Development of a Dual-Attention U-Net Model for Sea Ice and Open Water Classification on SAR Images

Yibin Ren, *Associate Member, IEEE*, Xiaofeng Li, *Fellow, IEEE*, Xiaofeng Yang, *Senior Member, IEEE*, and Huan Xu

***Abstract*— This study develops a deep learning (DL) model to classify the sea ice and open water from synthetic aperture radar (SAR) images. We use the U-Net, a well-known fully convolutional network (FCN) for pixel-level segmentation, as the model backbone. We employ a DL-based feature extracting model, ResNet-34, as the encoder of the U-Net. To achieve high accuracy classifications, we integrate the dual-attention mechanism into the original U-Net to improve the feature representations, forming a dual-attention U-Net model (DAU-Net). The SAR images are obtained from Sentinel-1A. The dual-polarized information and the incident angle of SAR images are model inputs. We used 15 dual-polarized images acquired near the Bering Sea to train the model and employ the other three images to test the model. Experiments show that the DAU-Net could achieve pixel-level classification; the dual-attention mechanism can improve the classification accuracy. Compared with the original U-Net, DAU-Net improves the intersection over union (IoU) by 7.48.% points, 0.96.% points, and 0.83.% points on three test images. Compared with the recently published model DenseNet$_{FCN}$, the three improvement IoU values of DAU-Net are 3.04.% points, 2.53.% points, and 2.26.% points, respectively.**

***Index Terms*— Dual-attention, sea ice and open water classification, synthetic aperture radar (SAR) image, U-Net.**

I. Introduction

THE changes in global sea ice volume, distribution, and movement reflect the interaction of the atmosphere–cryosphere–hydrosphere and the global climate change [1]. Sea ice study is also significant because it causes marine navigation and transportation safety concerns. Since the classification of sea ice and open water provides valuable information for safe navigation, sea ice classification, and monitoring draw extensive attention [2]. Satellite remote sensing, such as optical camera, microwave radiometer, and synthetic aperture radar (SAR), has been the most effective way to monitor sea ice in the polar regions. SAR images have been the primary source for sea ice classification and monitoring due to their high spatial resolution, wide coverage, and ability to penetrate clouds [3].

A series of studies have been devoted to classifying sea ice and open water on SAR images, including threshold-based methods [4], expert systems [5], and machine learning (ML) methods. Since the 21st century, ML has become the mainstream method, such as the neural network (NN) model [6], [7], support vector machine (SVM) [8], [9], and random forest (RF) classifier [10].

With the innovation of the method, the accuracy of sea ice classification is improving. However, the methods above need prior expert knowledge and sophisticated manual engineering. This drawback has been a common challenge faced by earth system science in the era of big data [11]. Deep learning (DL) addresses this challenge. A typical DL model accepts input data in a raw format and automatically discovers the required features [12]. DL has been successfully applied in oceanography and remote sensing [11], [13]–[16]. A deep convolution NN (CNN) is a particular type of DL model composed of CNN layers. A CNN layer connects to the local patches of the previous layer through convolution kernels to extract local spatial features. Researchers employed CNN to improve the accuracy and efficiency of sea ice classification.

Xu and Scott [17] introduced an early CNN-based model AlexNet to classify sea ice and open water. Li *et al.* [18] proposed a CNN-based model to classify sea ice and open water from Chinese Gaofen-3 SAR images. Wang *et al.* [19] constructed a three-layer CNN model. Gao *et al.* [20] integrated transfer learning and DenseNet to form a transferred multilevel fusion network (MLFN). Boulze *et al.* [21] proposed a model composed of three CNN layers, two max-pooling layers, and four fully connected (FC) layers to classify the detailed sea ice types.

