DS-UNet: Dual-Stream U-Net for Oil Spill Detection of SAR Image

Chunshan Li, Mingzhi Wang, Xiaofei Yang, and Dianhui Chu

***Abstract*— The oil spill detection of synthetic aperture radar (SAR) images has had great success. Existing deep-learning-based methods make predictions mainly based on the U-Net structure and Transformer, which fail to blend the local and global information generated by other different feature maps. In this letter, we proposed a dual-stream Unet (DS-Unet) for oil spill detection of SAR images. In particular, the proposed DS-Unet consists of two modules: an edge feature extraction module for extracting the local information and an interscale alignment module for capturing the global information. Moreover, an edge extraction branch is applied to handle the speckle noise of SAR images. Extensive experiments on two real-world datasets (Palsar and Sentinel) have shown that the proposed DS-Unet outperforms many existing state-of-the-art methods.**

***Index Terms*— Oil spill, semantic segmentation, Transformer, U-Net.**

I. Introduction

THE ocean is a part of the whole world, which plays an important role in global climate and environmental change. Meanwhile, the oil spill is one of the most widespread and harmful marine pollution, which would cause serious damage to the marine ecosystem. Therefore, it is very important to monitor oil spills timely and accurately. An oil spill detection task could be defined as a pixel-wise classification task, and the existing approaches for oil spill detection would divide the oil spill pixels from other pixels, which can be classified into three categories. The first category is traditional unsupervised segmentation methods. Otsu [1] proposed a threshold-based synthetic aperture radar (SAR) image segmentation method "OTSU" that can divide the image by setting a grayscale threshold and then cut it into two regions with different characteristics. Duan et al. [2] developed an unsupervised oil spill detection method based on isolation forest. However, these methods only consider the pixel-level features of the image typically and ignore the spatial and semantic information of the SAR image itself. The second category is convolutional neural network (CNN)-based methods. As an efficient visual feature extraction method, the CNN has been driving progress in the field of image processing as well as the field of image semantic segmentation. In the early times, Long et al. [3] proposed the FCN model, which is the first end-to-end, pixels-to-pixels fully convolutional network for image semantic segmentation. Then, the U-Net [4] structure extends CNNs and has become the default structure for most image segmentation. It consists of a symmetric encoder–decoder network with skip connections to enhance detail retention. Although CNN-based methods perform excellently in extracting local features, they exhibit limitations for modeling explicit long-range relations. Recently, researchers have gradually noticed the Transformer structure to capture the global connection between different tokens (the third category). TransUnet [5] is the first image segmentation framework of the "CNN + Transformer," which establishes self-attention mechanisms from the perspective of sequence-to-sequence prediction. It leveraged both detailed spatial information from CNN features and the global context encoded by Transformers. Duan et al. [6] proposed a self-supervised learning method to learn the deep neural network from unlabeled hyperspectral data for oil spill detection.

Although TransUnet [5] aggregates local features into the deepest feature map produced by convolution operations and captures long-distance dependencies through attention operations, it may ignore the local and global information generated by other different feature maps, which may lead to loss of performance. To integrate the information from different feature maps, Mao et al. [7] proposed a dual-stream network that can dynamically propagate information between local and global scales in both directions. It only requires one multihead attention operation and three cross-attention mechanism operations to fuse two different dimensional feature maps and find the connection between tokens in different dimensions. However, DS-Net failed to consider the local high-level features and global dependencies of the images and cannot be applied to the SAR image segmentation field. Besides, SAR images are prone to generating speckle noise and the above methods ignore the task of extracting image edge features. Generally speaking, reducing noise and improving the positioning accuracy of edge features are two parallel and unified tasks.

