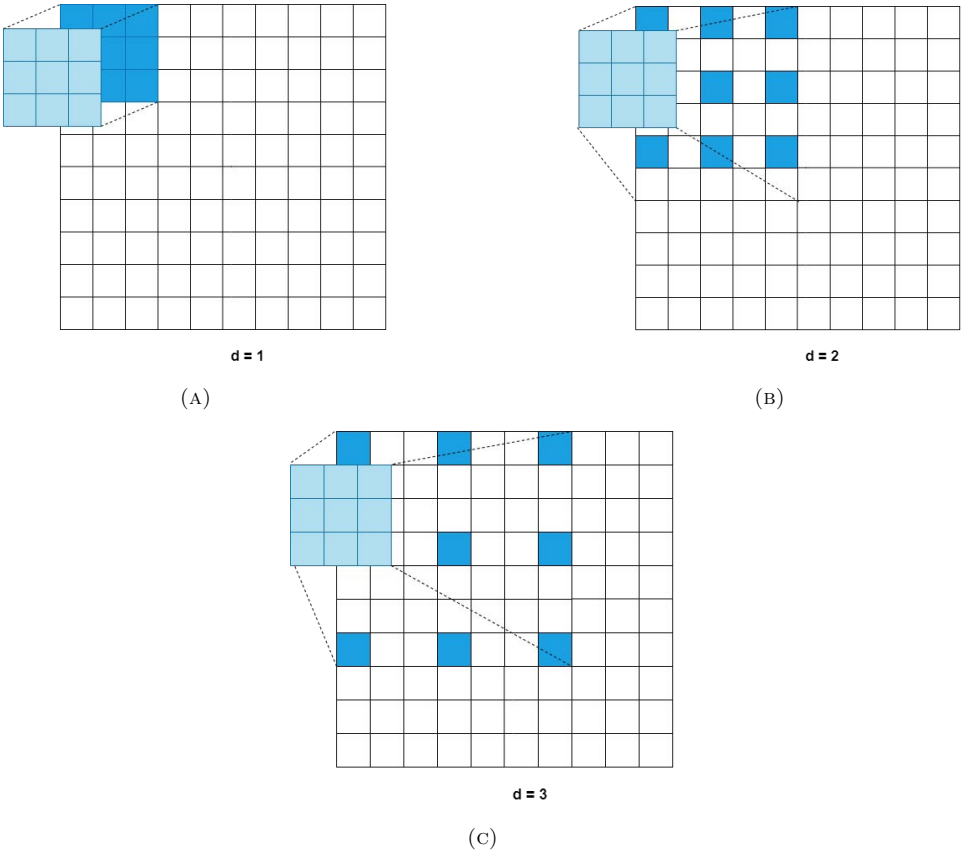
$$y^{l,j} = \mathrm{conv}_{3\times3}(\mathrm{conv}_{3\times3}(x^{l,j})), \tag{4}$$

where $\mathrm{conv}_{3\times3}$ represents the standard $3 \times 3$ convolution (with BN and ReLU).

This study proposes to apply a $3 \times 3$ dilated convolution to secure a larger receptive field than the standard $3 \times 3$ convolution (Figure 4B). The output of the $j^{th}$ convolution block at the $l^{th}$ horizontal level, $z^{l,j}$, is calculated as follows:

$$z^{l,j} = \mathrm{conv}^{d^1}_{3\times3}(\mathrm{conv}^{d^2}_{3\times3}(x^{l,j})), \tag{5}$$

where $\mathrm{conv}^{d^1}_{3\times3}$ and $\mathrm{conv}^{d^2}_{3\times3}$ stand for $3 \times 3$ dilated convolutions with dilation rates $d^1$ and $d^2$, respectively (with BN and ReLU). The experiments employ dilated convolutions with different rates. The flowchart of the entire process is shown in Figure 5.




(A) (B)

(C)

FIGURE 3. Schematics of dilated convolution. (A) Standard convolution. (B) Dilated convolution ($d = 2$). (C) Dilated convolution ($d = 3$).

## 3. Image Sets

In the experiments, multispectral images are in the form of Normalized Difference Vegetation Index (NDVI) to reflect green vegetation accurately, with the formula depending on the relation between the red (abbraviated as R) and the near-infrared (abbraviated as NIR) spectral bands as delineated:

$$NDVI = \frac{NIR - R}{NIR + R}. \tag{6}$$

The following sections describe the image sets used in the experiments, with more detailed information available in the reference study [22].