CNN-based models achieve end-to-end classification between sea ice and open water on SAR images. However, most of these CNN-based models use an image classification framework, not a pixel-level segmentation framework. A typical procedure of these CNN-based models includes

Manuscript received November 13, 2020; revised January 6, 2021; accepted February 3, 2021. Date of publication February 23, 2021; date of current version December 22, 2021. This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) under Grant XDA19060101 and Grant XDB42040401, in part by the China Postdoctoral Science Foundation under Grant 2019M662452, in part by the Key Research and Development Project of Shandong Province under Grant 2019JZZY010102, in part by the Key Deployment Project of Center for Ocean Mega-Science, CAS under Grant COMS2019R02, in part by the CAS Program under Grant Y9KY04101L, and in part by the National Natural Science Foundation of China under Grant 41776183. *(Corresponding author: Xiaofeng Li.)*

Yibin Ren and Xiaofeng Li are with the Key Laboratory of Ocean Circulation and Waves, Institute of Oceanology, CAS, Qingdao 266071, China, and also with the Center for Ocean Mega-Science, CAS, Qingdao 266071, China (e-mail: yibinren@qdio.ac.cn; xiaofeng.li@ieee.org).

Xiaofeng Yang is with the State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, CAS, Beijing 100101, China, and also with the Key Laboratory of Earth Observations of Hainan Province, Sanya 572029, China (e-mail: yangxf@radi.ac.cn).

Huan Xu is with the School of Geomatics and Marine Information, Jiangsu Ocean University, Lianyungang 222005, China (e-mail: 2018224050@jou.edu.cn).

Digital Object Identifier 10.1109/LGRS.2021.3058049

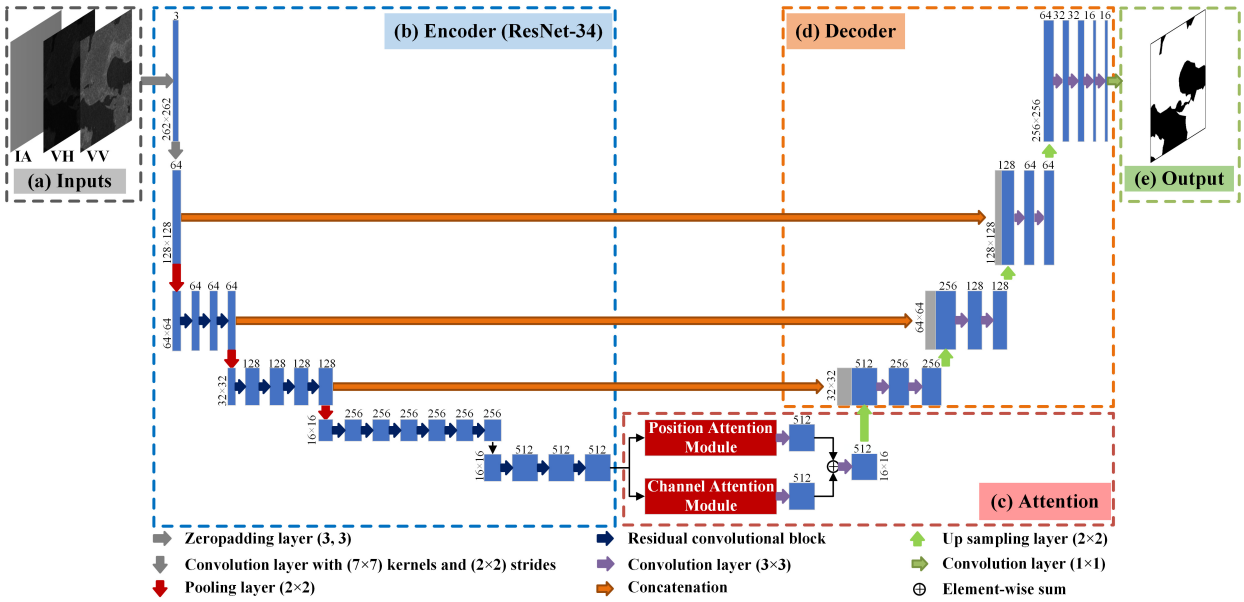

Fig. 1. Overall structure of the proposed DAU-Net model. (a) Inputs of DAU-Net: VV, VH, and incident angle. (b) Encoder of DAU-Net, based on ResNet-34. (c) Dual-attention structure we employed to optimize the original U-Net, including PAM, CAM, and three CNN blocks. (d) Decoder of DAU-Net. (e) Output module of DAU-Net.