In this letter, we proposed dual-stream-Unet (DS-Unet) for the oil spill detection of SAR images, which is a novel framework that combines edge, local, and global features with different dimensional feature maps. The DS-Unet employs an "U"-type architecture to combine deep and shallow features of images. For the encoder section of the DS-Unet, the model

Manuscript received 12 July 2023; revised 15 October 2023; accepted 3 November 2023. Date of publication 8 November 2023; date of current version 17 November 2023. This work was supported by the Major Scientific and Technological Innovation Project of Shandong Province of China under Grant 2021ZLGX05 and Grant 2020CXGC010705. *(Corresponding authors: Xiaofei Yang; Mingzhi Wang.)*

Chunshan Li, Mingzhi Wang, and Dianhui Chu are with the Department of Computer Science, Harbin Institute of Technology, Weihai 264209, China (e-mail: wangmingzhi618@gmail.com).

Xiaofei Yang is with the School of Electronic and Communication Engineering, Guangzhou University, Guangzhou 510006, China (e-mail: xiaofeiyang@gzhu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3330957

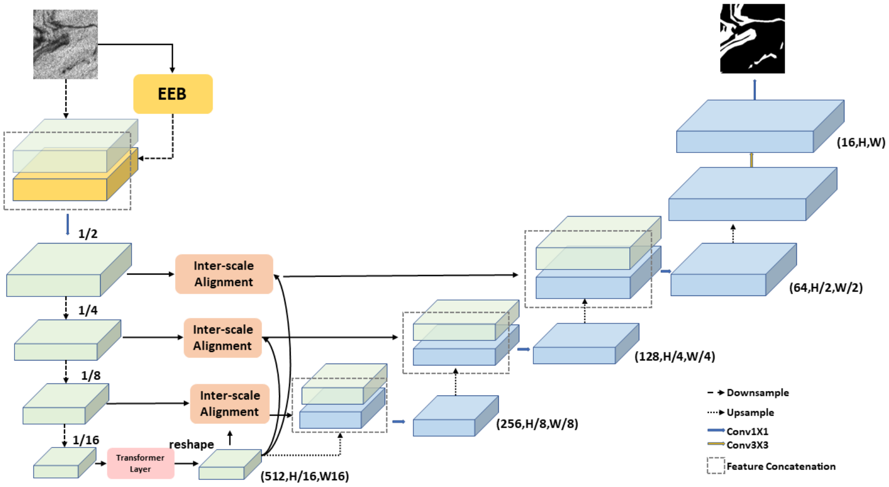

Fig. 1. Architecture of DS-Unet, which is based on U-Net and composed of EEB and interscale alignment module.

first performs an edge extraction module of the Sobel operator to identify the edge feature. Then, a novel interscale alignment module can generate hybrid feature maps with dual-scale features. Finally, the "U"-type structure can learn the local and global features. Comprehensive experiments on the Palsar and Sentinel datasets illustrate the excellent performance of the proposed DS-Unet. The key contributions of the letter are summarized as follows.

1) Unlike previous methods, we designed a new interscale alignment module to seamlessly integrate both local and global features from various feature maps.
2) We introduce the Sobel operator as an edge feature extraction module to handle the speckle noise of SAR images.

The rest of this letter is organized as follows. The DS-Unet is introduced in Section II. Section III shows the experimental analysis. The conclusion is given in Section IV.

II. Proposed Method

In this section, we first define the input of the DS-Unet, which is a SAR image $x \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ is the resolution of the image and C is its channels. Then, the final output of the DS-Unet is $H \times W \times 2$, where 2 denotes the number of categories to which the segmented object belongs.

A. Dual-Stream-Unet

The overall architecture of the DS-Unet is presented in Fig. 1. It consists of four parts: a block for edge feature extraction, a CNN–Transformer hybrid encoder for local and global feature extraction, an interscale alignment module for the fusion of dual-scale images, and a decoder for the restoration of image resolution.

