# End-to-End Pixel-Wisely Detection of Oceanic Eddy on SAR Images With Stacked Attention Network

Ming Xu, Hongping Li, Yuying Yun, Fan Yang, and Cuishu Li

***Abstract*—Oceanic eddies are ubiquitous phenomena carrying large amounts of energy, thus of great importance for marine ecology and sea–air exchange. As an all-weather, high-resolution sensor, synthetic aperture radar (SAR) could provide valuable observations for oceanic eddies. However, there are few well-established methods for eddy detection on SAR images except for manual seeking, which is laborious and time-consuming. In combination with deep learning, this study is among the earliest in the literature that attempts end-to-end eddy detection on SAR images. Due to obscure pictures and indistinct eddy boundaries, ordinary deep learning models are not adaptable to the objective. Therefore, an customized model, stacked attention network (SANet), is designed to recognize the unique eddy pattern presented on radar images automatically. SANet is a two-unit stacking architecture, with each an hourglass structure for bottom–up, top–down reference and the overall stacking network for iterative extractions of eddy textures contained in shallow layers of each unit. Besides, SANet has included the inner-hourglass attention gates and the outer-hourglass GCblock for the extracted features to be more concentrated on the interested areas. Using SANet, we have identified 87.75% of eddies in the constructed dataset collected from ESA-2 and ENVISAT SAR products. The result is much better than the no-stacking counterpart U-net, as well as state-of-the-art deep learning models DeepLabV3+ and SegFomer, thus verifying the superiority of the proposed method. The generalization ability of the algorithm has also been tested. The code and the constructed SAR dataset have been made public for broader use.**

***Index Terms*—Deep learning, end-to-end detection, oceanic eddies, synthetic aperture radar (SAR).**

## I. Introduction

OCEANIC eddies are rotating structures with scales of tens to hundreds of kilometers and tens to hundreds of days. As relatively concentrated water masses, eddies transport momentum, heat, and substances, such as carbon, phytoplankton, and salt, thereby contributing to the general circulation, large-scale water mass distributions, and ocean biology [1], [2], [3]. As a subject of identifying eddy as a mesoscale phenomenon, eddy detection is extremely important for understanding the ocean dynamic system. And satellite remote sensing, with its advantage of large coverage, is one of the most common ways to accomplish this task.

Oceanic eddies exhibit unique geographic and physical characteristics and can therefore be extracted on a variety of remote sensing imagery, such as sea surface height (SSH), sea surface temperatures (SSTs), seawater flow fields, and synthetic aperture radar (SAR). Among them, however, SSH is of low spatial resolution, and SST variation may also be caused by other oceanic phenomena. In addition, seawater flow fields are either computed by SSH, thus suffering the same defect of low resolution, or acquired by in situ sensors, thus being laborious and discontinuous. Therefore, as a long-lasting, high-resolution, and accurate data source, SAR has been increasingly noticed for eddy detection research, especially for observing the sub- or small-scale eddies, that are found to be ubiquitous but had not attracted much attention until recently.

Various feasible eddy detection methods have been developed for decades, despite the significant contribution they have made to scientific comprehension, these traditional methods have their drawbacks like elaborate selection for thresholds, poor generalization capability, high-level expertise knowledge demanding, etc. Driven by the prosperity of artificial intelligence (AI), some marine scientists are seeking solutions for automatic eddy detection in deep learning communities.

In combination with the advantage of remotely sensed SAR data source and the deep learning application, this article proposes a new automatic eddy detection method. There are some other works committed to the similar channel. Of which DeepEddy [4], [5], and Xia [6] achieve high accuracy, and Yan et al. [7] can distinguish five types of oceanic phenomena. These works perform relatively well in the classification task on potential image patches, while a more comprehensive method is still awaited to identify possible eddies on the entire SAR image, where multiple instances may exist. Furthermore, unlike some previous works that apply existing deep learning models to the eddy detection matter, we focus on eddy features on SAR imagery to design a customized network inspired by stacking architecture, namely stacked attention network (SANet).

The proposed SANet consists of two independent units stacked together, drawing on shallow layers of each unit to extract eddy textures, and refining the information by repeatedly bottom–up, top–down propagating through all units. In addition, two kinds of attention mechanisms within and out of the

Manuscript received 26 March 2023; revised 5 June 2023 and 31 July 2023; accepted 2 October 2023. Date of publication 6 October 2023; date of current version 25 October 2023. This work was supported by the National Program on Global Change and Air-Sea Interaction (Phase II) -Parameterization assessment for interactions of the ocean dynamic system. *(Corresponding author: Hongping Li.)*

Ming Xu, Hongping Li, Yuying Yun, and Fan Yang are with the College of Marine Technology, Ocean University of China, Qingdao 266100, China (e-mail: xuming@stu.ouc.edu.cn; lhp@ouc.edu.cn; 543602074@qq.com; 1589704409@qq.com).

Cuishu Li is with Nanjing Lishui District Garden Management Institute, Nanjing 211200, China (e-mail: 418273715@qq.com).

Code and dataset are available online at https://github.com/xuming1212/SANet-for-eddy-in-SAR.

Digital Object Identifier 10.1109/JSTARS.2023.3322404

individual unit are deployed to further highlight the interested areas containing eddies and suppress the irrelevant backgrounds. Using the architecture, we are able to identify eddies pixel-wisely in a whole massive SAR imagery. Furthermore, by combining the postprocessing based on mathematical morphology, the final detection result achieves a recognition rate of 87.75% on the constructed dataset, much better than several state-of-the-art deep learning networks.

## II. Related Works

### A. Eddy Detection in Traditional Ways

Many studies have been devoted to eddy detection since it was first discovered by a radiation thermometer [8]. From then on, SST has become the earliest source for eddy interpretation. For example, Peckinpaugh and Holyer [9] applied several circle detectors to the reduced advanced very high resolution radiometer (AVHRR) edge images, and compared their ability of defining size and position of eddies. Fernandes and Nascimento [10] developed a three-stage procedure. First, they calculated the vectorial field with the SST map and a zeros matrix. Then, the vectorial field was binarized using an iterative thresholding algorithm. Finally, five edge points classified by their gradient vector direction were selected to fit an ellipse corresponding to the eddy. Oram et al. [11] developed an edge detection algorithm that is insensitive to noise and utilized it to classify cyclonic and anticyclonic eddies in satellite images of the Southern California Bight. Lemonnier et al. [12] used a multiscale analysis of isotherm curvature and a characterization through phase portraits to detect eddy outlines and extract related information about the structures.

Although SST has played an important role in early eddy research. Satellite altimeter data are the dominant resource for eddy detection, since the merged products of two or more altimeters were available. One of the most classic methods based on SSH is the Okubo–Weiss (OW) [1], [13], [14], [15] algorithm. The OW parameter is defined with the geostrophic velocity components, and areas with OW parameter less than the specified threshold are considered eddy regions. Another popular approach is the winding angle (WA) method [16]. It takes the SSH minima or maxima as the eddy center, and the eddy contour is outlined by classifying streamlines with winding angles exceeding 360°. Chelton et al. [2] also used the SSH or sea level anomaly (SLA) extreme-value point as the eddy core, and the boundary was searched from the opposite polarity contour toward the core by deliberate criteria. Yi et al. [17] developed a hybrid algorithm combining the OW method and SSH topology that is capable of recognizing eddy multicore structures. To simplify the recognition process and narrow the search range, Liu et al. [18] divided the global SLA map into several regions and has greatly improved the efficiency.

Oceanic eddies are also visible on the flow field as unique rotating patterns, which inspires research works detecting with the flow velocity. The representative is the vector geometry (VG) method [19]. It derived four constraints characterizing the spatial features of velocity vectors in eddy presence, and the pixel that satisfied these constraints was detected as the eddy center. In addition, eddies carrying chlorophyll makes it possible to identification by ocean color. And there are studies speculating eddies in the North Pacific [20] and the Western South China Sea [21] in this regard.

Due to high-resolution, all-day, all-weather observational advantage, SAR has recently raised more attention in eddy detection domain. Johannessen et al. [22] explored eddy expression on SAR images in relation to wave–current interactions, surface film damping the Bragg waves, and the varying wind field. Based on this, Johannessen et al. [23] proposed a radar imaging model (RIM) concerning surface current and temperature fields, which quantitatively explained eddy signature in SAR images. Xu et al. [24] illustrated the characteristics of oceanic eddies in the Luzon Strait and its adjacent seas by visual interpretation of 426 SAR images, valuable but laborious. Chen et al. [25] used Canny detector to extract eddy edges and estimated the center position with structure characteristics.

### B. Eddy Detection With Deep Learning

Methods mentioned above are all traditional ways that may suffer deficiencies like haphazard parameter setting, erratic threshold initialization, unstable detection accuracy, and low operational efficiency to varying degrees. Therefore, many scientists are pursuing automatic eddy identification through deep learning or AI frames. Benefited from increased computing power, deep learning is gaining popularity in many practical domains [26]. Lots of advanced architectures were constructed for object detection or semantic segmentation, such as SPP-net [27], YOLO [28], Faster R-CNN [29], and U-net [30], DeepLab [31], [32], [33], PSPNet [34], etc.