two steps: 1) inputting an SAR image chip (e.g., with the size of 13 × 13, 45 × 45, or 50 × 50 [10], [17]–[20]) into several stacking pairs of CNN + max-pooling layers to extract downscaled feature using FC NN layers to connect the downscaled feature maps and output the class of the central pixel of the input chip. The drawback is that we need manually construct training/testing samples for each pixel of an image, increasing the workload and the storage burden of memory. The FC NN layers have more trainable parameters and increase computational complexity. Recently, fully convolutional networks (FCNs) have been proposed to replace the FC layers of image classification frameworks to achieve pixel-level segmentation. In an FCN-based segmentation framework, the downscaled feature maps are rescaled to the same size as the input image by upsampling layers. The FCN works on the rescaled feature maps and directly outputs the classification results of all pixels, not only the central pixel.

To solve the mentioned drawback, we proposed an FCN-based model to classify sea ice and open water on SAR images. The proposed model integrates the dual-attention mechanism into the U-Net and is named dual-attention U-Net model (DAU-Net). The U-Net is a well-known FCN-based segmentation model and is named for its almost symmetric encoder–decoder network architecture like a "U" shape [22]. The dual-attention mechanism [23] is employed to enhance the feature characteristics to improve the classification accuracy further. We use SAR images near the Bering Sea to train and evaluate the model. The main contributions of this study are: we propose FCN-based sea ice and open water classification model DAU-Net, and integrating the dual-attention mechanism into the original U-Net is helpful to improve the classification performance.

The rest of this letter is organized as follows. Section II describes the proposed DAU-Net in detail. In Section III, comparison experiments are conducted to evaluate the effectiveness of the DAU-Net. Section IV concludes this letter.

II. Method

A. Overall Structure of DAU-Net

The overall structure of DAU-Net is shown in Fig. 1, including five parts: inputs, encoder, attention, decoder, and output. An input unit consists of three channels of a 256 × 256 pixel SAR image: VV (vertical emitting and vertical receiving), VH (vertical emitting and horizontal receiving), and incident angles [see Fig. 1(a)]. The encoder is the residual-CNN-network-34 (ResNet-34) [24]. It extracts abstracted, downscaled high-level feature maps. The attention part includes the position attention module (PAM) and channel attention module (CAM). The decoder restores the resolution of the feature maps. Skip connections pass the intermediate feature maps extracted by the encoder to the decoder. The output module generates pixel-level classification results. Following, we will detail each part.

B. Encoder

The ResNet-34 encoder includes five stages: one zero-padding layer with a 3 × 3 kernel, one CNN layer, and four ResNet stages containing 3, 4, 6, and 3 ResNet blocks [see Fig. 1(b)]. The number of convolutional kernels in the five stages is 64, 64, 128, 256, and 512, respectively. We retain the first three max-pooling layers of the ResNet-34. After encoding, the inputs are transformed into 16 × 16 × 512 feature maps.

C. Attention

Discriminant feature representations are essential for high accuracy classification. We integrate PAM and CAM into the original U-Net to improve the feature representations of sea ice. The PAM captures long-range dependencies in spatial dimensions. The CAM captures the channel dependencies between any two channels.