1) Edge Extraction Block: We first use an edge extraction block (EEB) and an operation of concat to find the entire contour of the target to be segmented. A SAR image is a remote-sensing grayscale image that is easily affected by speckle noise, in which significant changes in grayscale values indicate the presence of edges. Hence, we construct the EEB using the Sobel operator, which can extract features in the horizontal and vertical directions of the image through two fixed weighted 3×3 convolutional kernels G_H, G_V. The detailed values of G_H and G_V are as follows:

$$G_H = \begin{bmatrix} -1 & 0 & 1 \\ 2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$
$$G_V = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}.$$

The final output of the EEB module can be written as

$$f_e = \sqrt{(f_i * G_H)^2 + (f_i * G_V)^2} \tag{1}$$

where f_i is the input image with a size $H \times W \times C$, f_e is the output image with a size $H \times W \times C$ through the EEB module, and $*$ denotes the convolutional operation. To add edge features to the original image, the two images simultaneously pass through a 7×7 convolutional operation which is the first stage of Resnet50. The final fused image can be calculated as

$$f_e' = \mathrm{Relu}(\mathrm{GN}(R_{\mathrm{cov}}(f_e))) \tag{2}$$
$$f_i' = \mathrm{Relu}(\mathrm{GN}(R_{\mathrm{cov}}(f_i))) \tag{3}$$
$$f_o = H_{\mathrm{cov}}\left(\mathrm{concat}\left(f_e', f_i'\right)\right) \tag{4}$$

where R_{cov} denotes the operation of 1×1 convolution, H_{cov} denotes the operation of 7×7 convolution, GN represents the GroupNorm layer, and f_o is the final output feature map with a size $H \times W \times 64$. This module relies on the Sobel operator as a constraint for displaying edges. It can guide subsequent feature extraction operations to focus more on the information associated with these edges while disregarding isolated areas characterized by small regions that might be influenced by speckle noise. The Sobel operator has been proven to be effective in removing noise in some denoising networks, such as AGSDNet [8].

2) CNN–Transformer Hybrid Encoder: Inspired by TransUnet, we think that a powerful encoder can make outstanding contributions to the application effect of a model. In DS-Unet, we use the CNN–Transformer hybrid model. The encoder consists of a variant of Resnet50 [9] and a Transformer layer. The number of bottlenecks in each block of the resnet50 variant is [3, 6, 9]. Through the CNN operation of resnet, the local features of the image will be extracted. This will generate four feature maps with a resolution of 1/2, 1/4, 1/8, and 1/16 of the original image resolution. To retain global patterns, the lowest feature map $((W/16) \times (H/16) \times 1024)$ go through a Transformer layer. The input data for the layer first goes through a position embedding operation with a default patch size of 16. The Transformer layer consists of layers of multihead self-attention (MSA), multilayer perceptron (MLP), and LayerNorm. The calculation operation of the entire Transformer layer is as follows:

$$\begin{aligned} t_i' &= \mathrm{MSA}(\mathrm{LN}(t_i)) + t_i \\ t_o &= \mathrm{MLP}\big(\mathrm{LN}\big(f_i'\big)\big) + t_i' \end{aligned} \tag{5}$$

where LN() denotes the LayerNorm, t_i represents the data that have been embedded, and t_o represents the data output from this layer.

3) Interscale Alignment: Interscale alignment can fuse high-level (1/2, 1/4, 1/8) images with local features and low-level (1/16) images with global features into three new feature maps. It is well known that the high-level image processed by CNNs has better local features and the low-level image has a larger receptive field and more semantic information, which is suitable for establishing long-distance dependencies between parts of the image. Hence, the interscale alignment module can find the correlation between each token of two different level images and adaptively integrate global and local features.

Specifically, the input is a local feature map f_h with the size of $(W/S) \times (H/S) \times C_h$ and a global feature map reshaped from the Transformer layer with the size of $(W/16) \times (H/16) \times C_l$. The values of S, C_h, and C_l are, respectively, [2, 4, 8], [64, 256, 512], and 1024. Then, we use a 1×1 convolution to make f_l to be $(W/16) \times (H/16) \times C_h$. To turn two features into sequences, we reshape them to be $(W \times H)/S^2 \times C_h$ and $(W \times H)/256 \times C_h$. After this, coattention is used to obtain the connection between them. The calculation process is as follows and is shown in Fig. 2:

