SeNet: Structured Edge Network for Sea–Land Segmentation

Dongcai Cheng, Gaofeng Meng, Guangliang Cheng, and Chunhong Pan

***Abstract*— Separating an optical remote sensing image into sea and land areas is very challenging yet of great importance to coastline extraction and subsequent object detection. Traditional methods based on handcrafted feature extraction and image processing often face this dilemma when confronting high-resolution remote sensing images for their complicated texture and intensity distribution. In this letter, we apply the prevalent deep convolutional neural networks to the sea–land segmentation problem and make two innovations on top of the traditional structure. First, we propose a local smooth regularization to achieve better spatially consistent results, which frees us from the complicated morphological operations that are commonly used in traditional methods. Second, we use a multitask loss to simultaneously obtain the segmentation and edge detection results. The attached structured edge detection branch can further refine the segmentation result and dramatically improve edge accuracy. Experiments on a set of natural-colored images from Google Earth demonstrate the effectiveness of our approach in terms of quantitative and visual performances compared with state-of-the-art methods.**

***Index Terms*— Deconvolution network (DeconvNet), local smooth regularization, sea–land segmentation, structured edge network (SeNet).**

I. Introduction

FOR remote sensing imagery, sea–land segmentation is aimed to separate a nearshore image into sea and land regions exactly. The segmentation result is of great importance to coastline extraction [1] and subsequent ship detection [2], [3].

Lots of works have been done on sea–land segmentation and coastline extraction for multispectral images based on thresholding methods [4], [5]. In these methods, the normalized difference water index (NDWI) [6] is an important metric, which takes advantage of the fact that the reflectance of water areas is near to zero in the near-infrared band and high in the green band. Recently, researchers have proposed a new index, combining the MDWI and morphological shadow index (NDWI-MSI) [7], to better differentiate water from shadows for eight-band WorldView-2 urban imagery. Based on that, an urban water classification method [8] has been proposed.

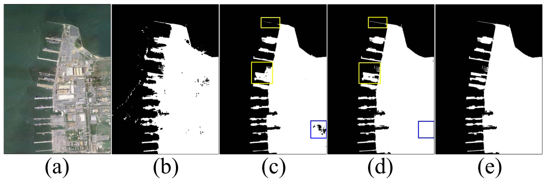

Fig. 1. Results of different segmentation methods. (a) Original image. Results of (b) SMS [10], (c) DeconvNet [11], and (d) our method. (e) Ground truth. Both traditional method and DeconvNet have problems in getting spatially consistent results. Our method can obtain results with very few mislabels in the land region and more accurate edges, such as the comparison of the rectangle regions in (c) and (d).

However, for panchromatic and natural-colored images, directly using global threshold often fails due to the complicated distribution of intensity and texture in land and sea areas. To this end, researchers make efforts on multithreshold selection or combine the thresholding method with other image processing techniques. Liu and Jezek [9] proposed an approach to determine the thresholds for local regions by integrating edge detection and Gaussian curve fitting [locally adaptive thresholding method (LATM)].

Though this method can achieve satisfactory results on the image of coarse spatial resolution, its convergence of fitting process is not guaranteed. You and Li [10] built a statistical model for the sea area (SMS) to decide the segmentation threshold. Connected mislabeled regions in the land results are rectified by evaluating their variances. However, it is unable to solve the problem of misclassification in the sea area.

With the increase in the spatial resolution of remote sensing images, traditional methods often fail when facing with clearer texture and more complicated sea background. To this end, supervised approach with powerful representation ability is needed to obtain accurate results both for the segmentation and edge detection tasks.

The last several years have seen the driving advance of convolutional neural networks (CNNs) in various visual recognition fields like object detection [12], [13], image classification [14], and semantic segmentation [11], [15]. CNNs have also been widely used in remote sensing imagery, such as the land use classification [16] and satellite image scene classification [17]. Instead of using hand-crafted features, CNNs learn hierarchical semantic features by training end to end. For semantic segmentation, fully convolutional network (FCN) [15] replaces the fully connected layers in classification tasks with fixed-size convolution layers, and implements an upsample operation to get the final prediction map. Deconvolution network (DeconvNet) [11] obtains finer results by constructing a symmetrical deconvolution structure on top of FCN to recover semantic features layer by layer.

Manuscript received September 1, 2016; revised November 11, 2016; accepted November 30, 2016. Date of publication December 29, 2016; date of current version January 19, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61370039 and Grant 91338202 and in part by the Beijing Nature Science Foundation under Grant 4162064.