By transferring the state-of-the-art networks, many eddy detection studies have achieved superior performance than in traditional ways. For example, Lguensat et al. [35] proposed EddyNet based on U-net [30], adding dropout and modifying the loss metric to better fit the eddy detection task. Similarly, Liu et al. [36] used U-net [30] in a multimodal manner applying to multisource remote sensing data. Santana et al. [37] also used U-net [30], together with the plain model and the residual U-net, to discuss how SSH and SLA data sources impact the final results. Xu et al. [38] employed PSPNet [34] as the core algorithm and validated its ability to detect small-scale eddies. Furthermore, Xu et al. [39] included three algorithms of PSPNet [34], DeepLabV3+[33], BiSeNet [40], and compared their detective abilities in terms of eddies numbers, sizes, and lifetimes. Lu et al. [41] applied the HRNet [42] and further refined the results using CascadePSP [43] module. Duo [44] proposed OEDNet based on RetinaNet [45] for eddy identification and enabled positioning and contour seeking. Sun et al. [46] proposed an encoder–decoder model including a modified Xception [47] backbone and several atrous convolutions [32]. Furthermore, Franz et al. [48] delved into the eddy tracking problem using the convolutional long short-term memory [49] network.

All methods mentioned above were based on SSH or SLA, largely because of the easily available altimeter fusion data and the well-established conventional detection algorithms serving

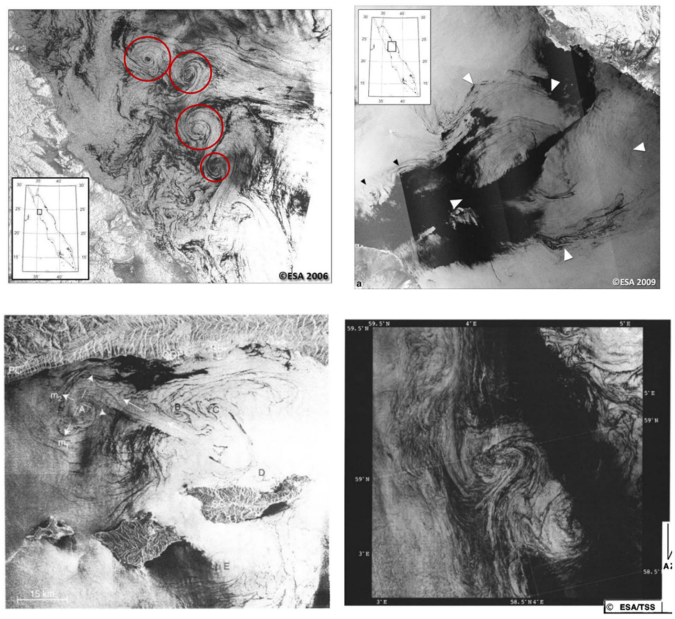



Fig. 1. Eddy morphological pattern on SAR images. The upper left and upper right images are from [53] Figs. 18.2 and 18.11. The bottom left image is from [52] Fig. 5. The bottom right image is from [22] Fig. 5(a). Only for the upper left sub-figure we add red circles for eddy notation, and all other sub-figures remain unchanged as they are already self-explained.

for the ground truth. On the other hand, there are some scientists having been dedicated to other data sources. For example, DeepEddy [4], [5] took SAR images as the study objective and applied PCANet [50], spatial pyramid pooling model [27], and supporting vector machine classifier in succession to distinguish eddies. Xia et al. [6] combined edge information fusion and multiscale detection strategy and obtained a good result on Sentinel-1 products. Also based on SAR images, Yan et al. [7] used ResNet-50 and atrous spatial pyramid pooling [27] to classify five oceanic phenomena, including eddy, rain cell, ship wake, front, and oil spill. These works primarily targeted classification problem on manually preselected patches rather than in an end-to-end manner, but still being inspirational for future research.

## III. Materials and Dataset

### A. Eddy Pattern on SAR Images

Oceanic eddy expression on SAR imagery strongly depends on the wind speeds and is best visualized when winds are between 2 and 7 m/s [22]. Under moderate wind speeds between 3 and 5 m/s, natural films with surfactants dampen small waves, leading to reduced radar backscattering from the sea surface [51]. In this case, eddies are often recognized as dark, narrow, curvilinear, concentric bands (slicks) that appear to spiral inward. At wind speeds of 5–7 m/s, surfactant films start to disrupt, and eddies expressions on SAR only result from wave–current interactions. In this circumstance, eddies are typically identified by a narrow band of enhanced brightness, usually associated with current shear [23], [52], [53], [54]. Fig. 1 provides some examples of eddy presentation on SAR images.

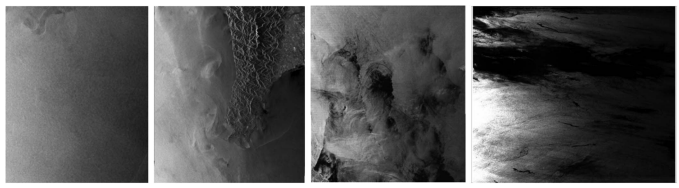


Fig. 2. Data samples. From left to right images are from SAR products: SAR_IMP_20050326_022750, ASA_IMP_20080105_015857, ASA_APP_20071117_134558, ASA_WSS_20110501_022248. (ProductID_startDay_startTime).

### B. Dataset

Materials used in this study include ERS-2 and ENVISAT (A)SAR images of the South China Sea and the Northwest Pacific from 2005 to 2012, although may not continuous in space and time. ERS-2 operated in wave mode or image mode. Under image mode, three Level 1 products were provided: single look complex product, precision product, and medium resolution product. As the successor of the ERS program, ENVISAT has expanded to five operational modes in addition to wave and image mode, including alternating polarization mode, wide swath mode, and global monitoring mode. All these modes provided multiple products based on different processes. This work only selects part of the products that are suitable for eddy observation. Table I briefly introduces these products used [55].

We employ the SNAP [56] tool offered by the European Space Agency to assist in building the dataset. The downloaded SAR products are read and only speckle filtered before being saved as. jpg images. We deliberately keep the manipulations on images as few as possible to facilitate future practical usage. The default filter in SNAP is Lee Sigma with a window size of $7 \times 7$, although we do not think any other filter would make much difference. The only speckle filtering operation imposed on SAR images is for human recognition. Two experts on marine science manually identify eddies and the images containing them are selected for our dataset as raw data. At the same time, the specific eddy area on these images is marked with Labelme [57] software, creating binary images indicating eddy-present or eddy-absent as the ground truth. We have collected a total of 137 images presenting 204 eddies with an average radius of 16.3 km, indicating that SAR images are capable of observing sub- or small-scale oceanic eddies. Fig. 2 provides some SAR image samples of the collected data, which typically have 7–10 K pixels in width or height.

After collection, two common data augmentation techniques, namely flipping and rotating are simultaneously applied to the raw data and the related ground truth. These two transformations should not affect as SAR images are not geographically calibrated at first. However, we do not apply cropping because it would change the contents of images, which goes against our intentions of no extra operation needed when the methodology is deployed. For storage and training time considerations, all images are scaled to 1/4 of the origin, resulting in a width or height of approximately 2–3 K pixels. The downscaling will not affect the authenticity of the dataset because the interrelationship between pixels are not changed during the process.

TABLE I
SAR PRODUCTS USED IN THE STUDY

| Product | Satellite | Mode | Process | Resolution range × azimuth (m) | Coverage range × azimuth (km) | Level |
|---|---|---|---|---|---|---|
| SAR_IMP_1P | ERS-2 | Image | Precision | 12.5 × 12.5 | 100 × at least 102.5 | 1 |
| ASA_IMP_1P | ENVISAT | Image | Precision | 30 × 30 | 56-100 × 100 | 1B |
| ASA_APP_1P | ENVISAT | Alternating Polarization | Precision | 30 × 30 | 56-100 × 100 | 1B |
| ASA_WSS_1P | ENVISAT | Wide Swath | Single Look Complex | 150 × 150 | 406 × 400-4000 | 1B |

Note: "range" refers to the direction of scanning, and "azimuth" is the direction of the satellite track, the two are perpendicular to each other.

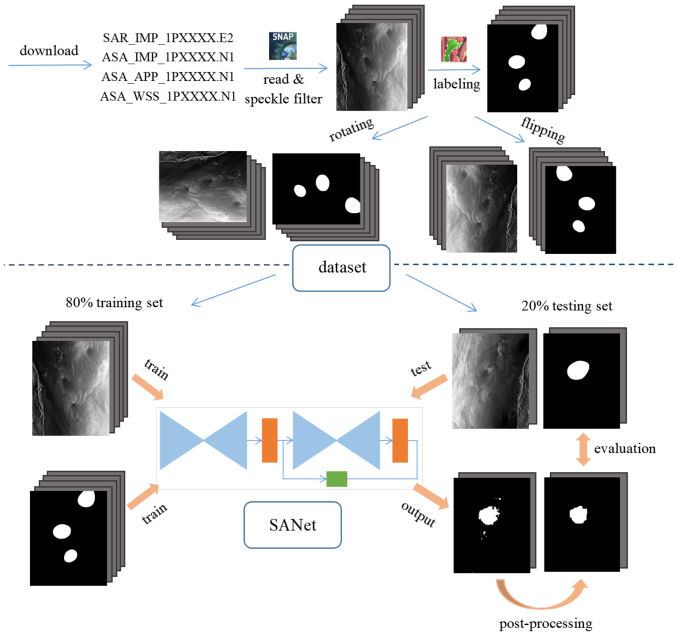



Fig. 3. Overall flowchart of the processing. The top side shows the construction of the SAR dataset, and the bottom side is the end-to-end algorithm for eddy detection.

Eventually, 822 sets of data have been gathered, each consisting of one raw data and one ground truth. This dataset has been made public, given the rarity of exiting eddy datasets based on SAR imagery. With its relative diversity regarding four products from two satellites, we hope it will play a greater role in the field.

In the experiment, the dataset are randomly divided in a ratio of 4:1, of which 80% is the training set for the designed network to learn the implicit function, and the remainder is the testing set to assess the validity of eddy detection. Fig. 3 demonstrates the overall flowchart of the processing.