$$\begin{aligned} Q_h &= f_h W_Q^h, \quad K_h = f_h W_K^h, \quad V_h = f_h W_V^h \\ Q_l &= f_l W_Q^l, \quad K_l = f_l W_K^l, \quad V_l = f_l W_V^l \end{aligned} \tag{6}$$

where $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{C_h \times C_h}$ and $W_Q^l, W_K^l, W_V^l \in \mathbb{R}^{C_l \times C_l}$ denote matrices to generate queries, keys, and values, respectively. Then, attention weights can be calculated as

$$\begin{aligned} W_{h \to l} &= \mathrm{softmax}\left(\frac{Q_h K_l^T}{\sqrt{d}}\right) \\ W_{l \to h} &= \mathrm{softmax}\left(\frac{Q_l K_h^T}{\sqrt{d}}\right) \end{aligned} \tag{7}$$

where the value of d is C_h. Then, we get new f_h and f_l that have both local and global features. We separately reshape them to $(W/S) \times (H/S) \times C_h$ and $(W/16) \times (H/16) \times C_h$. After the following operations, we obtain the final output:

$$f_{\mathrm{out}} = H_{\mathrm{cov}}(\mathrm{concat}(f_h, f_l)) \tag{8}$$

where H_{cov} denotes the operation of 1×1 convolution and f_{out} is the output feature map with a size $(W/S) \times (H/S) \times C_h$. Interscale alignment can realize the alignment and fusion of biscale information. The final output is further combined with the deep image of the decoder to compensate for the loss of details.

4) Decoder: The core of the decoder has three levels of convolutional blocks. Each block has two inputs: a feature map from the interscale alignment module in the same level and a feature map 2× upsampling from the lower level. Skip connections between the two feature maps can retain more detailed information. In each block, the model sequentially operates 2× upsampling, concat, the operation of 3×3 convolution, and Relu(). The encoder and the decoder of DS-Unet form an approximately symmetrical structure. The interscale alignment module connects the two like a bridge.

III. Experiments

A. Experimental Setup

1) Dataset: We employ the Palsar dataset and the Sentinel dataset to verify the performance of the proposed model, which can be downloaded at https://grzy.cug.edu.cn/zhuqiqi/en/zhym/32383/list/index.htm [10]. The Palsar dataset is sourced from oil spill events in the Gulf of Mexico between May 2010 and August 2010. The dataset is taken by the ALOS satellite, equipped with PALSAR sensors operating in HH polarization. It consists of 14 raw SAR images that are PALSAR (level 1.5) data products, employing a polarimetry observation mode with a pixel spacing of 12.5 m. Following the removal of samples with a lower proportion of oil spill areas, the initial set of 14 SAR images was divided into 3807 individual images, each measuring 256 × 256 pixels. Among these, 3101 images were allocated for training purposes, while the remaining 776 were designated for testing. The Sentinel dataset is sourced from the Persian Gulf region and captured by the Sentinel-1A satellite in August 2017, which is equipped with a C-band SAR sensor operating in VV polarization. The dataset comprises seven raw SAR images, specifically Level-1 Interferometric Wide Swath GRD Products, with a spatial resolution of 5 m in the cross-track direction and 20 m in the along-track direction. Following preprocessing operations, such as radiation correction, terrain correction, and the removal of samples with a lower proportion of oil spill areas, the original set of seven SAR images was divided into 4193 individual images, each measuring 256 × 256 pixels. Out of these, 3354 images were allocated for training purposes, while the remaining 839 were designated for testing. It is worth mentioning that the Palsar dataset and the Sentinel dataset are entirely independent, with SAR images carefully chosen under diverse weather, seasonal, lighting, imaging conditions, and scales. These images were acquired from multiple remote-sensing satellites featuring varying spatial resolutions and spectral coverages.

