Extended U-Net for Satellite Image Semantic Segmentation

Jin Won Jung
School of Electronic Engineering
Soongsil University
Seoul 06978, Korea
Email: jinwonj@soongsil.ac.kr

Yoan Shin[†]
School of Electronic Engineering
Soongsil University
Seoul 06978, Korea
Email: yashin@soongsil.ac.kr

***Abstract*— In this paper, we propose an extended model of encoder-decoder structure using U-Net to improve accuracy in satellite image semantic segmentation. The U-Net is a deep learning-based semantic segmentation scheme, and research is being conducted in various applications such as road sign detection, medical image analysis, and tumor detection. The conventional U-Net suffers from losses during feature compression-expansion due to shallow structures of encoder-decoder. This creates a problem of reduced segmentation accuracy. In order to address this issue, the proposed scheme exploits an extended structure using concatenate upsampling and residual learning, which remedies the loss of information and improves the accuracy of semantic segmentation. In the experiments, segmentation was performed on various satellite images, and it was shown that the proposed U-Net was superior to the conventional counterpart.**

Keywords*— Image processing, Semantic segmentation, Deep learning, U-Net, Concatenate upsampling, Residual learning

I. Introduction

As one of main research areas of image processing, semantic segmentation which is also called pixel-level classification is the task of clustering parts of images belonging to the same object class together [1]. Recently, deep learning techniques have been widely applied in semantic segmentation with noticeable adoption of the convolutional neural networks (CNNs). The U-Net is a representative CNN model for this purpose [2]. This model delivers the encoder's information to the decoder through skip connections, which leads to minimizing information loss. This is an excellent feature for obtaining fine-grained details from complex backgrounds and segmenting hidden objects. However, in the case of satellite image segmentation, even a small error can raise a big problem, and there is an issue that the segmentation of the object is wrong or fails due to lack of feature map information extracted from the encoder.

The extended U-Net proposed in this paper exploits the pre-trained model of the InceptionResNetV2 and the residual learning to efficiently use the feature map information [3][4]. Moreover, we remedy the loss of spatial information contained in the feature map by concatenating the feature map information with a part of the decoder, while reducing the feature loss of the encoder.

II. Related Works

A. U-Net

As one of popular deep learning models for semantic segmentation, the U-Net is a U-shaped network designed for segmentation and follows the encoder-decoder structure [2]. Figure 1 shows a general structure of the U-Net.

Each step in the convolutional U-Net consists of two convolution encoders and one pooling, and one transposed convolution and two convolution decoders. The encoder extracts the feature map information of the image by gradually reducing the size of the image, while increasing the depth of the image. The decoder converts the information back to the original pixel position by upsampling the image. At this time, the size of the image is increased, while reducing the depth. Furthermore, a skip connection is used to connect the feature map of the encoder to the output of the transposed convolution layer to improve the upsampling of the image. This U-Net has significantly less convolutions compared to other CNN models such as the InceptionResNetV2, and this leads to a problem of poor segmentation accuracy due to poor feature map information extraction and learning ability.

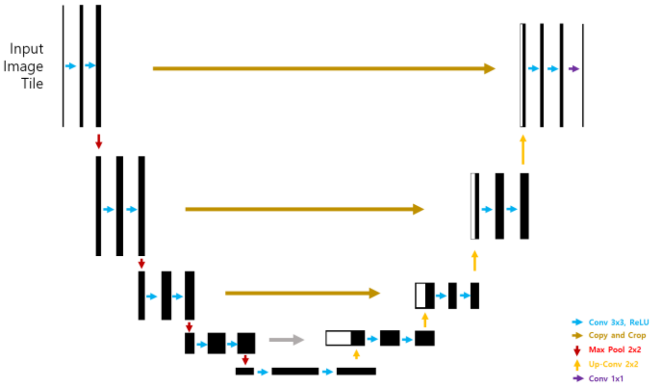

Fig. 1. General U-Net architecture [2]

III. Proposed Scheme

The proposed U-Net exploits the InceptionResNet V2 block as the respective encoder, as shown in Fig. 2. The skip connections connecting the encoder's feature map to the decoder are used to improve the upsampling. The skip connections are made in the concatenate block, and the feature maps contain original spatial information lost during compression of the encoder and help the decoder to construct more precise segmentation results. After that, it passes through the decoder block and finally outputs through the convolution layer and the softmax layer.