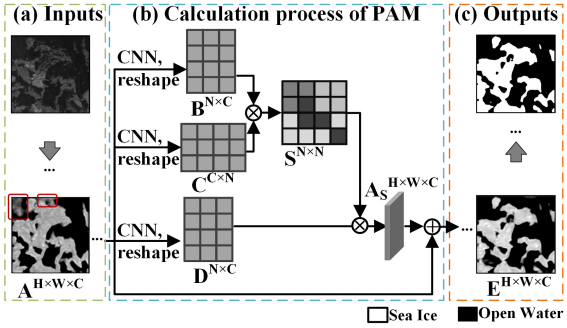

Fig. 2. Detailed calculation process of PAM in the DAU-Net. (a) Feature map A without PAM, some water pixels are inaccurately encoded as ice pixels, marked in the red rectangle. (b) Calculation process of PAM. (c) Feature map E after PAM, some inaccurate ice pixels are optimized.

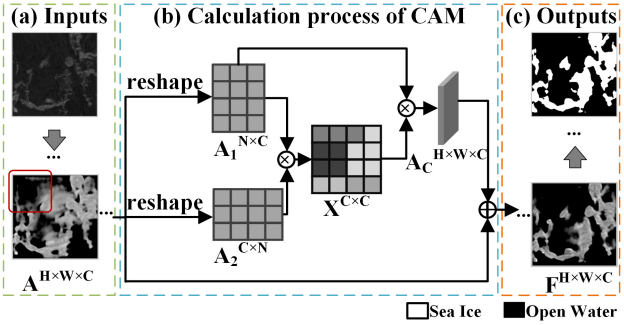

Fig. 3. Detailed calculation process of CAM in the DAU-Net. (a) Feature map A without CAM, some water pixels are inaccurately encoded as ice pixels, marked in the red rectangle. (b) Calculation process of CAM. (c) Feature map F after CAM, some inaccurate ice pixels are optimized.

1) PAM: For semantic segmentation, local features captured by standard CNN could lead to misclassifications. The PAM updates the feature value at a specific position by aggregating feature values at all positions with a weighted summation. The global spatial dependencies of any two positions could be captured.

As shown in Fig. 2(a), $A^{H\times W\times C}$ is the feature map extracted by the encoder. H, W, and C are the rows, columns, and channels, respectively. The gray/dark regions represent ice/water features. The red rectangle marks some inaccurate features, which should be water but are encoded to be sea ice. A is transformed and reshaped to $B^{N\times C}$, $C^{C\times N}$, and $D^{N\times C}$ ($N = H \times W$) [see Fig. 2(b)]. A matrix multiplication and softmax activation is performed on B and C to obtain the spatial attention map $S^{N\times N}$. The more similar features of two positions (i, j) generate a large element at S_{ij}. S is multiplied by D and reshaped to $A_S^{H\times W\times C}$. For each channel of A_S, the element of a position is the weighted sum of elements across all positions in D based on the weights in S. A_S is multiplied by a scale parameter α and added to the input A in element-wise to obtain the output $E^{H\times W\times C}$

$$E = \alpha A_S + A \tag{1}$$

where α is a trainable parameter that gradually learns to assign weight from 0 by the backpropagation. E integrates the local features (A) and the global features (A_S). Intuitively, as shown in Fig. 2(c), feature map E optimizes the inaccurate features in A by the PAM, resulting in accurate segmentation.

2) CAM: There are 512 channels in the high-level feature map output by the ResNet-34. The CAM updates the feature value at a position by aggregating all channels' feature values with a weighted sum. The channels corresponding to the same or similar class responses will be assigned with large weights. The interdependencies among channels can be captured.

As shown in Fig. 3(a), $A^{H\times W\times C}$ is the feature map output by the encoder. A is reshaped to $A_1^{N\times C}$ and $A_2^{C\times N}$. A_2 is multiplied by A_1 and is activated by a softmax function to obtain the channel attention $H^{C\times C}$. A_1 is multiplied by the transpose of X and is reshaped to Ac. For each position of Ac, the element of a channel is the weighted sum of elements across all channels in the corresponding position of A. Ac is multiplied by a scale parameter β and added to A element-wise to obtain the output $F^{H\times W\times C}$:

$$F = \beta A_C + A \tag{2}$$

where β is a trainable parameter that gradually learns a weight from 0 by the backpropagation. The final feature map is a weighted sum of the original feature and the global semantic dependencies between different channels, boosting feature discriminability. In Fig. 3(c), CAM optimizes the inaccurate features in Fig. 3(a), resulting in accurate segmentation.