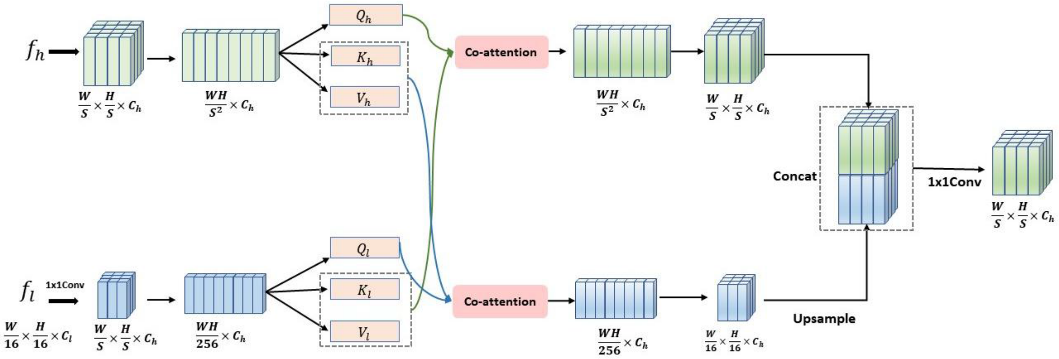
f_h
$\frac{W}{S} \times \frac{H}{S} \times C_h$
$\frac{WH}{S^2} \times C_h$
Q_h
K_h
V_h
Co-attention
$\frac{WH}{S^2} \times C_h$
$\frac{W}{S} \times \frac{H}{S} \times C_h$
Concat
1x1Conv
$\frac{W}{S} \times \frac{H}{S} \times C_h$
f_l
1x1Conv
$\frac{W}{16} \times \frac{H}{16} \times C_l$
$\frac{W}{S} \times \frac{H}{S} \times C_h$
$\frac{WH}{256} \times C_h$
Q_l
K_l
V_l
Co-attention
$\frac{WH}{256} \times C_h$
$\frac{W}{16} \times \frac{H}{16} \times C_h$
Upsample

Fig. 2. Architecture of interscale alignment module.

2) Implementation Details: For all experiments, we use random rotation and flipping to increase data diversity. Our model is trained on a single NVIDIA GeForce RTX 3090 with an SGD optimizer with a learning rate of 0.01, a momentum of 0.9, and a weight decay of 1e-4. The default batch size is 24 and the default number of training iterations is 21k for the Palsar dataset and 39k for the Sentinel datasets.

3) Evaluation Metrics: We use conventional semantic segmentation evaluation metrics: average Dice-Similarity coefficient (DSC), average Hausdorff distance (HD), and binary classification evaluation metrics F1 score. DSC is a measure of set similarity used to calculate the similarity between two samples. HD describes the similarity between two sets of point sets. The F1 score takes into account both the accuracy and recall of the classification model. It can be seen as a harmonic average of model accuracy and recall. We also measured the running time of different methods used to infer a 256 × 256 pixel image. The calculation methods for DSC, HD, and F1 are as follows:

$$\mathrm{DSC} = \frac{2 \times |\boldsymbol{X} \cap \boldsymbol{Y}|}{|\boldsymbol{X}| + |\boldsymbol{Y}|} \tag{9}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$
$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{10}$$

$$\mathrm{HD}(\boldsymbol{X}, \boldsymbol{Y}) = \max(h(\boldsymbol{X}, \boldsymbol{Y}), h(\boldsymbol{Y}, \boldsymbol{X}))$$
$$h(\boldsymbol{X}, \boldsymbol{Y}) = \max_{x \in \boldsymbol{X}} \left\{ \min_{y \in \boldsymbol{Y}} \|x - y\| \right\}$$
$$h(\boldsymbol{Y}, \boldsymbol{X}) = \max_{y \in \boldsymbol{Y}} \left\{ \min_{x \in \boldsymbol{X}} \|y - x\| \right\} \tag{11}$$

where $\boldsymbol{X}$ denotes the prediction results and $\boldsymbol{Y}$ denotes the true label. TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

4) Compared Methods: We compared DS-Unet with eight other methods, including: R2UNet [11], Swin-Unet [12], TransUnet [5], FCN [2], U-Net [4], NestedUNet [13], AttnUNet [14], and SegNet [15].