## IV. METHODOLOGY

### A. Overview of the SANet

Oceanic eddies in SAR images are either presented as dark, narrow, curvilinear, concentric slicks or a narrow band of enhanced brightness in varying sizes [23], [52], [53], [54]. Finding these signatures differs from normal natural image segmentation tasks in two aspects. First, there is no distinct boundary between eddy areas and the other parts, and second, extra care should be taken weighing texture and semantics information. The first issue is inevitable partly from the fact that natural oceanic phenomena do not have clear boundaries themselves, and partly from the obscurity of SAR images. This will negatively affect the detection task, and making it even more difficult and urgent to solve the second problem.

While global semantic information is essential to understand a natural image, it is not that important when seeking eddies in a geographic map. For example, we need the head, neck, body, limbs, and possible surrounding pastures to be able to determine a horse. However, in an remotely sensed SAR images, what other water parts look like has little to do with eddies appearing in this specific area, so local texture, rather than global semantics, is the ultimate key to identifying oceanic eddies. For a deep learning network, the deeper layers often focus on global semantic features and shallower layers on local texture features, thus the rotating presentation of eddies actually lies in the model first few layers. On the other hand, going deeper and deeper is the trend in developing convolutional neural networks mainly because of the more powerful and diverse nonlinear potential, which is also necessary for eddy identification. Thus, the primary challenge of our work is how to utilize shallow features fully and comprehensively while preserving the nonlinearity of deep networks simultaneously.

To achieve the above goal, this article sets up a convolutional neural network architecture of image segmentation customized for eddy detection, named SANet. The very principle of SANet is to stack two independent units, each is relatively shallow in order to extract the rotating texture of eddies, while the overall stacked network is still deep enough to provide the desired mapping ability. To further promote performance, SANet is additionally implemented with internal and external attention mechanisms to highlight the regions of interest that contain eddies.

An overview of SANet is illustrated in Fig. 4. A simple convolutional layer initially extends the image to multiple feature channels. And then these features will go through two stacked units interpolated by an intermediate supervision module. The individual unit is in the hourglass shape for its effective bottom–up, top–down inference. Besides, we incorporate the attention gate within, and a GCblock followed by each hourglass unit to impel the network to focus on the region of interest. Finally, postprocessing is added at the end of the SANet to further refine the eddy detection results. The following will demonstrate the stacked network, the internal attention mechanism (attention gates), the external attention mechanism (GCblock), and the postprocessing in detail.

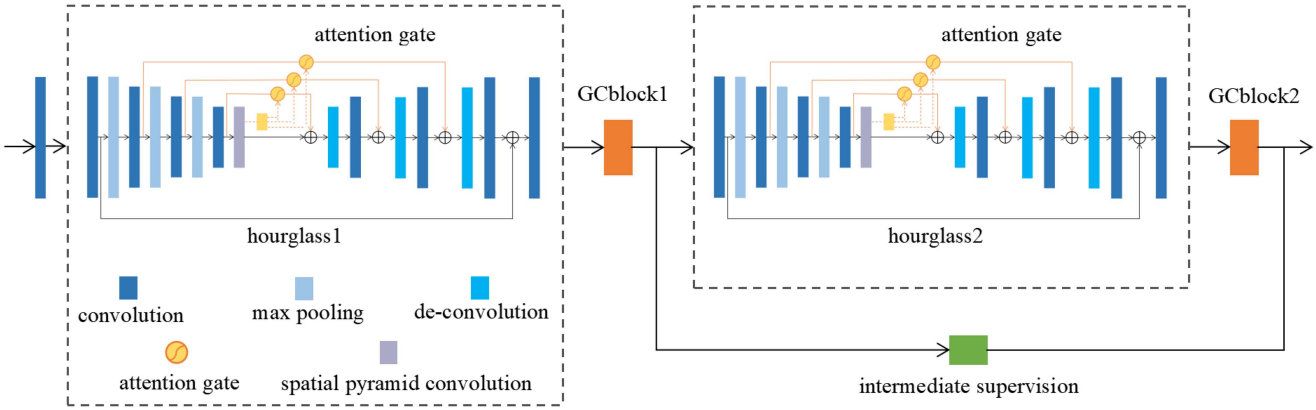



Fig. 4. Overview of the SANet.

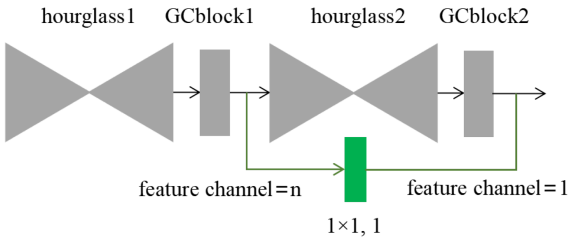



Fig. 5. Intermediate supervision.

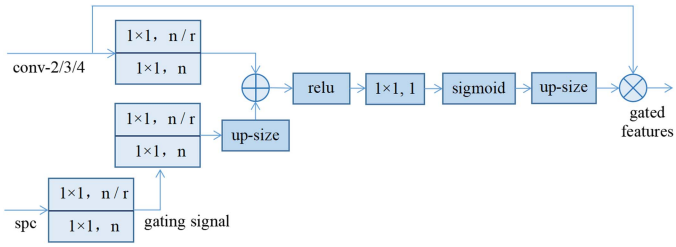



Fig. 6. Attention gate installation.

### B. Stacked Network With Intermediate Supervision

The extraction of eddy textures mainly depends on the first few layers of convolutional neural networks, indicating the discriminative model should not be a very deep one. While one single shallow unit is not enough to capture the complex expression in obscure and noisy radar imagery, stacking two hourglasses together not only deepens the overall network, but also allows the inferences to be drawn double times, so as to refine the features repeatedly.

However, stacked architecture leads to the optimization problem on network training, Newell et al. [58] applied intermediate supervision to settle the matter. We follow their ideas but greatly simplify the procedure considering the heavy computational burden already imposed by the huge radar images. The pruned intermediate supervision is illustrated in Fig. 5, it uses only one convolutional layer to reduce the feature channels after the first hourglass plus GCblock from $n$ to 1. These will be provided as a set of predictions attributing to the final loss, allowing the feedback to be connected directly, rather than back-propagating in sequence tediously.

### C. Hourglass Unit With Attention Gates

The individual unit is the hourglass model that was first proposed for human pose estimation [58]. The name reflects its symmetric encoder–decoder architecture that pools down the input to a low resolution (encoder); and then upsamples and combines features across multiple scales back to the original resolution (decoder). This specific topology effectively captures information at every scale, including local evidence at shallower layers and holistic cognition at deeper layers. The original model reaches the lowest resolution of $4 \times 4$ at the end of the encoder, which is suitable for understanding human poses, but is not proper for our task where local textures are more indispensable. Thus, the modified hourglass unit (dotted box in Fig. 4) only downsamples three times, so that even the lowest resolution remains 1/8 of the input image. Besides, unlike the original version, which employs the same number of feature channels for each layer, we make it 32, 64, 128, and 256 from the input to the end of the encoder, thus considerably easing the computation. Furthermore, given the varying size of eddies, we have added a spatial pyramid convolution [27] after the fourth layer (end of the encoder path), with dilation rates of 6, 12, and 18, respectively, making sure no eddy being missed in the abstraction.

Inspired by "attention U-Net" [59], the attention gate as the hourglass-internal attention mechanism is another improvement of our work. The key is the gating signal, which should constrain the entire representation. As the encoder goes to the deepest layer, the information is the most coherent available to the network, so it is appropriate to transmit the gating signal here (after the spatial pyramid convolution). Multiscale features from the encoder path will be gated by this signal before being concatenated to the decoder path, thus being more focused on the target structure.

The detailed installation of the attention gate is shown in Fig. 6. Since the first scale features (conv-1) usually do not

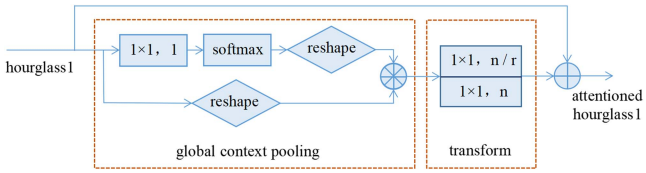



Fig. 7. GCblock installation.

capture enough information, the attention gate is only applied to the second, the third, and the fourth scale features (conv-2, conv-3, and conv-4). One single $1 \times 1$ convolution after the spatial pyramid convolution produces the contextual gating signal. Then, features from the encoder are pruned by aligning them with this gating signal. This operation would provide coefficients which, after some transformations, are multiplied by the original features element-wisely to obtain the gated features that identify salient regions. The consecutive $1 \times 1$ convolutions with channels of $n/r$ and $n$ ($n$ equals the channel number of the input features and $r$ is a specified ratio) perform the nonlinear transformation in a parameter-reduction manner, and this special design is also used in the GCblock and will be further illustrated in the next section.

### D. GCblock as the External Attention

To further highlight the area containing eddy textures and suppress irrelevant backgrounds, SANet apply Global Context block (GCblock) [60], [61], [62] as the hourglass unit-external attention mechanism. GCblock integrates both the pixelwise and channelwise feature attentions, yet with only lightweight parameters, thus can be flexibly plugged into any network. As shown in Fig. 7, the GCblock consists of a global context pooling component, a transform component and a broadcast additive fusion. The global context pooling takes as input the first hourglass's outcome. The $1 \times 1$ convolution and softmax activation are used to obtain the global context attention map, and by matrix-multiplying it with the previous features, the global context is modeled. Then, the transform component captures the channelwise dependencies with consecutive $1 \times 1$ convolutions of $n/r$- and $n$-channels, respectively ($n$ equals the channel number of the global context pooling output and $r$ is a specified ratio). Finally, these enhanced transformation will be added back on the basic features, giving emphasis to the region of interest.