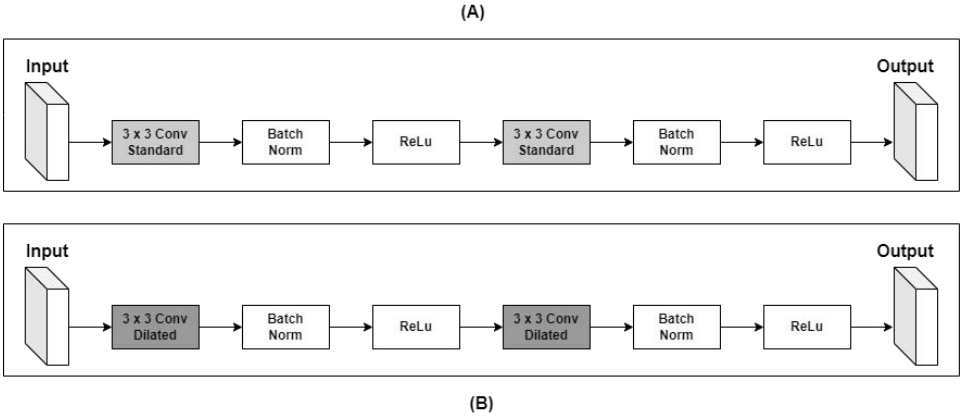



FIGURE 4. (A) The details of the convolution block in Nested U-net architecture. (B) The details of the proposed convolution block incorporating dilated convolutions in Nested U-net architecture.

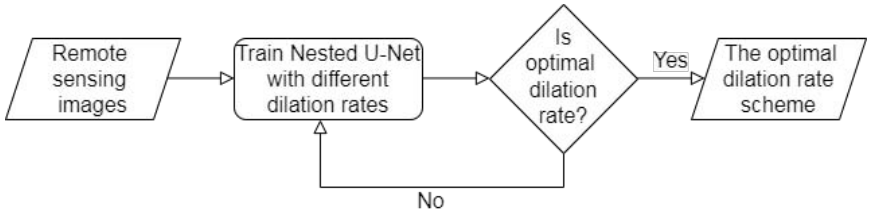
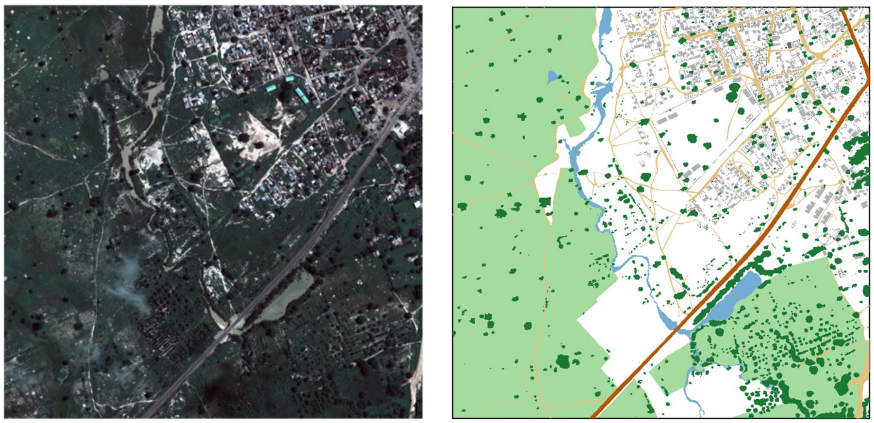



FIGURE 5. The flowchart of the entire process.

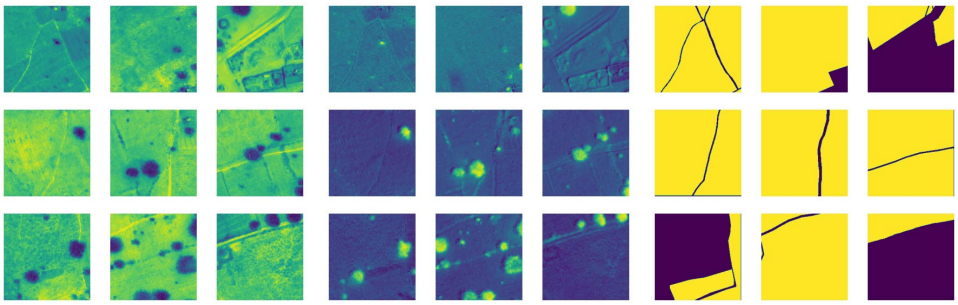
3.1. **DSTL Satellite Image Set.** The DSTL (Defense Science and Technology Laboratory) provides this image set for the Kaggle competition, which includes 25 satellite images with 1 km × 1 km size. Experiments perform binary semantic segmentation of crops in this image set. Figure 6 shows a sample image with its ground truth mask by representing crop pixels in light green. DSTL contains RGB imagery with a resolution of 3348 × 3392 pixels and a spatial precision of 0.31 m. Multispectral images in DSTL cover the wavelength range of 400–1040 nm, with 837 × 848 pixels and 1.24 m spatial resolution. It contains 5985 patches of 224 × 224 pixels, where Figure 7 depicts some examples of RGB and NDVI image patches together with the ground truth masks. These masks include pixels of wheat, potato, and turnip crops, all appearing in yellow.