5) Loss Functions: A mixed loss consisting of dice loss and cross-entropy loss is used. In training, dice loss pays more attention to the mining of segmentation targets, which may cause loss saturation. Cross-entropy loss calculates the loss of each pixel equally. Therefore, using two kinds of loss in combination will have a better effect. The total loss is calculated as follows:

$$\mathrm{Loss} = 0.4 \times \mathrm{CrossEntropyLoss} + 0.6 \times \mathrm{DiceLoss}. \tag{12}$$

6) Ablation Methods: To explore the role of EEB in oil spill detection tasks, we also implement a variant of DS-Unet. The model is named DS-Unet-N and removes the EEB branch from the original DS-Unet.

B. Performance Comparison

We compare our method with eight state-of-the-art semantic segmentation methods. The results are shown in Tables I and II. Red represents the best result, and blue represents the second best.

As can be seen in Table II, our proposed DS-Unet has achieved state-of-the-art performance on all segmentation evaluation metrics. Specifically, for the DSC metric, the DS-Unet outperforms the AttnUNet model by 1.49% on the Sentinel dataset and outperforms the R2U-Net model by 0.98% on the Palsar dataset. For the F1 metric, our DS-Unet exceeds the NestedUNet model by 0.37% on the Palsar dataset and exceeds the NestedUNet model by 0.79% on the Sentinel dataset. For the HD metric, DS-Unet outperforms the NestedUNet model by 0.73 on the Palsar dataset and outperforms the NestedUNet model by 0.48 on the Sentinel dataset. The reason for this phenomenon is that our DS-Unet can combine edge, local, and global features with different dimensional feature maps to make an accurate prediction. At the same time, DS-Unet overcomes its variant, DS-Unet-N, and the image EEB can further enhance the segmentation ability of the DS-Unet.

In terms of running time (Table I), it is found that the U-Net method performs the fastest since it is the most basic segmentation framework with 3 × 3 convolution operations. The running time of the R2U-Net is the longest since it requires multiple recurrent convolution operations. It is worth noting that our DS-Unet is not the fastest method but exhibits the highest efficiency. The running time of the DS-Unet is acceptable. Our future research goal will focus on enhancing the running time of this method.

Finally, we also visually compare the segmentation results of all the methods on the two datasets. As can be seen in Fig. 3,

TABLE I
RUNNING TIME OF ALL METHODS

	Swin-Unet	TransUnet	FCN	NestedUnet	U-net	AttnUNet	R2U-Net	SegNet	DS-Unet
Time	10.86	16.25	8.58	5.50	**4.22**	11.30	19.04	**4.98**	15.39

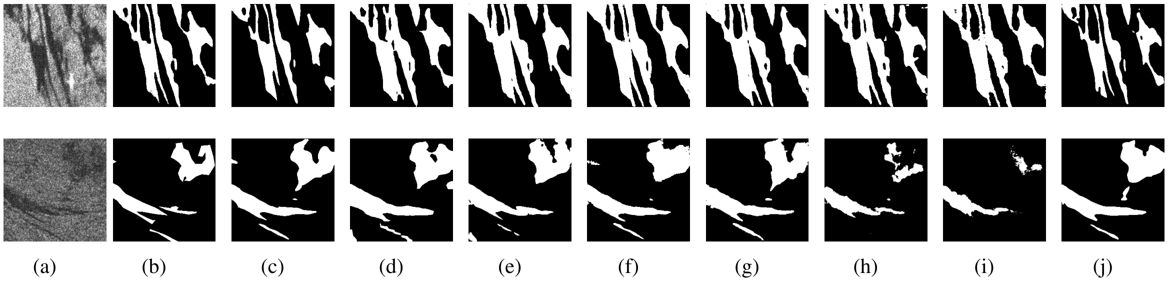

Fig. 3. Experimental results of different semantic segmentation methods on two datasets. (a) Original image. (b) True segmentation results. (c) TransUnet. (d) FCN. (e) NestedUNet. (f) U-Net. (g) AttnUNet. (h) R2U-Net. (i) SegNet. (j) Proposed DS-Unet.