3) Fusion: Fig. 1(c), the PAM output, and CAM output are transformed by two CNN layers. An element-wise summation is performed on the two results. A CNN layer further fuses the summation.

TABLE I
INFORMATION ON TESTING DATA SETS

Data	Date	Coverage			
		Min_Lon	Max_Lon	Min_Lat	Max_Lat
$Data_1$	31/03/2019	-171.2750	-167.4583	64.1508	65.8461
$Data_2$	24/04/2019	-170.8583	-167.2417	60.8167	62.5114
$Data_3$	30/04/2020	-171.8583	-168.2417	60.8489	62.5914

D. Decoder

Fig. 1(d), the DAU-Net, has five decoder modules. Each decoder module contains one upsampling layer and two CNN layers. The number of convolutional kernels of the five decoders is 256, 128, 64, 32, and 16, respectively. After decoding, the 16×16 feature maps are rescaled to the same size of the input.

E. Output

The output module contains one CNN layer and one sigmoid layer. The CNN layer transforms the last feature map of the decoder into a $256 \times 256 \times 1$ feature map. The sigmoid layer activates the feature map and output values between [0, 1]. If it is larger than 0.5, the pixel is sea ice; otherwise, it is open water. The loss function is binary cross-entropy.

III. EXPERIMENTS

A. Experiment Data and Setting

The experiment data locate near the Bering Strait, with a spatial extent of [173°W, 163°W, 69°N, 61°N].