The same strategy of transformation is also deployed in the hourglass-internal attention gates. A single $n$-channel convolution may achieve a similar goal, but would require a vast number of parameters ($n^2$). However, replacing it with two consecutive $n/r$- and $n$-channel ones will significantly reduce the parameters from $n^2$ to $2n^2/r$. Our work's default setting of $r$ is 4, which should cut down half of the parameters, saving approximately 2/5 of the training time.

### E. Postprocessing

The direct output of SANet is a binary image, where the gray value 1 is rendered white, indicating the eddy area, and the gray value 0 is rendered black, indicating the ocean background. In these results, however, the identified eddy areas are uneven with burrs, bumps, and dips on the boundary, and hollows in the interior. Also, the background is not clean with some white dots appearing there and here. This is due to the opaque nature of SAR imagery, the indistinct eddy features, and of course, the inevitable incompetence of the deep learning model itself. Therefore, a postprocessing is needed to congregate the eddy presence and eradicate the noise on the massive background.

The well-established image analysis method, namely, mathematical morphology, is applied as the postprocessing. Specifically, the Open and Close operation is imported and imposed successively to the direct output of the network. The Open operation is defined as first Dilating and then Eroding, and the Close operation is the opposite, first Eroding and then Dilating, both based on the interrelationship of pixels in the binary image [63]. Simply put, the Open operation will remove the individual white dots on the background and on the eddy boundaries. After that, the Close operation is applied to fill in the hollows and to cluster these potentially separate parts that actually belong to one eddy. The Open–Close order is determined because the direct output of the network is generally congregated, and those individual or small scattered dots outside the eddy region must be removed first, otherwise large portions of the image might be catastrophically permeated and connected all together.

The detailed procedures for these operations will not be elaborate redundantly since they have been maturely integrated into programming frameworks such as Python and MATLAB, where a simple built-in function could do the job. The basic formula is as follows, in which $B$ is the squared kernel that scans the original image $A$, and the sizes of $B$ are empirically set to 80 $\times$ 80 for both Open and Close operations

$$\begin{aligned} A \circ B &= (A \ominus B) \oplus B \\ A \bullet B &= (A \oplus B) \ominus B \end{aligned} \tag{1}$$

where $\circ$ is the Open operation, $\bullet$ is the Close operation, $\ominus$ is for Dilating operation, and $\oplus$ is for Eroding operation.

### F. Loss Matrix

The loss function is one of the most decisive factors in deep learning. After several trials, we found the mixture of dice and focal loss is proper for our task. Dice loss is commonly used in image segmentation. It is the opposite of the dice coefficients, which measures the similarity between two volumes. And focal loss [45] is explicitly proposed to address the class imbalance problem, which often leads to degenerate models dominated by negative backgrounds. The loss function used in the experiment is as follows:

$$p_i^* = \begin{cases} p_i & \text{if } \ q_i = 1 \\ 1 - p_i & \text{otherwise} \end{cases} \qquad F = -\sum_i^N ((1 - P_i^*)^\gamma \log(p_i^*)) \tag{2}$$

$$D_c = \frac{2\sum_i^N (p_i q_i)}{\sum_i^N p_i^2 + \sum_i^N q_i^2} \qquad D = -\log(D_c) \tag{3}$$

$$L = \alpha \cdot F + D. \tag{4}$$

In the above, $p$ is the network's output and $q$ specifies the ground truth, both have the same dimension of $N$ pixels, with each denoted by $i$. The focal loss $F$ is defined by adding a modulating factor $(1-p_i^*)^\gamma$ ($\gamma$ is the focusing parameter) on the basis of the normal cross entropy $\log(p_i^*)$, thus to up-weight the misclassified samples. $D_c$ represents the dice coefficients calculated with two times the intersection between $p$ and $q$ divided by the sum of their respective elements. The actual dice loss $D$ is the opposite of $\log(D_c)$, in which the logarithm is used to balance the final loss combination $L$, with also the help of the balanced variant $\alpha$.

## V. Experiments and Results

### A. Implementation Details

This article aims to build an end-to-end deep learning framework that performs minimal pre- or postprocessing on SAR images, so that little effort is required when the method is implemented. Given this, raw images of the training set are directly fed into the presented network without cutting or extra enhancing operations. Considering that huge SAR images would consume a bunch of memory, thus training by stochastic gradient descent is the best choice under the condition (batch size=1). Furthermore, the Adam optimizer is also used for stabilizing parameter updating. Using the loss function defined in Section IV, the training process begins with the initial learning rate of 1e-5, decreases to 1e-6 after 200 epochs, and further drops to 1e-7 after 300 epochs. The loss value stops falling after 350 epochs, which completes the training process. Before the images are sent to the network, normalization on gray values is performed by scaling the original range 0–255 to 0–1 to accelerate convergence of the network, and scaled back to 0–255 before output. There is no uniformity on the data size since SANet is a fully convolutional network capable of inputting images of arbitrary size. The direct output will be postprocessed with the kernel size for both Open and Close operations empirically set to $80 \times 80$, producing the final result that could be examined compared to the ground truth.

The experiments are implemented on the TensorFlow platform, and the hardware includes an NVIDIA DGX station with one 20-core CPU and four Tesla V100 GPU.

### B. Objective Metrics

Since the proposed method achieves pixel-wisely eddy detection, the evaluation should first be performed on the pixel level. Moreover, the ultimate goal of this work is to correctly identify eddy as a whole instead of as individual pixels, thus the evaluation on the target level is also essential.

First, for pixel level evaluation, accuracy, precision, recall, Intersection over Union (IoU), Dice coefficients are used, most of which could be calculated through the confusion matrix consisting of true positive, false positive, true negative, and false negative. Fig. 8 helps to understand their meanings.

1) True positive (tp): when pixel $i$ on the output and the ground truth both indicate eddy-present (the green part in Fig. 8);

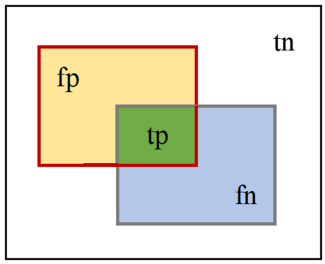



Fig. 8. Confusion matrix. The outer black box represents the entire image, the top-left red box is the network-determining eddy area, and the bottom-right gray box is the actual eddy area, also known as the ground truth.

2) False positive (fp): when pixel $i$ on the output indicate eddy-present but the ground truth actually eddy-absent (the yellow part in Fig. 8);
3) True negative (tn): when pixel $i$ on the output and the ground truth both indicate eddy-absent (the white part in Fig. 8);
4) False negative (fn): when pixel $i$ on the output indicate eddy-absent but the ground truth actually eddy-present (the blue part in Fig. 8);

$$\text{accuracy} = \frac{\text{tp}+\text{tn}}{\text{tp}+\text{fp}+\text{tn}+\text{fn}}. \tag{5}$$

Accuracy is the ratio of all correctly classified pixels, whether eddy-present or eddy-absent, to the total number of pixels in the image.

$$\text{precision} = \frac{\text{tp}}{\text{tp}+\text{fp}}. \tag{6}$$

Precision is with regard to all model-determining eddy-present pixels, the ratio of the actual eddy-present pixels

$$\text{recall} = \frac{\text{tp}}{\text{tp}+\text{fn}}. \tag{7}$$

Recall is regarding all actual eddy-present pixels, the ratio of the correctly identified pixels

$$\text{IoU} = \frac{\text{tp}}{\text{tp}+\text{fp}+\text{fn}}. \tag{8}$$

IoU is the intersection of the model-determining eddy areas and the actual eddy areas to the union of them

$$\begin{aligned}\text{dice coefficient} &= \frac{2\text{tp}}{2\text{tp}+\text{fp}+\text{fn}} = D_c \\ &= \text{F1 score} = 2/\left(\frac{1}{\text{precision}}+\frac{1}{\text{recall}}\right).\end{aligned} \tag{9}$$

Dice coefficient is the same as $D_c$ defined in Section IV, calculated by two times the intersection of the model output and the ground truth divided by the sum of their respective pixels. In addition, algebraically, Dice coefficient is equal to F1 score, an index to balance precision and recall.

In terms of the evaluation on the target level, recognition rate and false alarm rate are used. The Dice coefficient is applied as