[†]Corresponding author.

This research was supported by the MSIT, Korea, under the ITRC support program (IITP-2023-2018-0-01424) supervised by the IITP.

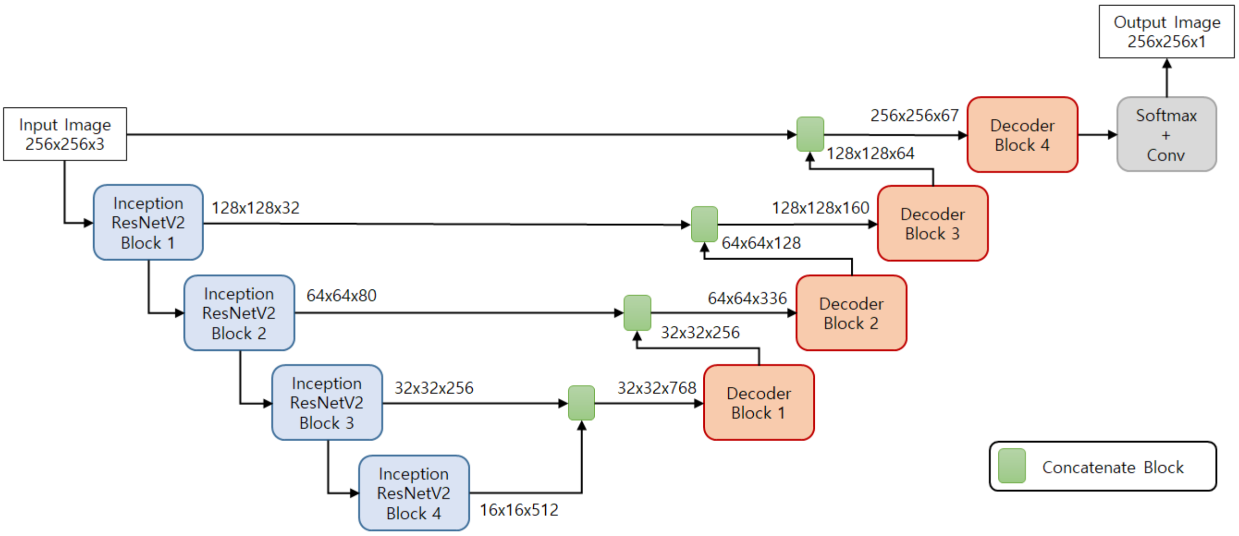

Fig. 2. Proposed U-Net architecture

A. InceptionResNetV2 Block and Concatenate Block

Figure 3 shows the structures of the InceptionResNetV2 block and the concatenate block. The InceptionResNetV2 which acts as an encoder, delivers the feature map information to the concatenate block through zero padding.

The concatenate block is a 2x2 transposed convolution with stride 2, which doubles the feature map. We then concatenate the intermediate feature map of the encoder with the same resolution as the extended feature map. Concatenation is the process of connecting two feature maps and is represented by (1).

$$\text{Concatenate}(T, I) = [T, I]_{w,h,c}\,. \tag{1}$$

Here, T is the feature map extended to the transposed convolution, and I is the intermediate feature map of the InceptionResNetV2 block in the encoder. In addition, w, h, and c represent width, height, and channel, respectively.