3.2. **RIT-18 Aerial Image Set.** The image obtained with an octocopter equipped with a Tetracam MicroMCA6 multispectral sensor, with dimensions of 9393 × 5642 pixels, is depicted in Figure 8. Experiments conduct only the semantic segmentation



(A) (B)

FIGURE 6. DSTL satellite set images (A) An example image is given. (B) A ground truth is given by highlighting the pixels of the class crop with light green.



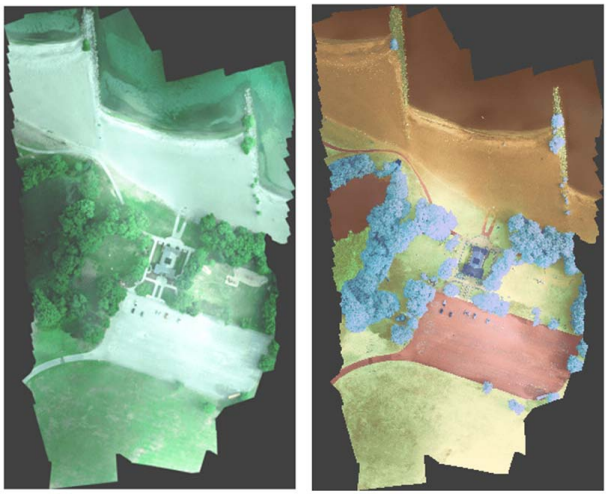
(A) (B) (C)

FIGURE 7. Illustrations of example image patches (DSTL) (A) False-color RGB image patches (B) False-color NDVI image patches (C) Crop class ground truth masks.

of the tree class from this image set, where the blue color identifies the image pixels belonging to this class in Figure 9. This set has a spatial image resolution of 0.047 m with RGB, 715–725 nm, 795–805 nm, and 890–910 nm wavelength ranges. The total sample size is 1778, with each patch having the same 224 × 224-pixel size.

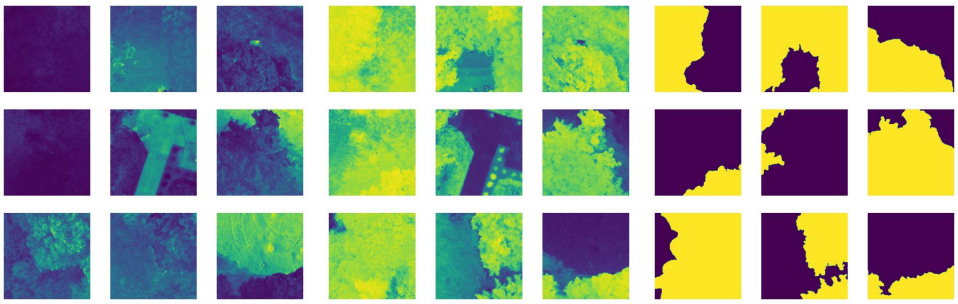
## 4. Experimental Setup

All the experiments are conducted with PyTorch using an NVIDIA Quadro RTX 5000 GPU. The hardware used is a server with CPU Intel(R) Xeon(R) Gold 6240R



(A) (B)

FIGURE 8. RIT-18 aerial set images (A) An example image is given. (B) A ground truth is given by highlighting the pixels of the class tree with blue.



(A) (B) (C)

FIGURE 9. Illustrations of example image patches (RIT-18) (A) False-color RGB image patches (B) False-color NDVI image patches (C) Tree class ground truth masks.

@2.40GHz and the size of RAM is 128 GB. The optimization algorithm chosen is Adam with corresponding batch size of 8. The learning coefficient value of $10^{-4}$ is reduced by 9% after every five epochs until the $70^{th}$ epoch. This study uses a cross-validation approach applied to all experiments. The image patches are divided to train, test, and validate the models as 72%, 20%, and 8%, respectively.

4.1. **Evaluation Metrics.** Performance evaluation relies on IoU (intersection over union) and $F_1$ score metrics. IoU is formulated as:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{7}$$

where $TP$ refers to correctly identified pixels, $FP$ represents the incorrectly identified predictions, and $FN$ corresponds to false negatives. $F_1$ score is computed as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}, \tag{8}$$

where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$.

## 5. Results

For comparison purposes, the experiments include several other U-Net architectures, including U-Net, Nested U-Net, AttU-Net, R2AttU-Net, InceptionU-Net, and scAGAttU-Net [23–27]. Table 1 presents the crop semantic segmentation test results on the DSTL image set using the IoU and $F_1$ score for RGB and NDVI images. Additionally, the quantity of floating-point operations performed per second expressed in billions (GFLOPs) metric helps infer the execution time based on the computational complexity of the model.