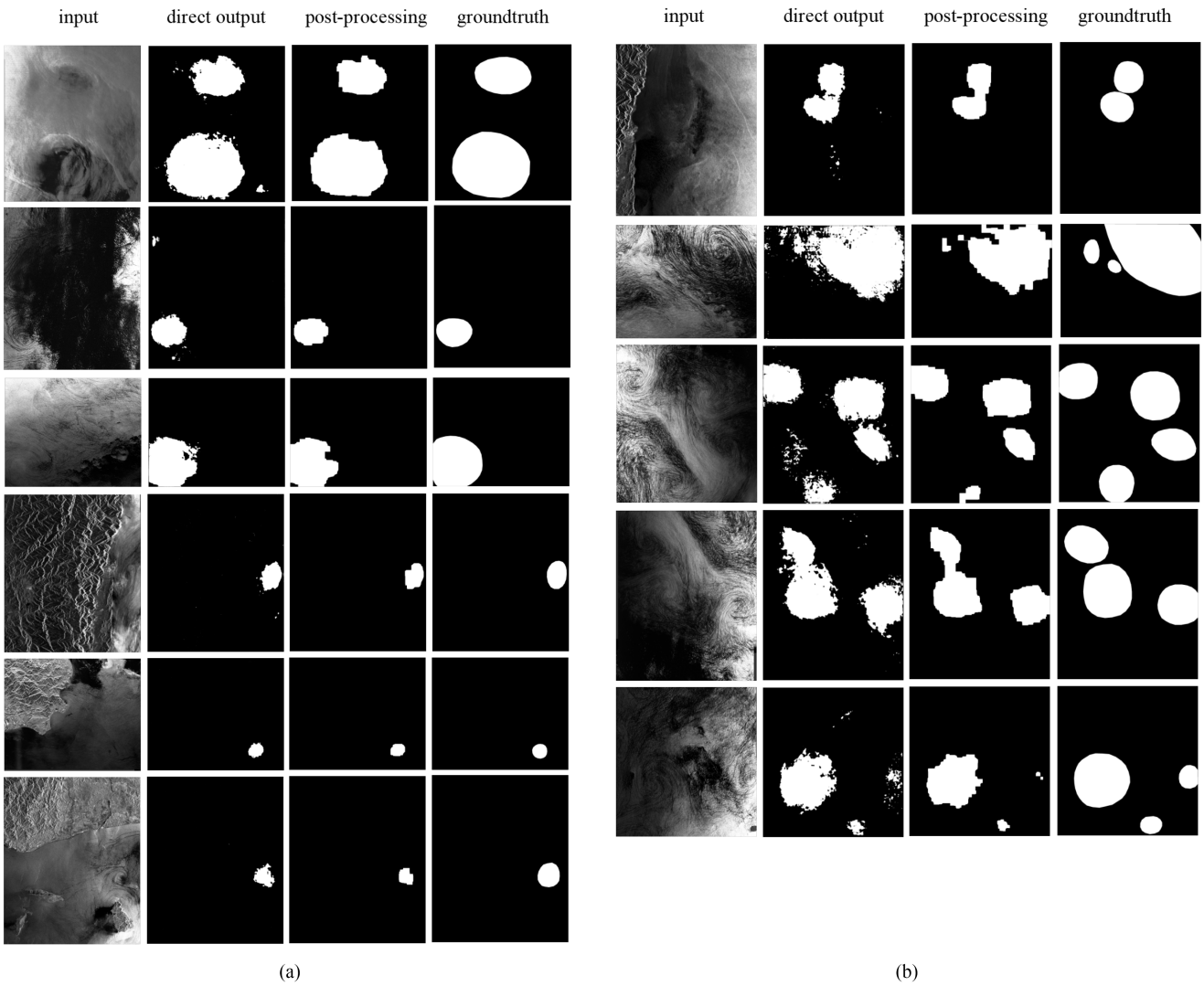



Fig. 9. Eddy detection results. From the left column to the right column, top to bottom, images are from SAR products: (ProductID_startDay_startTime) ASA_APP_20070218_014816, SAR_IMP_20090819_022317, SAR_IMP_20101013_022407, ASA_IMP_20080628_015827, SAR_IMP_20050829_022545, SAR_IMP_20070502_022249, SAR_IMP_20080225_022513, SAR_IMP_20090522_235858, SAR_IMP_20090607_235558, SAR_IMP_20090607_235554, and ASA_IMP_20080711_014929.

the decisive indicator for eddy recognition, whichever is greater than 0.5 and considered the related eddy to be identified. Under this premise, recognition rate is computed by the number of successfully detected eddies divided by the total number of eddies in the ground truth. And false alarm rate is the ratio of the number of wrongly identified signals that are actually not eddies, to the total number. Just note that when more than one eddy regions are connected together, all of them are considered correctly identified as long as the overall Dice coefficient exceeds 0.5.

### C. Eddy Detection Results

Eddy detection is performed on the testing set, which is retained aside when SANet is being trained. Fig. 9 shows the direct output and postprocessed final results for some samples, along with their associated input raw images and ground truths.

Since we see eddy detection as a semantic segmentation task that assigns each pixel of the picture into a category not only could we identify the presence of eddy, but also outline the full profile. When the raw SAR image is simple and monotonous, as in the six rows on the left in Fig. 9, the direct output of SANet is already relatively complete with legible centroids, despite some burrs appearing on the eddy boundaries and unexpected noise scattered in the background. After postprocessing, the result is better and more concrete in shape so that the eddy radius can be measured accordingly. However, when the raw image is more complex, such as with two or more eddies presenting or covered by some curves and lines of the sea surface, as shown in the five rows on the right of Fig. 9. In addition to the abovementioned problems, the detected contours may be inseparable between adjacent eddies even after postprocessing. Fortunately, even so, we can still determine the existence of eddies and their approximate locations.

TABLE II
OBJECTIVE COMPARISON OF ABLATION EXPERIMENTS

| Network | Accuracy ↑ | Precision ↑ | Recall ↑ | IoU ↑ | Dice ↑ | False alarm rate ↓ | Recognition rate ↑ |
|---|---|---|---|---|---|---|---|
| SAN: 2 hourglasses (each 4 scales) + attention gating + GCblock | **97.49** | **85.78** | **50.40** | **48.30** | **61.23** | **1.96** | **87.75** |
| 2 hourglasses (each 4 scales) + GCblock | 97.36 | 73.75 | 45.70 | 42.85 | 54.38 | 9.31 | 64.71 |
| 2 hourglasses (each 4 scales) | 96.57 | 64.48 | 34.99 | 32.20 | 42.89 | 16.67 | 54.41 |
| 2 hourglasses (each 5 scales) | 96.25 | 60.21 | 23.01 | 21.92 | 30.64 | 8.33 | 50.49 |
| 1 hourglass (8 scales) / U-net | 96.39 | 61.43 | 33.77 | 29.72 | 40.17 | 26.47 | 48.04 |

Note: all data are in %. Arrow following each indicator points toward its betterment.
The bold values indicate the best performance.

### D. Ablation Experiment

The special topology we design is not without reason. As said before, the repeated bottom–up, top–down hourglasses are for iterative extractions of shallow-layer features. The attention gates within and the GCblock followed by the hourglass units are employed to further constrain the expression. In order to testify tenable of each component, we set up several ablation experiments with the gadgets removed and compare their abilities of eddy detection.

Table II lists the objective comparison of the ablation experiments. The first row is the proposed SANet stacked by two hourglasses, each comprising the gating mechanism, and followed by the GCblock, as well demonstrated in Section IV. The extremely high Accuracy seems promising, but it mainly is contributed by the dominant backgrounds, or true-negatives in SAR images. IoU is acceptable, although slightly lower than the most advancing paper in computer vision, typically lying in 60%–70%, considering our work is based on only a few hundred obscure radar images that are utterly unparalleled to the standard competition datasets comprising hundreds of thousands of high-quality images with well-defined boundaries. Precision, recall, and their combination Dice coefficient determine whether an eddy is recognized by the network. With the cut off value of 0.5 for Dice coefficient announced before, the proposed SANet achieves the recognition rate of 87.75%, while only 1.96% false alarms, which we believe is an encouraging result.

Removing the attention gates within the hourglass unit (the second row in Table II) will decrease performance, and further removal of the GCblock (the third row in Table II) damages capability even more significantly, with the recognition rate dropping to only 54.41%. In addition, adding more layers to each hourglass (the fourth row in Table II) does not bring any improvement, verifying the reasonable adaptation to 4-scale for this specific task. Last but not least, in order to demonstrate the necessity of the stacking structure as the most powerful element for extracting eddy texture, we experiment on the one-hourglass model with eight scales (the last row in Table II), so as to have approximately the same total depth of layers and volume of parameters as in the replaced two-hourglass stacked model. In fact, since the modified hourglass illustrated in Section IV chooses to double the feature channels as layer deepens, instead of keeping the constant feature channels as in the original hourglass version, we found the one-hourglass model coincidentally share a similar architecture to the well-known and widely used network U-net [30]. However, the classic U-net only identifies less than half of the eddies, much lower than our SANet, thus proving the absolute superiority of the proposed stacking structure for eddy detection in SAR images.

In addition, to better interpret each component's role in the SANet, a feature visualization experiment is performed. Fig. 10 displays the intermediate outputs of some critical nodes as two data samples go through the SANet model. The left column images are features at the encoder–decoder junction within the hourglass unit that should recognize the overall structure but are obscure in textures. After recovering by the decoder path of the hourglass, the middle column images considerably reconstruct the fine-grained texture information. Moreover, compared to the right column images when the results are enhanced by the attention mechanism, where regions related to eddies are highlighted, it is sufficient to prove the functionality of the GCblock. Besides that, the most important insight is that the spatially two-stack architecture does improve the feature representation when comparing the output of the first and second hourglasses, confirming the crucial role of the stacking design.

### E. Comparison With Other Methods

An intuitive way to show the superiority of our method is to compare it with other classic oceanic eddy identification methods. Unfortunately, however, the widely accepted approaches based on physical oceanography are mainly applicable to SSH/SLA or flow field data (which have their own intrinsic drawbacks, as discussed in Section II) and not to SAR images. As for some other methods that benefit from the deep learning community, such as DeepEddy [4], [5], Xia et al. [6], Yan et al. [7], all of them first crop the whole SAR image into patches and select those containing eddies, and then classify the specific patch or frame out the target with a box, rather than performing an end-to-end detection on the whole SAR image, nor at the pixel-level. Therefore, none of these approaches are suitable for comparison with our method.