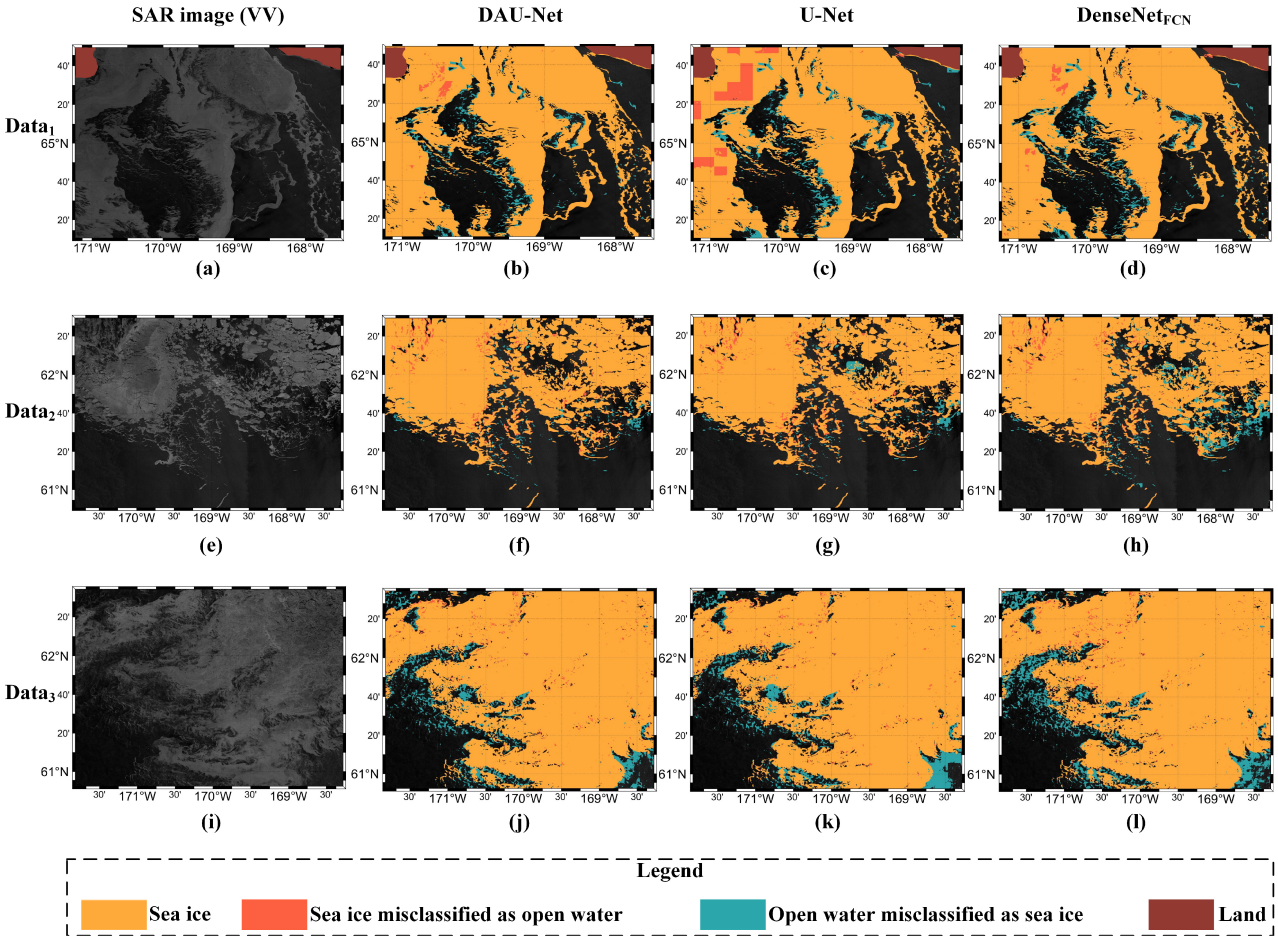

Fig. 4. SAR image (VV channel), the classification results of DAU-Net, U-Net, and DenseNet$_{FCN}$. (a)–(d) Data$_1$. (e)–(h) Data$_2$. (i)–(l) Data$_3$.

The SAR images are the ground range detected high resolution (GRDH) products of Sentinel-1A in the interferometric wide-swath (IW) mode with VV + VH polarizations. We use Sentinel Application Platform (SNAP) 3.0 to perform radiometric calibration. We downscale each image to 1/3 of its original size. The downscaled pixel spacing is 30 m × 30 m (rg × az). The images are labeled by LabelMe [25] based on visual interpretation to obtain ground truths:1 for sea ice and 0 for water. We choose image blocks that are easily distinguishable to build the training set. Inevitably, a few mislabeling pixels, especially some small sea ice objects, exist; such mislabeled pixels account for a small proportion of all pixels and do not affect the model's convergence. The training set includes 15 SAR images, from December to April of the next year. We move a 256 × 256 box across the entire SAR image to clip the image into input chips. We select 4684 chips to train the model. The other three SAR images are used as the test set, named Data$_1$, Data$_2$, and Data$_3$. Fig. 4(a), (e), and (f) shows the VV channels of the test images. Table I shows the details.

The hardware is a workstation with one NVIDIA TESLA V100 32 GB GPU. We use Keras as the DL packages. The mean intersection over union (IoU), accuracy, precision, and recall are four metrics.

B. Comparison Experiments Against Other Models

We compare DAU-Net with the other two FCN models. The first one is the original U-Net, without a dual-attention mechanism. The other one is the DenseNet$_{FCN}$, modified based on the MLFN proposed in 2019 [20]. To make a fair comparison, we replace the FC layers in the MLFN with FCN layers and add upsampling blocks. All three models output the pixel-level classification results of the input chip directly.