This satellite image set contains multispectral images with lower spatial resolutions than the RGB counterparts, leading to some performance loss. When used with the dilation rates as $d_1 = 2, d_2 = 2$, the proposed Nested U-Net model outperforms all other architectures on the DSTL image set and demonstrates stable performance under different spectral schemes. With these rates, there is an improvement of about 0.6% in the IoU metric for RGB images and 1.1% for NDVI images compared to the traditional Nested U-Net design. The efficacy of the proposed model with the dilation rates of $d_1 = 1, d_2 = 2$ is almost equal to that of the original Nested U-net. The dilated Nested U-Net with rates of $d_1 = 1, d_2 = 3$ shows an IoU performance improvement of about 0.4% for RGB images compared to the conventional Nested U-Net model. On the other hand, relatively large dilation rates of $d_1 = 3, d_2 = 3$ result in IoU performance losses for both RGB and NDVI images. Similarly, the model performance degrades further with increasing the dilation rates to $d_1 = 3, d_2 = 4$ and $d_1 = 4, d_2 = 4$.

Experimental results show that the obtained optimal dilation rate scheme of $d_1 = 2, d_2 = 2$ outperforms others that employ smaller or larger dilation rates. Small dilation rates only capture the local features, which impacts the effectiveness of semantic segmentation in the original Nested U-Net model. Nevertheless, larger receptive fields with dilation rates as $d_1 = 3, d_2 = 3$, $d_1 = 3, d_2 = 4$ and $d_1 = 4, d_2 = 4$ also decrease the semantic segmentation performance since they capture too much irrelevant information. Examining the GFLOPs values concerning computational complexity reveals that varying dilation rates do not indirectly impact the model's execution time.

TABLE 1. Test results with DSTL set.

| 2*Architectures | GFLOPS | RGB Images | | NDVI Images | |
|---|---|---|---|---|---|
| | | IoU | $F_1$ | IoU | $F_1$ |
| U-Net | 190.07 | 0.894 ± 0.237 | 0.904 ± 0.240 | 0.857 ± 0.285 | 0.883 ± 0.263 |
| AttU-Net | 408.12 | 0.893 ± 0.243 | 0.902 ± 0.220 | 0.879 ± 0.275 | 0.898 ± 0.260 |
| R2AttU-Net | 943.42 | 0.857 ± 0.306 | 0.874 ± 0.296 | 0.804 ± 0.337 | 0.836 ± 0.314 |
| InceptionU-Net | 482.26 | 0.897 ± 0.236 | 0.919 ± 0.213 | 0.886 ± 0.252 | 0.913 ± 0.225 |
| scAGAttU-Net | 101.03 | 0.895 ± 0.262 | 0.910 ± 0.248 | 0.884 ± 0.279 | 0.906 ± 0.267 |
| UNetFormer | 17.95 | 0.880 ± 0.283 | 0.896 ± 0.268 | 0.865 ± 0.307 | 0.879 ± 0.297 |
| Nested U-Net | 849.3 | 0.897 ± 0.234 | 0.919 ± 0.209 | 0.880 ± 0.268 | 0.900 ± 0.251 |
| Nested U-Net ($d_1 = 1, d_2 = 2$ ) | 849.3 | 0.897 ± 0.253 | 0.914 ± 0.236 | 0.880 ± 0.263 | 0.902 ± 0.245 |
| Nested U-Net ($d_1 = 2, d_2 = 2$ ) | 849.3 | **0.903 ± 0.237** | **0.922 ± 0.216** | **0.891 ± 0.262** | **0.908 ± 0.240** |
| Nested U-Net ($d_1 = 1, d_2 = 3$ ) | 849.3 | 0.901 ± 0.246 | 0.918 ± 0.227 | 0.873 ± 0.290 | 0.889 ± 0.279 |
| Nested U-Net ($d_1 = 3, d_2 = 3$ ) | 849.3 | 0.892 ± 0.262 | 0.909 ± 0.245 | 0.861 ± 0.292 | 0.882 ± 0.276 |
| Nested U-Net ($d_1 = 3, d_2 = 4$ ) | 849.3 | 0.886 ± 0.260 | 0.901 ± 0.266 | 0.869 ± 0.299 | 0.884 ± 0.288 |
| Nested U-Net ($d_1 = 4, d_2 = 4$ ) | 849.3 | 0.889 ± 0.276 | 0.903 ± 0.264 | 0.867 ± 0.301 | 0.882 ± 0.291 |

TABLE 2. Test results with RIT-18 set.