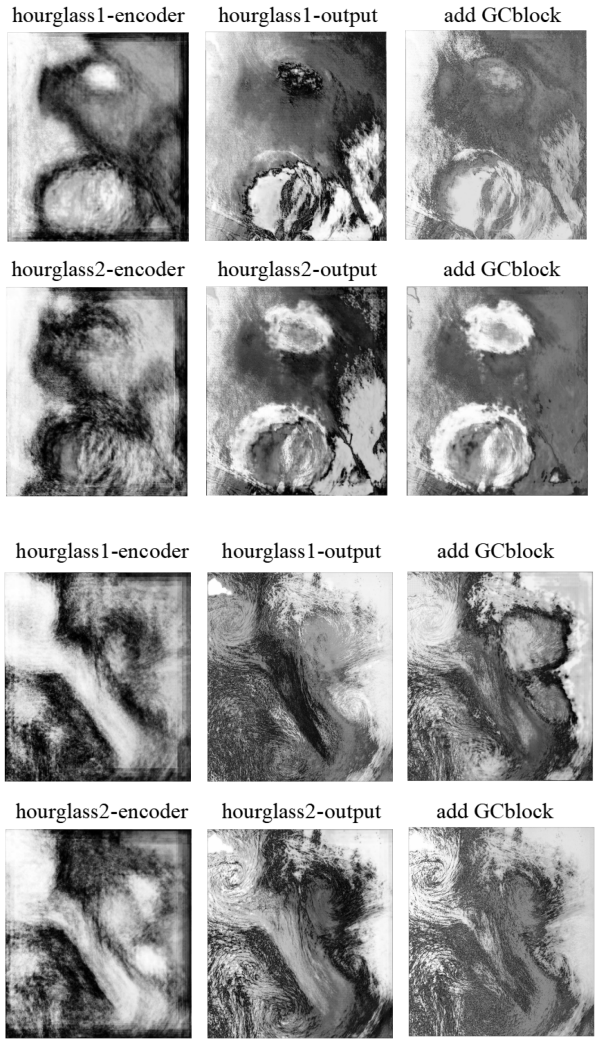



Fig. 10. Feature visualization. The first two rows and the last two rows are intermediate outputs of image corresponding to SAR product: ASA_APP_20070218_014816, SAR_IMP_20090607_235558, (ProductID_startDay_startTime). For each product, the subfigures from top to bottom, left to right are intermediate outputs of: the encoder–decoder junction inside the first/second hourglass, the endpoint of the first/second hourglass, after the first/second GCblock. Note that for alignment, the figures does not reflect the actual size relationship between these intermediate outputs, and that the intermediate outputs of the inner-hourglass attention gates cannot be generated because these gates are distributed and intertwined in every scale in the hourglass unit.

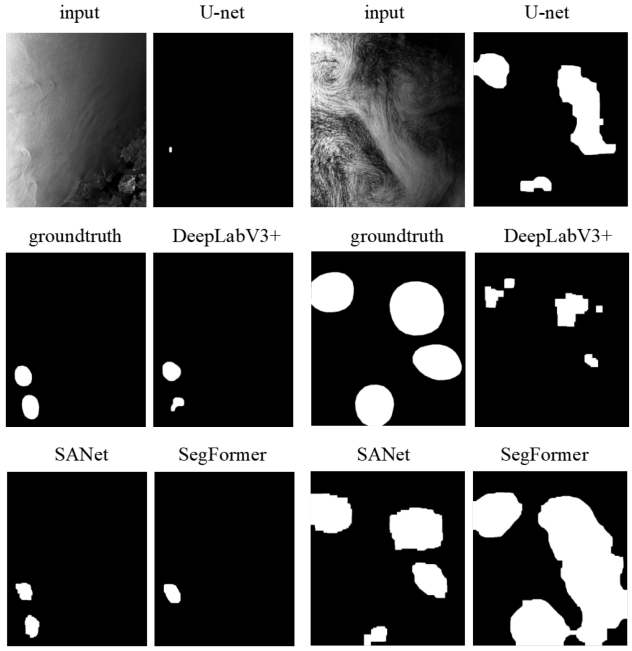



Fig. 11. Visual comparison with other deep learning networks. The left two columns and the right two columns corresponds to SAR product: SAR_IMP_20051202 and SAR_IMP_20090607 (ProductID_startDay_startTime).

However, in terms of the capability of the proposed SANet that belongs to the semantic segmentation genre, we can compare it with other existing state-of-the-art networks with respect to solving the problem of oceanic eddy identification. Since some authors [35], [36], [37] have applied U-net [30] as the backbone of their work (although implemented on SSH/SLA data), and considering its potential connection with SANet as discussed in the previous section, we decided to include U-net [30] for comparison. U-net was the first to propose the U-shaped encoder–decoder architecture combined with skip connections that has the advantage of comprehending the overall context while recovering fine-grained information, and is credited as the foundation of many later models in the field of semantic segmentation. Another network we draw attention to is the latest version of the DeepLab series, DeepLabV3+[33], developed by the Google Team. DeepLabV3+ is a relatively new high achiever in semantic segmentation tasks, exploring the atrous separable convolution and atrous spatial pyramid pooling on top of previous models, and is commonly recognized as a benchmark in many applications because of its robustness and efficacy. Besides, have been noticing the promising performance of Transformer in the Computer Vision field recently, the SegFormer [64] model is also included as one of the competitors. Transformer-based model differs from the aforementioned convolution-based models (including the proposed SANet) in its heavy reliance on the self-attention mechanism, which facilitates the acquisition of global information at the outset, rather than accumulating features through multiple layers of convolution computation.

Fig. 11 shows two randomly selected samples and their corresponding detection results for U-net, DeeplabV3+, SegFormer, and the proposed SANet. For the three convolution-based models, it is apparent that SANet outperforms U-net and DeeplabV3+, as the latter two either fail to detect the target or the detected regions are far from the ground truth in shapes and sizes. As for the SegFormer, the good news is that the direct output is much smoother on eddy boundaries and few scattered white dots appear outside the eddy region, thus the postprocessing seems unnecessary (although it is also performed for a fair comparison). However, in terms of the practical recognition ability, it is polarized for different samples, with some being precisely identified and others not recognized at all, as you can see in the first sample displayed in Fig. 11. Another drawback is that it tends to permeate large areas when multiple eddies are present in one image. Nevertheless, the transformer-based

TABLE III
OBJECTIVE COMPARISON WITH OTHER DEEP LEARNING NETWORKS

| Network | Accuracy ↑ | Precision ↑ | Recall ↑ | IoU ↑ | Dice ↑ | False alarm rate ↓ | Recognition rate ↑ |
|---|---|---|---|---|---|---|---|
| U-net | 96.39 | 61.43 | 33.77 | 29.72 | 40.17 | 26.47 | 48.04 |
| DeepLabV3+ | 96.74 | 67.51 | 39.02 | 38.01 | 49.70 | 14.22 | 66.85 |
| SegFormer | 97.40 | 72.34 | **51.53** | 44.41 | 55.80 | 4.90 | 69.61 |
| SANet | **97.49** | **85.78** | 50.40 | **48.30** | **61.23** | **1.96** | **87.75** |

Note: Results of U-net and SANet are the same as in Table II; All data are in %.
The bold values indicate the best performance.

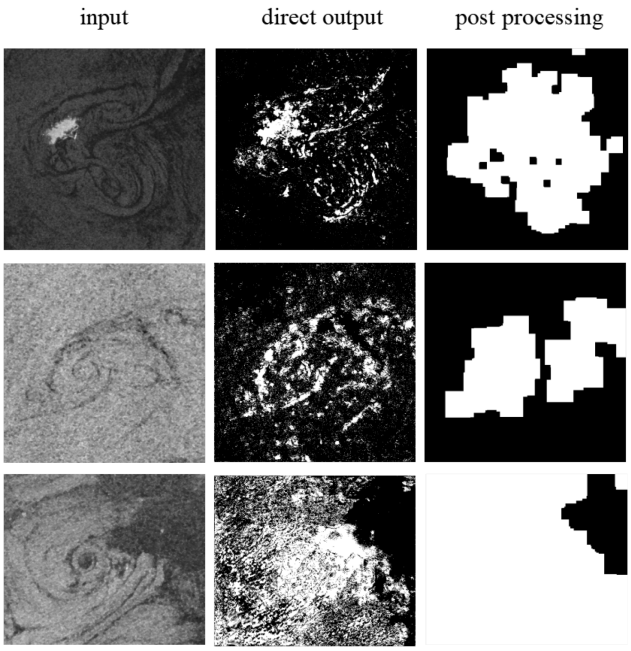



Fig. 12. Detection of Sentinel-1 SAR images.

method deserves more exploration. Table III lists the objective results of the four methods, in which SANet has an absolute advantage in almost all objective metrics, except for Recall on pixel-level, with SegFormer slightly higher, mainly because of its tendency to saturate large areas that may not even be eddies, as we explained above.

### F. Generalization Ability

Generalization ability refers to the adaptability of deep learning algorithms to new samples that are distant from the training set but share inherent commonalities. A few images from Xia et al. [6] derived from the Sentinel-1 satellite are leveraged for generalization testing. Naturally, Sentinel-1 differs from ERS-2/ENVISAT in many ways, such as the operation of the mission, the resolution and swath of the products, and most importantly, the signal-to-noise ratio of the image, which will severely affect the performance of the model trained exclusively on our dataset. As can be seen from the middle column in Fig. 12, the direct outputs of SANet for Sentinel-1 images are obviously not as good as for our dataset. However, what we appreciate is that it does highlight the eddy area, although poorly concentrated, mainly due to various noise distributions caused by instruments' differences. After postprocessing, the results are expected to be more concrete. Note that the postprocessing is slightly different from that in all other experiments mentioned in this article. Here, we choose the order of Close–Open operation since the direct output of the model is no longer congregated but rather scattered in this situation. In addition, the kernel size is changed to 30 × 30 for Close operation and 60 × 60 for Open operation.

The important thing is that the proposed SANet has no restrictions on the size or resolution of the input, so it can be developed successively or simply start from scratch with other source SAR images. Thus, we are confident that the network's performance will be further improved if more varieties of SAR images are available, and that the model will be more accommodating for Sentinel-1 or other SARs if these products are used for training from the beginning.

