Skip Connection Model for SAR Ship 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 a new encoder-decoder model to enhance the performance of semantic segmentation in satellite synthetic aperture radar (SAR) images. Traditional semantic segmentation suffers from significant information loss during feature compression and expansion, primarily because of its shallow structure. In particular, SAR ship images are noisy and difficult to interpret due to the class imbalance problem. This issue reduces segmentation accuracy and hinders proper object distinction. The proposed model addresses these problems by incorporating multiple layers of features to create a connected structure with the decoder, thereby enhancing the preservation of feature information. In addition, the Gaussian filter is used to address the issue of texture bias problem and improve the accuracy of semantic segmentation. This connection method reduces information loss in the feature learning process and addresses the issue of texture bias problem. Compared to the existing techniques including U-Net, feature pyramid networks, and LinkNet, the experimental results of semantic segmentation of SAR ships demonstrate that the proposed method achieves an intersection over union of 78.2%, which is higher than U-Net's 67.8%. Additionally, the proposed model also achieves the highest Dice coefficient.**

***Index Terms*—image processing, semantic segmentation, synthetic aperture radar, deep learning, gaussian filter, texture bias, residual learning**

I. Introduction

Synthetic aperture radar (SAR) is an advanced microwave sensor that is widely used for ocean monitoring. Unlike optical imagery, SAR can observe climate changes such as clouds and fog, and it can acquire images regardless of the time. However, SAR ship images are noisy and only have two classes: background and ship. As a result, there is a class imbalance problem, making it challenging to apply semantic segmentation. In addition, SAR ship images suffer from a texture bias problem. This issue arises due to the limited dataset available for SAR ship images, as well as the insufficient feature information extracted by the encoder. Consequently, this leads to inaccurate segmentation of objects. In this paper, we propose a method to address the issue of texture bias by connecting the feature information of each layer using skip connection and reduce the noise of each layer using a Gaussian filter. In addition, we aim to address the class imbalance problem caused by the high background ratio through this process. Finally, we check whether the segmentation performance of the SAR ship image is improved.

†Corresponding author.

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

II. Related Works

A. Semantic Segmentation

Semantic segmentation is the task of classifying pixels into specific categories of objects or areas, and it is being studied in various fields such as tumor detection, landmark classification, and road sign detection. Semantic segmentation is primarily studied using convolutional neural network (CNN) deep learning models, with the U-Net [1] being the representative method.

B. U-Net

As one of the popular deep learning models for semantic segmentation, the U-Net is a U-shaped network designed for segmentation and follows the encoder-decoder structure. Each step in the convolutional U-Net consists of two convolution encoders and one pooling layer, as well as one transposed convolution and two convolution decoders. The encoder extracts the feature map information of the image by progressively reducing the image size and increasing its depth. The decoder upsamples the image to convert the information back to its original pixel position. At this time, the image size 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 in order to enhance the upsampling of the image. U-Net has significantly fewer convolutions compared to other CNN models, which leads to a problem of poor segmentation accuracy. This is because U-Net has limited ability to extract and learn feature map information.

C. Texture Bias

It was generally thought that CNN use shape information to learn, similar to humans. However, recent studies have shown that CNN rely heavily on texture for object recognition and are texture bias [2], [3]. The texture bias problem is an issue that gives more importance to texture rather than shape in image

recognition. It is characterized by the fact that the texture bias problem is worse when the amount of training data is small. Especially in SAR ship images, there is a texture bias problem due to the limited amount of data and the presence of noisy or distorted images. This significantly hampers the performance of semantic segmentation.