| 2*Architectures | GFLOPS | RGB Images | | NDVI Images | |
|---|---|---|---|---|---|
| | | IoU | $F_1$ | IoU | $F_1$ |
| U-Net | 190.07 | 0.860 ± 0.285 | 0.887 ± 0.243 | 0.841 ± 0.306 | 0.878 ± 0.269 |
| AttU-Net | 408.12 | 0.881 ± 0.265 | 0.906 ± 0.243 | 0.883 ± 0.252 | 0.907 ± 0.227 |
| R2AttU-Net | 943.42 | 0.878 ± 0.243 | 0.904 ± 0.210 | 0.683 ± 0.464 | 0.710 ± 0.364 |
| InceptionU-Net | 482.26 | 0.864 ± 0.269 | 0.891 ± 0.239 | 0.873 ± 0.260 | 0.900 ± 0.236 |
| scAGAttU-Net | 101.03 | 0.873 ± 0.271 | 0.898 ± 0.251 | 0.860 ± 0.294 | 0.892 ± 0.262 |
| UNetFormer | 17.95 | 0.870 ± 0.260 | 0.899 ± 0.232 | 0.842 ± 0.296 | 0.871 ± 0.272 |
| Nested U-Net | 849.3 | 0.885 ± 0.254 | 0.910 ± 0.228 | 0.893 ± 0.242 | 0.918 ± 0.220 |
| Nested U-Net ($d_1 = 1, d_2 = 2$ ) | 849.3 | 0.885 ± 0.259 | 0.906 ± 0.243 | 0.889 ± 0.278 | 0.914 ± 0.260 |
| Nested U-Net ($d_1 = 2, d_2 = 2$ ) | 849.3 | **0.888 ± 0.256** | **0.908 ± 0.236** | **0.894 ± 0.264** | **0.918 ± 0.244** |
| Nested U-Net ($d_1 = 1, d_2 = 3$ ) | 849.3 | 0.884 ± 0.257 | 0.906 ± 0.236 | 0.886 ± 0.275 | 0.912 ± 0.254 |
| Nested U-Net ($d_1 = 3, d_2 = 3$ ) | 849.3 | 0.880 ± 0.261 | 0.903 ± 0.207 | 0.884 ± 0.253 | 0.907 ± 0.232 |
| Nested U-Net ($d_1 = 3, d_2 = 4$ ) | 849.3 | 0.879 ± 0.266 | 0.901 ± 0.247 | 0.825 ± 0.340 | 0.843 ± 0.329 |
| Nested U-Net ($d_1 = 4, d_2 = 4$ ) | 849.3 | 0.879 ± 0.261 | 0.902 ± 0.239 | 0.823 ± 0.332 | 0.845 ± 0.318 |

Table 2 shows the tree semantic segmentation test results for RGB and NDVI images on the RIT-18 set using IoU and $F_1$ score metrics. The Nested U-Net model with dilation rates $d_1 = 2, d_2 = 2$ in this image set demonstrates a similar trend as with the DSTL by outperforming all other models. The model with dilation rates of $d_1 = 1, d_2 = 2$ achieves almost the same performance as the original Nested U-Net for RGB images. However, the IoU performance of this case is about 0.889 for NDVI images, slightly lower than its original Nested U-Net counterpart. When the dilation rates become large with $d_1 = 3, d_2 = 3$, the Nested U-Net suffers from a performance degradation of about 0.9% for NDVI images compared to the original model. With a similar trend, IoU values for RGB images at $d_1 = 3, d_2 = 4$ and $d_1 = 4, d_2 = 4$ dilation rates decrease by approximately 0.9% compared to the one at the optimal rate of $d_1 = 2, d_2 = 2$.

As the dilation rates continue to increase, the false alarms introduced by highlighting the irrelevant regions become a significant drawback [28–30]. Furthermore, limiting feature extraction to the central image pixels with rates close to the feature

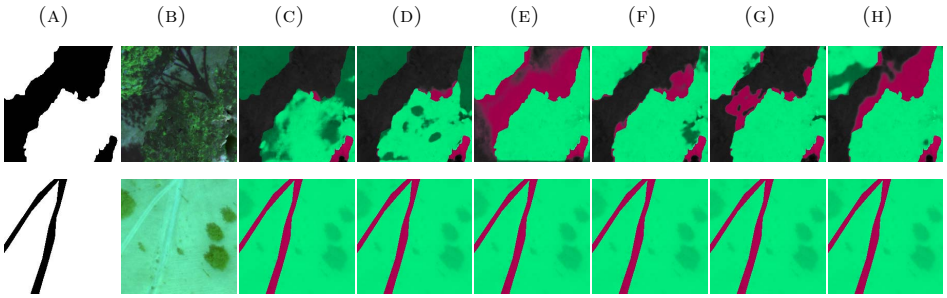



FIGURE 10. Semantic segmentation predictions. Light greens are hits, dark green are misses, and reds are false alarms. First row illustrates tree class predictions (NDVI) and second row shows crop class predictions (RGB). (A) Ground-truth mask. (B) Original image. (C) U-Net. (D) AttU-Net. (E) R2AttU-Net. (F) InceptionU-Net. (G) scAGAttU-Net. (H) UNetFormer.