TABLE II
METRICS OF DIFFERENT MODELS

Data	Model	IoU	Accuracy	Precision	Recall
Data$_1$	DenseNet$_{FCN}$	91.62%	95.55%	94.28%	97.01%
	U-Net	87.18%	93.17%	93.73%	92.58%
	DAU-Net	**94.66%**	**97.22%**	**96.14%**	**98.40%**
Data$_2$	DenseNet$_{FCN}$	87.07%	94.20%	93.85%	92.33%
	U-Net	88.64%	95.04%	96.50%	91.58%
	DAU-Net	**89.60%**	**95.49%**	**97.24%**	**91.93%**
Data$_3$	DenseNet$_{FCN}$	89.35%	92.19%	93.07%	94.24%
	U-Net	90.78%	93.27%	94.21%	96.15%
	DAU-Net	**91.61%**	**93.95%**	**95.26%**	**95.17%**

The metrics and classification results of the test images are shown in Table II and Fig. 4. For three images, the IoUs and accuracies of DAU-Net are the highest. Some differences occur among different test sets. Data$_1$ is an entire block of sea ice with an open water region inside. The upper left corner of the image has some dark pixels. U-Net performs the worst as

it misses the most sea ice pixels [see Fig. 4(c)]. Comparing Fig. 4(b) and (d), DAU-Net misses less sea ice than U-Net. The results of DAU-Net and $\text{DenseNet}_{\text{FCN}}$ are very close. In terms of metrics, DAU-Net performs a little better than $\text{DenseNet}_{\text{FCN}}$. As shown in Fig. 4(e), Data_2 contains many small floes and could be challenging for the ice/water classifier. Fig. 4(g) and (h) shows that the U-Net and $\text{DenseNet}_{\text{FCN}}$ misclassify some water pixels as sea ice pixels. DAU-Net corrects the misclassifications in the middle of the image and improves the classification results in the lower right corner [see Fig. 4(f)]. Data_3 has a complex sea ice boundary. Fig. 4(j)–(l) shows that all three models generate false alarms in the lower right corner of the image. The false alarms of DAU-Net are the lowest. Compared with the original U-Net, DAU-Net improves the IoU by 7.48.% points, 0.96.% points, and 0.83.% points on three test images. For $\text{DenseNet}_{\text{FCN}}$, the three improvement values are 3.04.% points, 2.53.% points, and 2.26.% points, respectively.

DAU-Net performs the best among the three models. The main reason is that the feature map extracted by DAU-Net integrates the local feature extracted by CNN and the global feature generated by the attention mechanism. In the other two models, there are only local features extracted by CNN, limiting the characterization ability of the feature map and affecting segmentation accuracy.

IV. Conclusion

This study proposes an FCN-based model, DAU-Net, to classify the sea ice and open water on SAR images. We integrate the dual-attention mechanism, PAM and CAM, into the original U-Net model to extract more characteristic features. The proposed model is trained and tested by Sentinel-1A SAR images near the Bering Strait. Experiments show that the DAU-Net achieves pixel-level classification on the SAR image. The dual-attention mechanism helps the DAU-Net achieve better classification results than the original U-Net and $\text{DenseNet}_{\text{FCN}}$. However, the generalizability of the results to larger data sets/wider range of ice conditions still needs to be proven. In the future, we will collect more SAR images to evaluate the applicability of the model.

Acknowledgment

The authors would like to thank the European Space Agency for providing the Sentinel-1 data (https://scihub.copernicus.eu) and the Sentinel Application Platform (SNAP) software (https://step.esa.int/).

References

[1] D. Olonscheck, T. Mauritsen, and D. Notz, "Arctic sea-ice variability is primarily driven by atmospheric temperature fluctuations," *Nature Geosci.*, vol. 12, no. 6, pp. 430–434, Jun. 2019, doi: 10.1038/s41561-019-0363-1.

[2] M. Dabboor and T. Geldsetzer, "Towards sea ice classification using simulated RADARSAT constellation mission compact polarimetric SAR imagery," *Remote Sens. Environ.*, vol. 140, pp. 189–195, Jan. 2014.

[3] P. Maillard, D. A. Clausi, and H. Deng, "Operational map-guided classification of SAR sea ice imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 12, pp. 2940–2951, Dec. 2005.