## VI. DISCUSSION

### A. Overfitting

Constructing SAR image datasets for eddy detection is a time-consuming and arduous work; on the one hand, because of the large size of SAR products and, therefore, the slow downloading and reading. On the other hand due to the stringent observation conditions, such as sea surface wind speed and surface oil film accumulation, and finally because of the limitations on the temporal and spatial resolution of the satellite, leading to possible missing of the eddy occurrence, relative to observation coverage. Regardless of the reason, it is justified to consider the potential risk of overfitting problem arising from the limited number of data samples. However, as this article sees eddy detection as an image segmentation task where each pixel is required to be assigned into a category, the number of data samples should be determined not only by the number of images, but also by the number of pixels in these images, and that is quite a lot! As we have described in Section III, the typical width of the input images has 2–3 K pixels in just one dimension, and exponential to that in the whole picture. Under this circumstances, we do not have to think much about the issue of overfitting, which is one of the advantages of treating it as an image segmentation problem.

### B. Plain Samples

In relation to the first discussion, since the dataset is not so rich, should we add more plain samples which have no eddies

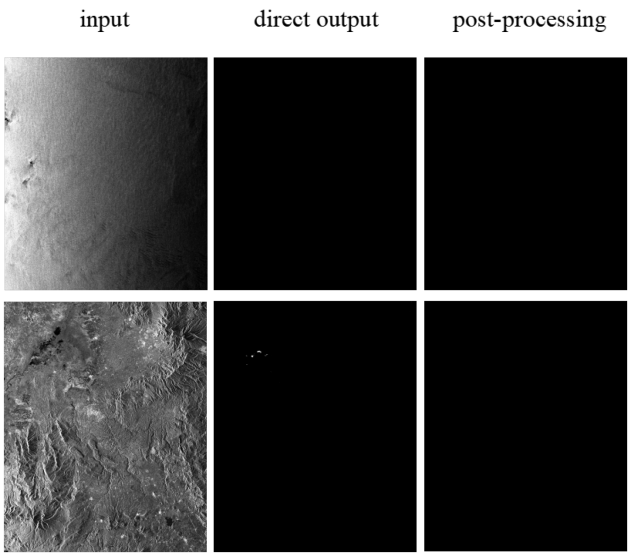



Fig. 13. Plain sample examination. The two SAR products are SAR_IMP_20050102_023557 and SAR_IMP_20050104_141535 (ProductID_startDay_startTime) for the smooth sea surface and for land, respectively.

in them? The answer is also in relation to the first consideration, that the data samples we are talking about are actually the product of the number of images and the number of pixels in images. If that is the case, we already have, and have much more plain samples than eddy samples, that we have had to bring in focal loss as mentioned in Section IV to compensate for this imbalance, instead of including more "fully plain sample" images. However, it is reasonable to include fully plain samples in the testing dataset and examine whether the output is empty. The experiment is conducted on two SAR images, one with a smooth sea surface and the other orienting to land. The results are displayed in Fig. 13. For the smooth sea surface, the direct output is all clean. And for the land area, the direct output presents some noise in the upper left of the image, but after the postprocessing, the final result is as expected.

### C. Future Improvements

During the study, we also realize some areas for improvement. First, since the original SAR images have not been geographically calibrated, and furthermore, we have rotated images for data augmentation. The consequence is that we can not tell whether the actual rotating direction is counterclockwise or clockwise, and thus to determine the type (cyclonic or anticyclonic) of oceanic eddies. This requires a more specific consideration regarding eddy attribution in our next step. The other crucial defect of this work is the small-scaled dataset, which needs to be expanded and enriched in terms of number and satellite sources. Finally, this work sees eddy detection as an image segmentation task performed only on the intensity channel of SAR products, but ignores many of their unique characteristics, which seems a big loss and should be reconsidered fully and completely in future work.

## VII. Conclusion

This article devotes to establishing an end-to-end pipeline from downloaded SAR images to the final eddy identification results that avoids all the hustle and bustle in the middle. For this purpose, a deep learning model is proposed, namely, SANet, customized for oceanic eddy detection on SAR images. As the name suggests, the constructed model is a stacking structure with additional functions providing attention mechanisms. The individual component for stacking is the hourglass unit, a symmetric encoder–decoder that first shrinks the image to acquire semantic signals and then expands it to restore the detailed information. Unlike most deep learning-applying cases where the holistic cognition of the whole image is more important, eddies' superficial texture contained in a model's shallow layers is the ultimate objective of our task. This gives a hint to the innovative usage of stacking two hourglass units together, thus allowing repeated refinement of shallow layers' embedded messages. Besides, SANet has included attention mechanisms for the network to be more concentrated on the eddy area. One of them is the attention gate installed within the hourglass unit, thus constraining the features extracted by its attached hourglass. The other is the GCblock followed by each hourglass, ensuring the validity of the overall abstraction. Also, for the feasibility of the model training, a trick called intermediate supervision is applied after the first hourglass, so that the back-propagating gradients will never vanish even more hourglasses are stacked if needed. The only postprocessing applied after SANet is based on mathematical morphology, which will congregate the eddy region and remove the noise of false alarms.

The proposed method achieves a recognition rate of 87.75% on the established dataset, which is higher than those similar models but without attention gates or GCblock, thus validating the reliability of attention mechanisms. More importantly, the stacked network is way better than its one-hourglass counterpart U-net, which only identifies less than half of eddies, thus confirming the efficacy of the stacked architecture. The same conclusion is also consolidated by the feature visualization of some intermediate outputs of the network. In addition, SANet has exceeded some other state-of-the-art deep learning models such as DeepLabV3+ and SegFormer by a large margin. Finally, a further generalization tests is conducted, verifying its adaptability to some extent.

## References

[1] D. B. Chelton, M. G. Schlax, R. M. Samelson, and R. D. Szoeke, "Global observations of large oceanic eddies," *Geophys. Res. Lett.*, vol. 34, 2007, Art. no. L15606, doi: 10.1029/2007GL030812.

[2] D. B. Chelton, M. G. Schlax, and R. M. Samelson, "Global observations of nonlinear mesoscale eddies," *Prog. Oceanogr.*, vol. 91, pp. 167–216, 2001, doi: 10.1016/j.pocean.2011.01.002.

[3] D. D'Alimonte, "Detection of mesoscale eddy-related structures through Iso-SST patterns," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 189–193, Apr. 2009, doi: 10.1109/LGRS.2008.2009550.

[4] Y. Du, W. Song, Q. He, D. Huang, A. Liotta, and C. Su, "Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection," *Inf. Fusion*, vol. 49, pp. 89–99, 2019, doi: 10.1016/j.inffus.2018.09.006.

[5] D. Huang, Y. Du, Q. He, W. Song, and A. Liotta, "DeepEddy: A simple deep architecture for mesoscale oceanic eddy detection in SAR images," in *Proc. IEEE 14th Int. Conf. Netw., Sens. Control*, Calabria, Italy, 2017, pp. 673–678, doi: 10.1109/ICNSC.2017.8000171.

[6] L. Xia, G. Chen, X. Chen, L. Ge, and B. Huang, "Submesoscale oceanic eddy detection in SAR images using context and edge association network," *Front. Mar. Sci.*, vol. 9, pp. 1023624, 2022, doi: 10.3389/fmars.2022.1023624.

[7] Z. Yan, J. Chong, Y. Zhao, K. Sun, Y. Wang, and Y. Li, "Multifeature fusion neural network for oceanic phenomena detection in SAR images," *Sensors*, vol. 20, 2020, Art. no. 210, doi: 10.3390/s20010210.

[8] P. M. Saunders, "Anticyclonic eddies formed from shoreward meanders of the gulf stream," *Deep Sea Res. Oceanographic Abstr.*, vol. 18, pp. 1207–1219, 1971, doi: 10.1016/0011-7471(71)90027-1.

[9] S. H. Peckinpaugh and R. J. Holyer, "Circle detection for extracting eddy size and position from satellite imagery of the ocean," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 2, pp. 267–273, Mar. 1994, doi: 10.1109/36.295041.

[10] A. Fernandes and S. Nascimento, "Automatic water eddy detection in SST maps using random ellipse fitting and vectorial fields for image segmentation," in *Proc. Int. Conf. Discov. Sci.*, Berlin, Germany, 2006, pp. 77–88, doi: 10.1007/11893318_11.

[11] J. J. Oram, J. C. Mcwilliams, and K. D. Stolzenbach, "Gradient-based edge detection and feature classification of sea-surface images of the Southern California bight," *Remote Sens. Environ.*, vol. 112, pp. 2397–2415, 2008, doi: 10.1016/j.rse.2007.11.010.

[12] B. Lemonnier, C. Lopez, E. Duporte, and R. Delmas, "Multiscale analysis of shapes applied to thermal infrared sea surface images," in *Proc. Int. Geosci. Remote Sens. Symp.*, 1994, pp. 479–481, doi: 10.1109/IGARSS.1994.399158.

[13] A. Okubo, "Horizontal dispersion of floatable particles in the vicinity of velocity singularities such as convergences," *Deep Sea Res. Oceanographic Abstr.*, vol. 17, pp. 445–454, 1970, doi: 10.1016/0011-7471(70)90059-8.

[14] J. Weiss, "The dynamics of enstrophy transfer in two-dimensional hydrodynamics," *Physica D: Nonlinear Phenomena*, vol. 48, pp. 273–294, 1991, doi: 10.1016/0167-2789(91)90088-Q.

[15] J. Isern-Fontanet, E. Garc'Ia-Ladona, and J. Font, doi: 10.1016/j.pocean.2008.10.013.

[16] A. Chaigneau, A. Gizolme, and C. Grados, "Mesoscale eddies off Peru in altimeter records: Identification algorithms and eddy spatio-temporal patterns," *Prog. Oceanogr.*, vol. 79, pp. 106–119, 2008, doi: 10.1016/j.pocean.2008.10.013.