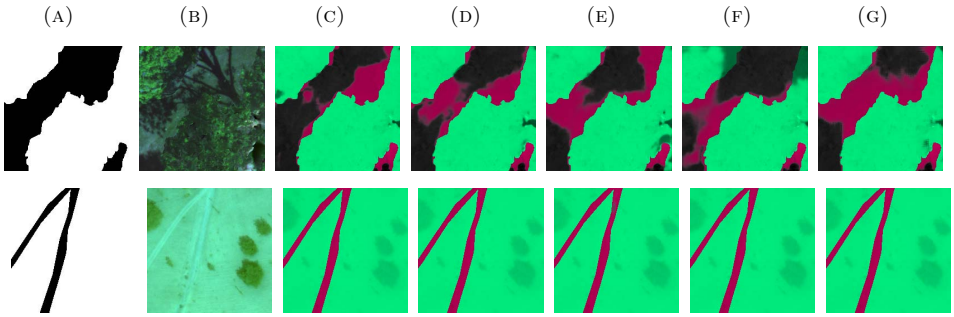



FIGURE 11. Semantic segmentation predictions. Light greens are hits, dark green are misses, and reds are false alarms. First row illustrates tree class predictions (NDVI) and second row shows crop class predictions (RGB). (A) Ground-truth mask. (B) Original image. (C) NestedU-Net. (D) NestedU-Net ($d_1 = 1, d_2 = 2$ ). (E) NestedU-Net ($d_1 = 2, d_2 = 2$ ). (F) NestedU-Net ($d_1 = 1, d_2 = 3$ ). (G) NestedU-Net ($d_1 = 3, d_2 = 3$ ).

map sizes reduces the valid convolution regions and can thus degrade the performance. On the other hand, standard convolution (with a rate of 1) tends to focus on local details, potentially neglecting global contextual information. Therefore, choosing the dilation rates to attain a good trade-off between small and large values can elevate the accuracy of the original model.

Experiments demonstrate that the dilation rate scheme of $d_1 = 2, d_2 = 2$ improves the performance when used with the Nested U-Net model. However, the resolution of the feature map decreases with the $d_1 = 3, d_2 = 3$ scheme, which makes it harder to capture details. Figures 10 and 11 show some example predictions on the DSTL and RIT-18 image sets. The first column visualizes the results

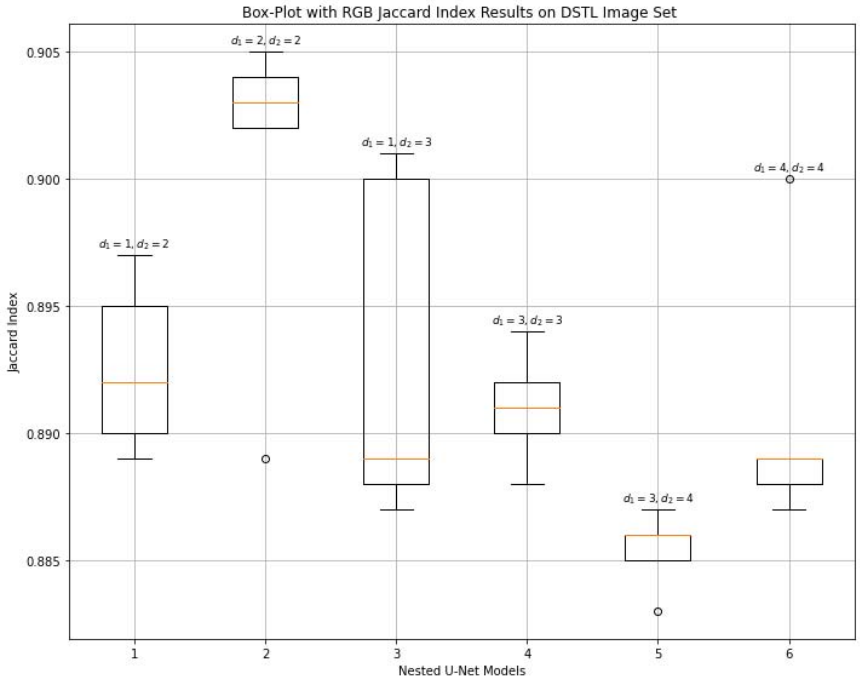



FIGURE 12. Box-plot comparison of IoU across different dilation rates in Nested U-Net model using RGB images from the DSTL image set.