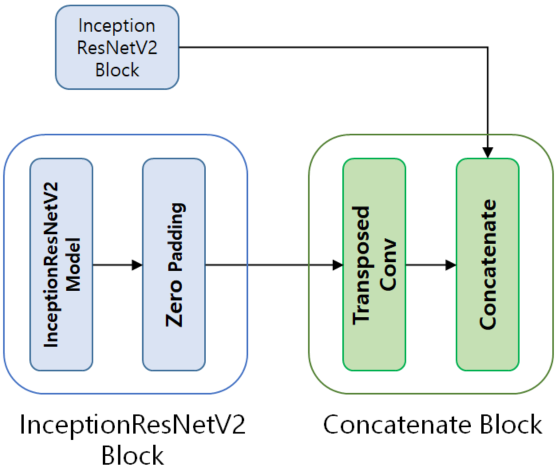

Fig. 3. InceptionResNetV2 block and concatenate block architecture

B. Decoder Block

Figure 4 shows the structure of the decoder block. Here, x is the feature map entering the decoder block, and $A(x)$ is the structure in which the batch normalization (BN) and the rectified linear unit (ReLU) are connected after the convolution layer. This structure adds a concatenated feature map to the last part of the decoder block by applying the residual learning [5][6].

The residual learning process is expressed in (2). This process minimizes the loss of features in the process of feature extraction and compression of continuous convolution, as well as the vanishing gradient problem which is likely to occur as the network deepens.

$$F(x) = A(x) + x. \tag{2}$$

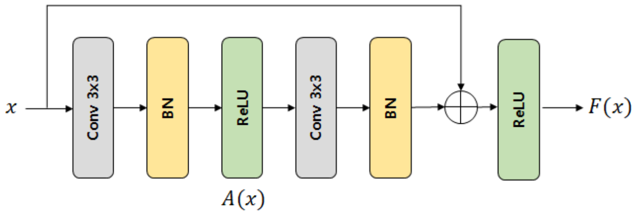

Fig. 4. Decoder block architecture

IV. Experimental Results

A. Training Strategy

We performed the experiments to compare segmentation performance of the conventional U-Net and the proposed one. The dataset consisted of Dubai's color satellite images obtained from MBRSC satellites [7]. This dataset was used for preprocessing 1,136 images with the size of 256x256 sizes. The batch size was set to 8, the epoch was set to 100, the initial learning rate was 0.001, and the weight decay was set to 0.00001. The Adam optimizer was used and data augmentation techniques were applied. Performance metrics include the accuracy, the mean intersection over union (MIoU) and the Dice coefficient. The MIoU is the average of IoU calculation for each class as follows [8].

$$\text{IoU} = \frac{|X \cap Y|}{|X \cup Y|}\,. \tag{3}$$

Moreover, the Dice coefficient is defined as [9]

$$\text{Dice} = 2 \cdot \frac{|X \cap Y|}{(|X| + |Y|)}\,. \tag{4}$$

In the above equations, X is the true pixel and Y is the prediction pixel. As two regions of X and Y become equal,