III. Proposed Scheme

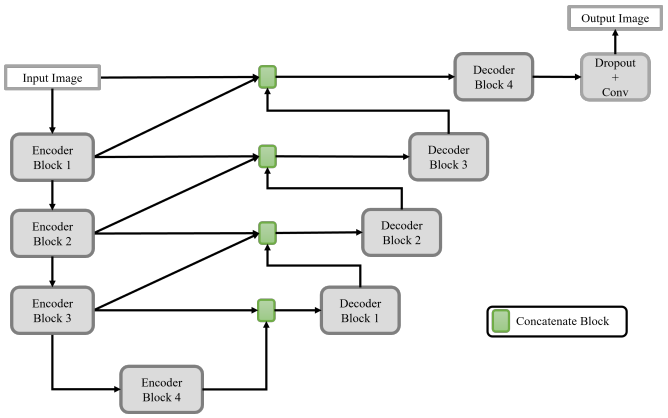

Fig. 1. Proposed skip connection model architecture

The proposed model utilizes encoder block as the encoder, as depicted in Figure 1. Additionally, a skip connection is employed to connect the feature maps of the two encoders to the decoder, thereby enhancing upsampling. The skip connection is established in concatenate block, and it allows the feature map to retain the spatial information that is lost during compression in the encoder. This information then assists the decoder in generating a more accurate segmentation result. It then passes through the decoder block with a Gaussian filter to denoise the feature map and utilize residual learning. Finally, the output is passed through the dropout and convolution layer.

A. Encoder Block and Concatenate Block

Figure 2 shows the structure of the encoder block and concatenate block. The InceptionResNetV2 model, acting as an encoder, passes the feature map information through zero padding to the concatenate block [4]. The concatenate block expands the feature map by 2 using a transposed convolution with a stride 2. It then concatenates the expanded feature map with an intermediate feature map from the encoder that has the same resolution as the expanded feature map, as well as the values from the decoder block. This operation performs the process of concatenating the feature maps of different blocks.

B. Decoder Block

Figure 3 shows the structure of the decoder block. Here, x represents the feature map entering the decoder block, while σ denotes the standard deviation of the Gaussian filter. $A(x)$ is the structure of the Gaussian filter, batch normalization (BN), rectified linear unit (ReLU), and convolution. This structure applies residual learning to add the concatenated feature map to the last part of the decoder block [5]. The standard deviation

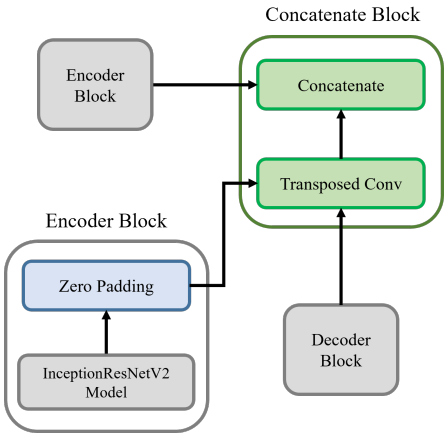

Fig. 2. Encoder block and concatenate block architectures

value of the Gaussian filter is different for each decoder block, which serves to reduce the noise of the feature map by decreasing its scale. Equation (1) shows the decoder's residual learning process. This process adds the feature map x to the feature map that has been denoised with a Gaussian filter. This helps minimize feature loss and texture bias during feature expansion.

$$F(x) = A(x) + x. \tag{1}$$

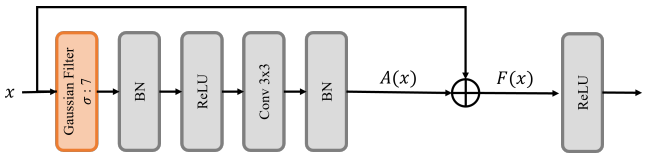

Fig. 3. Decoder block architecture

IV. Experimental Results

Experiments were conducted to compare the segmentation performance of the proposed model with the existing U-Net, feature pyramid networks (FPN) [6], and LinkNet [7]. The dataset is the high resolution SAR images dataset (HRSID) [8] which consists of high-resolution SAR images. This dataset contains 5604 high-resolution SAR images and 16951 ship instances. We set the batch size to 4, the epoch to 100, the initial learning rate to 0.001, and the weight decay to 0.00001. The Adam optimizer was used and the data augmentation technique was applied. For performance metrics, we used accuracy, Intersection over Union (IoU), and Dice coefficient [9], [10].

$$IoU = \frac{|X \cap Y|}{|X \cup Y|}. \tag{2}$$

$$Dice = 2\frac{|X \cap Y|}{(|X| + |Y|)}. \tag{3}$$

In the above equations, X represents the true pixel and Y represents the predicted pixel. Both performance metrics