for the NDVI images of the tree class, and the second column presents the results for the RGB images of the crop class. Figure 10 includes the predictions of U-Net models used for comparison, while Figure 11 provides those of the Nested U-Net model in different dilation rate schemes. Despite the promising experimental results, the best-performing Nested U-Net model with dilation rates $d_1 = 2, d_2 = 2$ still has limitations in avoiding false alarms, especially in complex detailed regions (Figure 11). Therefore, future work will address this specific issue concerning false alarms by incorporating some attention mechanisms. Additionally, the box-plot statistical analysis is beneficial to visualize the distribution of IoU performances with different dilation rates. Figure 12 presents the IoU values for different Nested U-Net models, each configured with various dilation rates $d_1$ and $d_2$. These are the DSTL image set RGB results obtained by applying 5-fold cross-validation for each dilation rate scheme. The box plot analysis further confirms that the dilation rate scheme of $d_1 = 2, d_2 = 2$ provides the best and most consistent Nested U-Net performance.

## 6. Conclusion

This study proposes an effective Nested U-Net architecture with dilated convolutional layers to capture detailed multi-scale objects within remote sensing data. Dilated convolution is a promising approach since it can enhance the resolution of receptive fields without increasing computational load. By selecting various dilation rates in convolution blocks, diverse dilation schemes can be generated, thereby analyzing the potential benefits of using dilated convolutions in the Nested U-Net model. According to the experimental results using two multispectral remote sensing image sets, integrating the optimal dilation rate scheme of $d_1 = 2, d_2 = 2$ can achieve higher semantic segmentation performance than the original Nested U-Net model. Furthermore, smaller dilation rates tend to neglect the global feature information, while larger ones contribute to higher false alarm rates. These results demonstrate that suitable dilation rates can enhance the effectiveness of the Nested U-Net model, which is extremely useful for capturing multi-scale details in remote sensing images.

**Declaration of Competing Interests** The author declares no known competing interests.

## References

[1] Piao, S., Liu, J., Accuracy improvement of UNet based on dilated convolution, *J. Phys. Conf. Ser.*, 1345 (5) (2019), 052066, https://doi.org/10.1088/1742-6596/1345/5/052066.

[2] Ma, B., Chang, C. Y., Semantic segmentation of high-resolution remote sensing images using multiscale skip connection network, *IEEE Sens. J.*, 22 (4) (2021), 3745-3755, https://doi.org/10.1109/JSEN.2021.3139629.

[3] Ding, L., Zhang, J., Bruzzone, L., Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture, *IEEE Trans. Geosci. Remote Sens.*, 58 (8) (2020), 5367-5376, https://doi.org/10.1109/TGRS.2020.2964675.

[4] Li, X., Lei, L., Kuang, G., Multilevel adaptive-scale context aggregating network for semantic segmentation in high-resolution remote sensing images, *IEEE Geosci. Remote Sens. Lett.*, 19 (2021), 1-5, https://doi.org/10.1109/LGRS.2021.3091284.

[5] Du, S., Du, S., Liu, B., Zhang, X., Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach, *Remote Sens. Environ.*, 261 (2021), 112480, https://doi.org/10.1016/j.rse.2021.112480.

[6] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J., Unet++: A nested U-Net architecture for medical image segmentation, *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, (2018), 3-11, https://doi.org/10.1007/978-3-030-00889-5_1.

[7] Zhong, Y., Shi, Z., Zhang, Y., Zhang, Y., Li, H., CSAN-UNet: Channel spatial attention nested UNet for infrared small target detection, *Remote Sens.*, 16 (11) (2024), 1894, https://doi.org/10.3390/rs16111894.

[8] Jiang, J., Liu, L., Cui, Y., Zhao, Y., A Nested UNet based on multi-scale feature extraction for mixed Gaussian-impulse removal, *Appl. Sci.*, 13 (17) (2023), 9520, https://doi.org/10.3390/app13179520.

[9] Kushnure, D. T., Tyagi, S., Talbar, S. N., LiM-Net: Lightweight multi-level multiscale network with deep residual learning for automatic liver segmentation in CT images, *Biomed. Signal Process. Control*, 80 (2023), 104305, https://doi.org/10.1016/j.bspc.2022.104305.

[10] Yang, B., Liu, Z., Duan, G., Tan, J., Residual shape adaptive dense-nested Unet: Redesign the long lateral skip connections for metal surface tiny defect inspection, *Pattern Recognit.*, 147 (2024), 110073, https://doi.org/10.1016/j.patcog.2023.110073.

[11] Wu, Z., Tang, Y., Hong, B., Liang, B., Liu, Y., Enhanced precision in dam crack width measurement: Leveraging advanced lightweight network identification for pixel-level accuracy, *Int. J. Intell. Syst.*, 2023 (2023), 9940881, https://doi.org/10.1155/2023/9940881.