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: dccheng@nlpr.ia.ac.cn; gfmeng@nlpr.ia.ac.cn; guangliang.cheng@nlpr.ia.ac.cn; chpan@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2016.2637439

However, the direct use of DeconvNet to sea–land segmentation has two problems. First, the land regions appear more complicated texture as well as intensity change compared with natural images. Both traditional method and DeconvNet have misclassifications in the land area. An example is illustrated in Fig. 1(b) and (c). Second, CNNs perform undesirably for some long and thin structures, such as wharfs as illustrated in Fig. 1(c). To address these problems, we propose a multitask network that combines segmentation and edge detection. Our work is distinguished by the following contributions.

1) A local smooth regularization term combined with softmax loss is proposed. We can achieve segmentation results with fewer misclassifications both in the land and sea regions without morphological operations.
2) To get finer edge results, we propose a structured edge detection network and combine it with segmentation net to form multitasks. The edge detection branch can not only help to obtain accurate edges but also further improve the segmentation results.

Both training and testing images in this letter are RGB colored, which are collected from Google Earth (GE). Since we focus on improving the edge accuracy, most of the training and testing images are harbor images with complicated distribution of wharfs and ships.

II. Architecture of the Proposed Network

We use DeconvNet [11] as our fundamental network in this letter. Fig. 2 gives a detailed illustration of the proposed model. The loss of our network consists of three parts: 1) softmax loss; 2) local regularization loss; and 3) structured edge loss. We call the overall architecture as structured edge network (SeNet), and will introduce these three parts specifically next.

A. Deconvolution Network

Compared with FCN, DeconvNet can get finer semantic segmentation results with fewer fragments and mislabels due to its layer-by-layer unpooling and deconvolution operations. The DeconvNet of our model is presented in Fig. 2. We implement totally three max poolings in the convolution net. Before each max pooling, we perform convolutions. Between each two convolution layers are batch normalization (BN) [18] layer and ReLU [19] layer sequentially. The numbers of the layer before each pooling are clearly presented in Fig. 2. We remove fully connected layers as [15]. The deconvolution part is mirrored version of convolution net, which contains a series of unpoolings, deconvolutions, BNs, and ReLUs. After the last 32-channel deconvolution layer, we put another two-channel convolution layer as well as the softmax layer to obtain final prediction maps for our two-class problem.

The inputs of our network are RGB images, the size of which are fixed at 300 × 300. The input size for each layer is also fixed in the network, which can be seen in Fig. 2.

B. Local Regularized DeconvNet

Due to the influence of sunlight, altitude, and objects on the ground, land regions often present complicated texture and intensity distribution. Moreover, sometimes the waves make it hard to get clean segmentation results of sea regions. Both of that increase difficulties for accurate segmentation. For instance, the first row of results of DeconvNet in Fig. 3(b) contains mislabels in the land, while the second row in Fig. 3(b) is unsatisfactory on the sea because of the existence of waves.

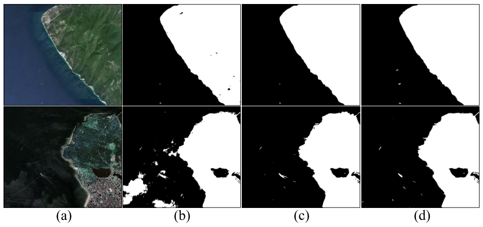

Fig. 3. Comparison of results of DeconvNet and local regularized DeconvNet. (a) Original image. (b) Results of DeconvNet. (c) Results of local regularized DeconvNet. (d) Ground truth. By incorporating local smooth regularization, our method achieves better performance in both land and sea.

The softmax loss in semantic segmentation considers only pixel-wise loss while ignores relationship between neighboring pixels that share similar color values. To better utilize the local relationship of neighboring pixels, we propose a local regularized DeconvNet. Structure of the proposed model is presented as the combination of the top two branches in Fig. 2. The loss function is defined as

$$\text{Loss}_{\text{seg}} = \frac{1}{N}\sum_{i=1}^{N}\left\{-\log p_{l_i,i} + \frac{\lambda}{2}\sum_{j\in Nb(i)}[(p_{0,i}-p_{0,j})^2 + (p_{1,i}-p_{1,j})^2]e^{\frac{-(x_i-x_j)^2}{\sigma}}\right\} \quad (1)$$

where l_i is the ground truth label of pixel i, $l_i = 0$ for sea pixels and $l_i = 1$ for land pixels; $p_{l_i,i}$ is the probability of $i^{'}$s ground truth label; N is the total number of points in the batch; $\boldsymbol{x}_i$ is the color value of i; and $Nb(i)$ are the eight neighbors of i. σ is the variance term, which is computed as the average squared distance between all neighboring pixels in each image : $\sigma = \langle(\boldsymbol{x}_i - \boldsymbol{x}_j)^2\rangle$.