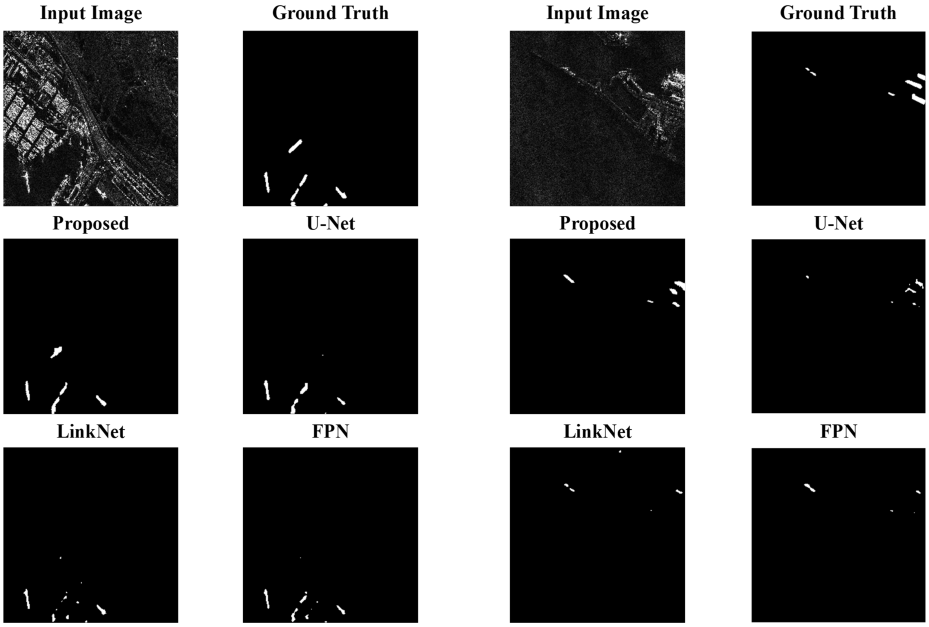

Fig. 4. Sample images of the experimental results

approach 1 as the two areas of X and Y become more equal, and approach 0 as they diverge.

A. Results

The experimental results compare the values of IoU and Dice coefficient for ships using U-Net, FPN, LinkNet, and the proposed model. As shown in Table 1, the Dice coefficient of the proposed method is 94.3%, which is the highest value. The proposed method achieves an IoU of 78.2% for ships, which is an 11%p improvement over U-Net. This demonstrates that the proposed model can identify objects more effectively than other models.

TABLE I
EXPERIMENTAL RESULTS.

Backbone	Model	Metrics	
		IoU (ship)	Dice Coefficient
Inception ResNetV2	U-Net	67.8%	90.8%
	FPN	60.1%	88.0%
	LinkNet	61.4%	88.3%
	Proposed	**78.2%**	**94.3%**

V. CONCLUSION

In this paper, we propose a method that utilizes a Gaussian filter and skip connection to enhance the performance of semantic segmentation in SAR ship images. This method improves texture bias by reducing noise and extracting information from multiple layers of encoders, which are then connected. In addition, residual learning is applied to the decoder with a Gaussian filter to reduce the loss of feature information. Experiments show that the proposed method can achieve significantly better performance than the existing models.

REFERENCES

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

[2] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," *Proc. Int. Conf. Learn. Representations*, Los Angeles, USA, May 2019.

[3] K. Hermann, T. Chen, and S. Kornblith, "The origins and prevalence of texture bias in convolutional neural networks," *Proc. Conf. Neural Info. Process. Syst.*, vol. 33, Virtual Conference, Apr. 2020, pp. 19000–19015.

[4] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inceptionv4, inception-ResNet and the impact of residual connections on learning," *Proc. AAAI Conf. Artif. Intell.*, San Francisco, USA, Feb., 2017, Art. no. 12.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. Conf. Comput. Vision & Pattern Recog.*, Las Vegas, USA, July 2016, pp. 770–778.

[6] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," *Proc. Conf. Comput. Vision & Pattern Recog.*, Honolulu, USA, July 2017, pp. 936-944.

[7] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," *Proc. IEEE Int. Conf. Visual Commun. Image Process.*, St. Petersburg, USA, Dec. 2017, pp. 1-4.

[8] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, June 2020.

[9] 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 Conf. Comput. Vision & Pattern Recog.*, pp. 658-666, Long Beach, USA, Feb. 2019.

[10] 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.