[12] Liu, Y., Liu, J., Ning, X., Li, J., MS-CNN: multiscale recognition of building rooftops from high spatial resolution remote sensing imagery, *Int. J. Remote Sens.*, 43 (1) (2022), 270-298, https://doi.org/10.1080/01431161.2021.2018146.

[13] Zhang, T., Yang, Z., Xu, Z., Li, J., Wheat yellow rust severity detection by efficient DF-UNet and UAV multispectral imagery, *IEEE Sens. J.*, 22 (9) (2022), 9057-9068, https://doi.org/10.1109/JSEN.2022.3156097.

[14] Zhou, G., Yu, J., Zhou, S., LSCB: a lightweight feature extraction block for SAR automatic target recognition and detection, *Int. J. Remote Sens.*, 44 (8) (2023), 2548-2572, https://doi.org/10.1080/01431161.2023.2203342.

[15] Ren, K., Chen, X., Wang, Z., Liang, X., Chen, Z., Miao, X., HAM-transformer: A hybrid adaptive multi-scaled transformer net for remote sensing in complex scenes, *Remote Sens.*, 15 (19) (2023), 4817, https://doi.org/10.3390/rs15194817.

[16] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.*, 40 (4) (2017), 834-848, https://doi.org/10.1109/TPAMI.2017.2699184.

[17] He, H., Zhang, C., Chen, J., Geng, R., Chen, L., Liang, Y., Xu, Y., A hybrid-attention nested UNet for nuclear segmentation in histopathological images, *Front. Mol. Biosci.*, 8 (2021), 614174, https://doi.org/10.3389/fmolb.2021.614174.

[18] Agnes, S. A., Anitha, J., Solomon, A. A., Two-stage lung nodule detection framework using enhanced UNet and convolutional LSTM networks in CT images, *Comput. Biol. Med.*, 149 (2022), 106059, https://doi.org/10.1016/j.compbiomed.2022.106059.

[19] Yang, K., Yi, J., Chen, A., Liu, J., Chen, W., ConDinet++: Full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images, *IEEE Geosci. Remote Sens. Lett.*, 19 (2021), 1-5, https://doi.org/10.1109/LGRS.2021.3093101.

[20] Safarov, S., Whangbo, T. K., A-DenseUNet: Adaptive densely connected UNet for polyp segmentation in colonoscopy images with atrous convolution, *Sensors*, 21 (4) (2021), 1441, https://doi.org/10.3390/s21041441.

[21] Zhao, H., Zhang, H., Zheng, X., A multiscale attention-guided UNet++ with edge constraint for building extraction from high spatial resolution imagery, *Appl. Sci.*, 12 (12) (2022), 5960, https://doi.org/10.3390/app12125960.

[22] Ulku, I., ResLMFFNet: a real-time semantic segmentation network for precision agriculture, *J. Real-Time Image Process.*, 21 (4) (2024), 101, https://doi.org/10.1007/s11554-024-01474-0.

[23] Ronneberger, O., Fischer, P., Brox, T., U-net: Convolutional networks for biomedical image segmentation, *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, (2015), 234-241, https://doi.org/10.1007/978-3-319-24574-4_28.

[24] Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., Asari, V. K., Recurrent residual U-Net for medical image segmentation, *J. Med. Imaging*, 6 (1) (2019), 014006, https://doi.org/10.1117/1.JMI.6.1.014006.

[25] Delibasoglu, I., Cetin, M., Improved U-Nets with inception blocks for building detection, *J. Appl. Remote Sens.*, 14 (4) (2020), 044512, https://doi.org/10.1117/1.JRS.14.044512.

[26] Khanh, T. L. B., Dao, D. P., Ho, N. H., Yang, H. J., Baek, E. T., Lee, G., Yoo, S. B., Enhancing U-Net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging, *Appl. Sci.*, 10 (17) (2020), 5729, https://doi.org/10.3390/app10175729.

[27] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P. M., UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, *ISPRS J. Photogramm. Remote Sens.*, 190 (2022), 196-214, https://doi.org/10.1016/j.isprsjprs.2022.06.008.

[28] Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T. S., Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, (2018), 7268-7277, https://doi.org/10.1109/CVPR.2018.00759.

[29] Yu, W., Zhuo, L., Li, J., GCFormer: Global Context-aware Transformer for Remote Sensing Image Change Detection, *IEEE Trans. Geosci. Remote Sens.*, 62 (2024), 1-12, https://doi.org/10.1109/TGRS.2024.3381738.

[30] Pang, Z., Hu, R., Zhu, W., Zhu, R., Liao, Y., Han, X., A Building Extraction Method for High-Resolution Remote Sensing Images with Multiple Attentions and Parallel Encoders Combining Enhanced Spectral Information, *Sensors*, 24 (3) (2024), 1006, https://doi.org/10.3390/s24031006.