TABLE II
RESULTS ON THE PALSAR/SENTINEL DATASET

Method	DSC↑	HD↓	F1↑
Swin-Unet	0.6856 / 0.6156	39.96 / 66.22	0.6353 / 0.5751
TransUnet	0.7290 / 0.7539	27.81 / **35.58**	0.7226 / **0.7372**
FCN	0.7380 / 0.7355	27.95 / 37.24	0.7238 / **0.7372**
NestedUNet	0.7378 / 0.7345	**26.62** / 40.71	0.7288 / 0.7298
U-Net	0.7167 / 0.7254	31.09 / 42.47	0.7167 / 0.7254
AttnUNet	0.7332 / 0.7431	27.08 / 48.92	0.7280 / 0.6815
R2U-Net	0.7395 / 0.7054	31.66 / 49.35	0.7176 / 0.5864
SegNet	0.6783 / 0.6303	34.34 / 54.28	0.6783 / 0.6303
DS-Unet-N	**0.7414** / **0.7573**	26.74 / 36.52	**0.7324** / 0.7311
DS-Unet(Ours)	**0.7493** / **0.7580**	**25.89** / **35.10**	**0.7325** / **0.7377**

the result of the DS-Unet can better display the contour and extract more detailed information. The predictions of R2U-Net and SegNet are very blurry and without important details. The visualization comparison of the datasets again demonstrates the superiority of the DS-Unet. However, the visualization result of Fig. 3(j) shows the presence of false positives in the predictions of the DS-Unet. The possible reasons for this phenomenon are limitations of the Sobel operator, which is a first derivative operator and cannot discern semantic information resulting in imperfect edge detection accuracy. Hence, while the Sobel operator effectively filters smaller and darker areas resembling oil spill regions, it may misidentify darker and denser areas, such as Fig. 3(j) or low-wind-speed sea surface areas. In the future, we will continue to study how to avoid the occurrence of this false-positive phenomenon.

IV. CONCLUSION

In this letter, we proposed a DS-Unet model for oil spill detection of SAR images. The proposed DS-Unet consists of two modules: an edge feature extraction module extracting oil spill contours to reduce noise influence, and an interscale alignment module integrating local and global features. These designs enable DS-Unet to extract dual-scale features and learn global representations. Finally, experiments on two real datasets show that the proposed DS-Unet model has excellent performance. Moreover, the proposed DS-Unet model represents a promising approach for real-world applications.

REFERENCES

[1] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[2] P. Duan, X. Kang, P. Ghamisi, and S. Li, “Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509711.

[3] 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.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Munich, Germany: Springer, Oct. 2015, pp. 234–241.

[5] J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, *arXiv:2102.04306*.

[6] P. Duan, Z. Xie, X. Kang, and S. Li, “Self-supervised learning-based oil spill detection of hyperspectral images,” *Sci. China Technol. Sci.*, vol. 65, no. 4, pp. 793–801, Apr. 2022.

[7] M. Mao et al., “Dual-stream network for visual recognition,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 25346–25358.

[8] R. K. Thakur and S. K. Maji, “AGSDNet: Attention and gradient-based SAR denoising network,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

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

[10] Q. Zhu et al., “Oil spill contextual and boundary-supervised detection network based on marine SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5213910.

[11] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, “Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation,” 2018, *arXiv:1802.06955*.

[12] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-UNet: UNet-like pure transformer for medical image segmentation,” 2021, *arXiv:2105.05537*.

[13] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, Sep. 2018, pp. 3–11.

[14] O. Oktay et al., “Attention U-Net: Learning where to look for the pancreas,” 2018, *arXiv:1804.03999*.

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