[17] J. Yi, Y. Du, Z. He, and C. Zhou, "Enhancing the accuracy of automatic eddy detection and the capability of recognizing the multi-core structures from maps of sea level anomaly," *Ocean Sci. Discuss.*, vol. 10, pp. 825–851, 2013, doi: 10.5194/osd-10-825-2013.

[18] Y. Liu, G. Chen, M. Sun, S. Liu, and F. Tian, "A parallel SLA-based algorithm for global mesoscale eddy identification," *J. Atmospheric Ocean. Technol.*, vol. 33, pp. 2743–2754, 2016, doi: 10.1175/JTECH-D-16-0033.1.

[19] F. Nencioli, C. Dong, T. Dickey, L. Washburn, and J. C. Mcwilliams, "Vector geometry-based eddy detection algorithm and its application to a high-resolution numerical model product and high-frequency radar surface velocities in the Southern California bight," *J. Atmospheric Ocean. Technol.*, vol. 27, pp. 564–579, 2010, doi: 10.1175/2009JTECHO725.1.

[20] G. Xu, C. Dong, Y. Liu, P. Gaube, and J. Yang, "Chlorophyll rings around ocean eddies in the North Pacific," *Sci. Rep.*, vol. 9, 2019, Art. no. 2056, doi: 10.1038/s41598-018-38457-8.

[21] F. Liu, S. Tang, and C. Chen, "Satellite observations of the small-scale cyclonic eddies in the western South China sea," *Biogeosciences*, vol. 11, pp. 13515–13532, 2014, doi: 10.5194/bgd-11-13515-2014.

[22] J. A. Johannessen et al., "Coastal ocean fronts and eddies imaged with ERS1 synthetic aperture radar," *J. Geophys. Res.*, vol. 101, pp. 6651–6667, 1996, doi: 10.1029/95JC02962.

[23] J. A. Johannessen, V. Kudryavtsev, D. Akimov, T. Eldevik, N. Winther, and B. Chapron, "On radar imaging of current features: 2. Mesoscale eddy and current front detection," *J. Geophys. Res.*, vol. 110, pp. C07017, 2005, doi: 10.1029/2004JC002802.

[24] G. Xu, J. Yang, C. Dong, D. Chen, and J. Wang, "Statistical study of submesoscale eddies identified from synthetic aperture radar images in the Luzon Strait and adjacent seas," *Int. J. Remote Sens.*, vol. 36, pp. 4621–4631, 2015, doi: 10.1080/01431161.2015.1084431.

[25] J. Chen, J. Yang, R. Tao, and Z. Yu, "Mesoscale eddy detection and edge structure extraction method in SAR image," *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 237, 2019, Art. no. 032010, doi: 10.1088/1755-1315/237/3/032010.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015, doi: 10.1038/nature14539.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361, doi: 10.1007/978-3-319-10578-9_23.

[28] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified; Real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput610.-Assist. Interv.*, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[31] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[32] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, doi: 10.48550/arXiv.1706.05587.

[33] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851, doi: 10.1007/978-3-030-01234-2_49.

[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6230–6239, doi: 10.1109/CVPR.2017.660.

[35] R. Lguensat, M. Sun, R. Fablet, P. Tandeo, E. Mason, and G. Chen, "EddyNet: A deep neural network for pixel-wise classification of oceanic eddies," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 1764–1767, doi: 10.1109/IGARSS.2018.8518411.

[36] Y. Liu, X. Li, and Y. Ren, "A deep learning model for oceanic mesoscale eddy detection based on multi-source remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Waikoloa, HI, USA, 2020, pp. 6762–6765, doi: 10.1109/IGARSS39084.2020.9323716.

[37] O. J. Santana, D. Hernández-Sosa, J. Martz, and R. N. Smith, "Neural network training for the detection and classification of oceanic mesoscale eddies," *Remote Sens.*, vol. 12, pp. 2625, 2020, doi: 10.3390/rs12162625.

[38] G. Xu et al., "Oceanic eddy identification using an AI scheme," *Remote Sens.*, vol. 11, 2019, Art. no. 1349, doi: 10.3390/rs11111349.

[39] G. Xu, W. Xie, C. Dong, and X. Gao, "Application of three deep learning schemes into oceanic eddy detection," *Front. Mar. Sci.*, vol. 8, 2021, Art. no. 672334, doi: 10.3389/fmars.2021.672334.

[40] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 334–349, doi: 10.1007/978-3-030-01261-8_20.

[41] X. Lu, S. Guo, M. Zhang, J. Dong, X. Chen, and X. Sun, "Mesoscale ocean eddy detection using high-resolution network," in *Proc. 11th Int. Conf. Awareness Sci. Technol.*, Qingdao, China, 2006, pp. 1–6, doi: 10.1109/iCAST51195.2020.9319490.

[42] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5686–5696, doi: 10.1109/CVPR.2019.00584.

[43] H. K. Cheng, J. Chung, Y. W. Tai, and C. K. Tang, "CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 8887–8896, doi: 10.1109/CVPR42600.2020.00891.

[44] Z. Duo, W. Wang, and H. Wang, "Oceanic mesoscale Eddy detection method based on deep learning," *Remote Sens.*, vol. 11, 2019, Art. no. 1921, doi: 10.3390/rs11161921.

[45] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[46] X. Sun, M. Zhang, J. Dong, R. Lguensat, Y. Yang, and X. Lu, "A deep framework for eddy detection and tracking from satellite sea surface height data," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7224–7234, Sep. 2021, doi: 10.1109/TGRS.2020.3032523.

[47] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2006, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.

[48] K. Franz, R. Roscher, A. Milioto, S. Wenzel, and J. Kusche, "Ocean eddy identification and tracking using neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Valencia, Spain, 2018, pp. 6887–6890, doi: 10.1109/IGARSS.2018.8519261.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[50] T. H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, Dec. 2015, doi: 10.1109/TIP.2015.2475625.

[51] H. A. Espedal, O. M. Johannessen, and J. Knulst, "Satellite detection of natural films on the ocean surface," *Geophys. Res. Lett.*, vol. 23, pp. 3151–3154, 1996, doi: 10.1029/96GL03009.

[52] P. M. DiGiacomo and B. Holt, "Satellite observations of small coastal ocean eddies in the Southern California bight," *J. Geophys. Res.*, vol. 106, pp. 22521–22544, 2001, doi: 10.1029/2000JC000728.

[53] S. Karimova and M. Gade, "Eddies in the Red Sea as seen by satellite SAR imagery," in *Remote Sensing of the African Seas*. Dordrecht, The Netherlands: Springer, 2014, pp. 357–378, doi: 10.1007/978-94-017-8008-7_18.

[54] S. Karimova, "Spiral eddies in the baltic; Black and Caspian Seas as seen by satellite radar data," *Adv. Space Res.*, vol. 50, pp. 1107–1124, 2012, doi: 10.1016/j.asr.2011.10.027.

[55] "ESA - Online dissemination - Homepage." Accessed: Sep. 1, 2021. [Online]. Available:https://esar-ds.eo.esa.int/oads/access/collection

[56] "SNAP - Earth online (esa.int)." Accessed: Sep. 1, 2021. [Online]. Available: https://earth.esa.int/eogateway/tools/snap

[57] "wkentaro/labelme: Image polygonal annotation with Python (polygon, rectangle, circle, line, point and image-level flag annotation)." Accessed: Sep. 1, 2021. [Online]. Available: https://github.com/wkentaro/labelme#anaconda

[58] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.

[59] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, doi: 10.48550/arXiv.1804.03999.

[60] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Seoul, South Korea, 2019, pp. 1971–1980, doi: 10.1109/ICCVW.2019.00246.

[61] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[62] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[63] P. Maragos, R. W. Schafer, and M. A. Butt, *Mathematical Morphology and its Applications to Image and Signal Processing*. New York, NY, USA: Springer, 2012, doi: 10.1007/978-1-4613-0469-2.

[64] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090, doi: 10.48550/arXiv.2105.15203.


**Ming Xu** received the B.S. degree in marine technology from Ocean University of China, Qingdao, China, in 2018. She is currently working toward the Ph.D. degree in marine technology with the Ocean University of China, Qingdao, China.

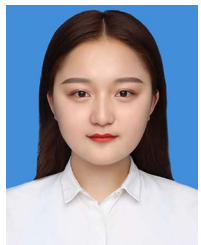
She has been studying ocean remote sensing technology and application.

**Hongping Li** received the M.S. degree in laser from TianJing University, Tianjing, China, in 1988, and the Ph.D. degree in computer science from the University of Oklahoma, Norman, OK, USA, in 2003.

He was a Lecturer with Tsinghua University, Beijing, China, from 1991 to 1997. In 2004, he joined the Faculty of Ocean University of China, Qingdao, China, and served as a Professor with the Department of Marine Technology. His research interests include ocean remote sensing and parallel computing.

**Yuying Yun** received the B.S. degree in geographic information science from Shandong University of Technology, Zibo, China, in 2019, and the M.S. degree in resources and environment from Ocean University of China, Qingdao, China, in 2023.

She has been studying ocean remote sensing and marine physics.



**Fan Yang** received the B.S. degree in land resource management from Shandong Agricultural University, Taian, China, in 2020, and the M.S. degree in resources and environment from Ocean University of China, Qingdao, China, in 2023.

Her current research interests include ocean remote sensing and ocean dynamics system interaction.

**Cuishu Li** received the B.S. degree in landscaping and environmental engineering from Yangzhou University, Yangzhou, China, in 2002.

From 2002 to 2007, she was a Manager with Nanjing Lishui District Garden Management Institute, Nanjing, China. From 2007 to 2018, she was a Section Member with Nanjing Lishui Urban and Rural Construction Bureau, Nanjing, China. She is currently working in Nanjing Lishui District Garden Management Institute, Nanjing, China. Her research interests are resource and environmental management.