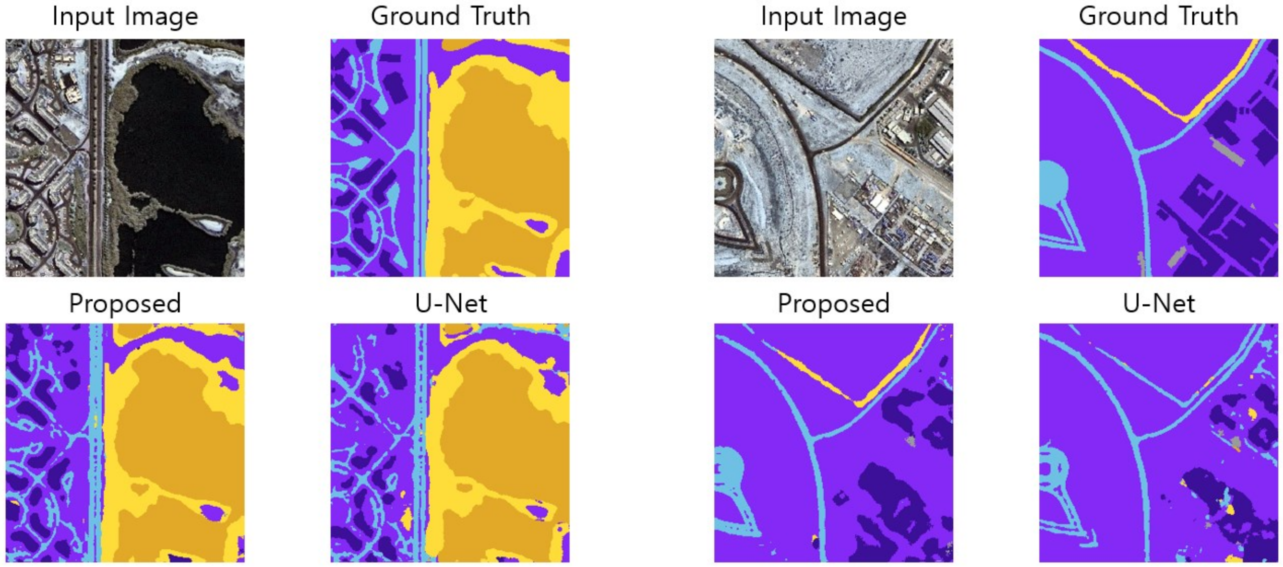

Fig. 5. Segmentation results on sample satellite images.

these performance metrics become closer to 1; Otherwise, they become closer to 0.

B. *Results*

Table 1 summarizes the experimental results. The Dice coefficient of the proposed scheme showed much higher accuracy of 84.5% compared to the conventional U-Net. In addition, the MIoU of the proposed scheme is 73.8%; an improvement by about 13%p over the conventional U-Net. Figure 5 illustrates sample image results after the semantic segmentation. We observe that the proposed scheme can well identify smaller objects than the U-Net.

TABLE 1. EXPERIMENTAL RESULTS

Scheme	Metrics		
	Dice Coefficient	MIoU	Accuracy
Proposed	84.5%	73.8%	88.3%
U-Net	80.2%	60.3%	84.8%

V. CONCLUSIONS

In this paper, we proposed a method using transfer learning, concatenate and residual learning to improve the U-Net-based semantic segmentation for satellite images. The conventional U-Net has a problem of feature loss due to its shallow structure and learning is difficult. To tackle these issues, we proposed an encoder using a transfer learning model on the conventional U-Net and a decoder utilizing concatenate and residual learning. Feature loss was reduced and feature map information was learned efficiently by the proposed scheme. The experimental results showed that the proposed scheme significantly outperformed the conventional U-Net in terms of various segmentation performance metrics.

REFERENCES

[1] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intell. Rev.*, vol. 52, pp. 1089-1106, Aug. 2019.

[2] O Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Comp. Sci.*, vol. 9351, Issue Cvd, pp. 234–241, Nov. 2015.

[3] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43-76, July 2020.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, InceptionResNet and the impact of residual connections on learning," *Proc. AAAI 2017*, pp. 4278-4284, San Francisco, USA, Feb. 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE CVPR 2016*, pp. 770-778, Las Vegas, USA, June 2016.

[6] S. Shin, S. Lee, and H. Han, "A Study on residual U-Net for semantic segmentation based on deep learning," *Jour.Digital Conv.*, vol. 19, no. 6, pp. 251–258, June 2021.

[7] https://humansintheloop.org/resources/datasets/ semantic-egmentation-dataset-2/

[8] H. Rez., N. Tsoi, J. Gwak,, A Sad., I. Reid, and S. Sav, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proc. IEEE/CVF CVRR 2019*, pp. 658-666, Long Beach, USA, Feb. 2019.

[9] L. R. Dice, "Measures of the amount of ecologic association between species," *Jour. Ecological Soc. Amer.*, vol. 26, no. 3, pp. 297-302, July 1945.