[4] F. Fetterer, C. Bertoia, and J. Ping Ye, "Multi-year ice concentration from RADARSAT," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Remote Sens.-Sci. Vis. Sustain. Develop. (IGARSS)*, Aug. 1997, pp. 402–404.

[5] L.-K. Soh, C. Tsatsoulis, D. Gineris, and C. Bertoia, "ARKTOS: An intelligent system for SAR sea ice image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 1, pp. 229–248, Jan. 2004.

[6] J. A. Karvonen, "Baltic sea ice SAR segmentation and classification using modified pulse-coupled neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1566–1574, Jul. 2004.

[7] S. Chen *et al.*, "MYI floes identification based on the texture and shape feature from dual-polarized Sentinel-1 imagery," *Remote Sens.*, vol. 12, no. 19, p. 3221, Oct. 2020.

[8] S. Leigh, W. Zhijie, and D. A. Clausi, "Automated ice–water classification using dual polarization SAR satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5529–5539, Sep. 2014, doi: 10.1109/TGRS.2013.2290231.

[9] X.-M. Li, Y. Sun, and Q. Zhang, "Extraction of sea ice cover by Sentinel-1 SAR based on support vector machine with unsupervised generation of training data," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 20, 2020, doi: 10.1109/TGRS.2020.3007789.

[10] J.-W. Park, A. A. Korosov, M. Babiker, J.-S. Won, M. W. Hansen, and H.-C. Kim, "Classification of sea ice types in Sentinel-1 synthetic aperture radar images," *Cryosphere*, vol. 14, no. 8, pp. 2629–2645, Aug. 2020, doi: 10.5194/TC-14-2629-2020.

[11] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019.

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

[13] X. Li *et al.*, "Deep-learning-based information mining from ocean remote-sensing imagery," *Nat. Sci. Rev.*, vol. 7, no. 10, pp. 1585–1606, 2020, doi: 10.1093/nsr/nwaa047.

[14] G. Zheng, X. Li, R. H. Zhang, and B. Liu, "Purely satellite data–driven deep learning forecast of complicated tropical instability waves," *Sci. Adv.*, vol. 6, no. 29, 2020, Art. no. eaba1482, doi: 10.1126/sci-adv.aba1482.

[15] X. Zhang and X. Li, "Combination of satellite observations and machine learning method for internal wave forecast in the sulu and Celebes Seas," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 16, 2020, doi: 10.1109/TGRS.2020.3008067.

[16] B. Liu, X. Li, and G. Zheng, "Coastal inundation mapping from bitemporal and dual-polarization SAR imagery based on deep convolutional neural networks," *J. Geophys. Res.: Oceans*, vol. 124, no. 12, pp. 9101–9113, 2019.

[17] Y. Xu and K. A. Scott, "Sea ice and open water classification of SAR imagery using CNN-based transfer learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3262–3265.

[18] J. Li, C. Wang, S. Wang, H. Zhang, Q. Fu, and Y. Wang, "Gaofen-3 sea ice detection based on deep learning," in *Proc. Prog. Electromagn. Res. Symp.-Fall (PIERS - FALL)*, Nov. 2017, pp. 933–939.

[19] C. Wang, H. Zhang, Y. Wang, and B. Zhang, "Sea ice classification with convolutional neural networks using Sentinel-L scansar images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 7125–7128.

[20] Y. Gao, F. Gao, J. Dong, and S. Wang, "Transferred deep learning for sea ice change detection from synthetic-aperture radar images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1655–1659, Oct. 2019, doi: 10.1109/LGRS.2019.2906279.

[21] H. Boulze, A. Korosov, and J. Brajard, "Classification of sea ice types in Sentinel-1 SAR data using convolutional neural networks," *Remote Sens.*, vol. 12, no. 13, p. 2165, Jul. 2020, doi: 10.3390/RS12132165.

[22] 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.*, Oct. 2015, pp. 234–241.

[23] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[24] 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.

[25] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, May 2008.