Consider that batch size is 1, (1) can be simplified as

$$\text{Loss}_{\text{seg}} = -\frac{1}{N}\sum_{i=1}^{N}\log p_{l_i,i} + \lambda(Tr(\boldsymbol{P}^T\boldsymbol{L}\boldsymbol{P})) \quad (2)$$

where $\boldsymbol{P} = (\boldsymbol{p_0}, \boldsymbol{p}_1) \in \mathcal{R}^{N\times 2}$, and $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are the pulled row-major probability vectors of the two output maps of softmax layer; $Tr(\cdot)$ is the trace of matrix; L is the Laplacian matrix; and λ is the regularization term.

The second term of (2) enforces neighboring pixels with similar color values to have similar label probabilities. By backpropagating the local relationship of neighboring pixels, the network can learn more meaningful as well as robust features for the segmentation task.

In this letter, we use Caffe [20] to run our network. To avoid out of memory when storing the Laplacian matrix L, we design a $w \times h \times 9$ structure to store L, because for eight-neighbor problem, each row of L has only nine nonzero values, where w and h are the width and height of the input image. In this way, each channel of a point stores the relationship between it and its corresponding nine neighbors (including itself).

The value of λ matters for the segmentation results. A large value of it will yield undersegmentation and decrease the edge accuracy. Moreover, the choice of λ should take the batch

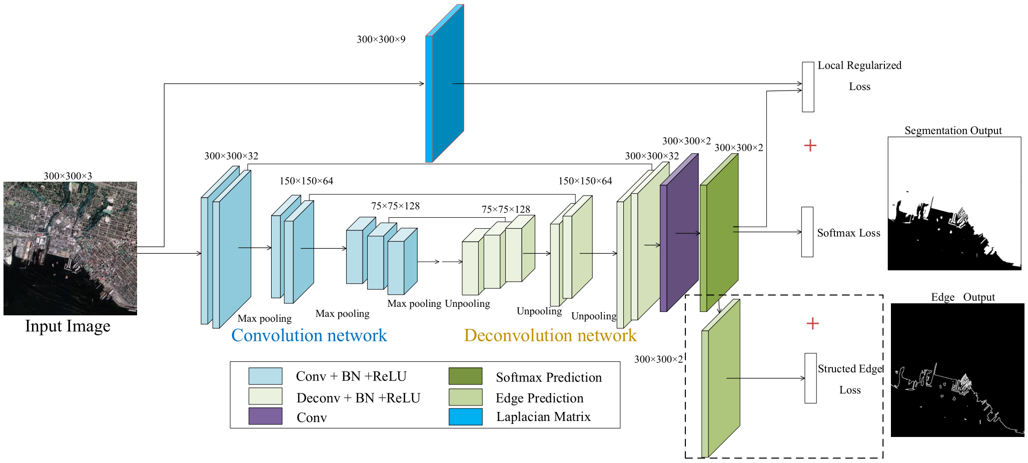

Fig. 2. Illustration of the overall architecture of the proposed model. DeconvNet [11] is employed as our fundamental network. There are three components in our net: local smooth regularization, DeconvNet segmentation, and structured edge detection. The overall architecture is called SeNet. Input size for each layer is fixed, for instance, 300 × 300 × 32 means that input of the layer has 32 channels with each of size 300 × 300.

size into consideration. In this letter, we fix the batch size at four in the training phase, and choose λ from coarse to fine. Satisfactory results on the validation set are acquired when λ is in [10, 50]. Therefore, we set it at ten in the training phase. From Fig. 3(c), we can see that the results of local regularized DeconvNet are more robust with fewer mislabels.

C. Structured Edge Network

For some fine structures, such as wharfs and ships, the segmentation network often has problems to get accurate results near the edge of sea and land.

We propose a multitask network that combines segmentation and edge detection. The bottom branch in Fig. 2 is for edge detection, which shares weights with the segmentation network and differs in prediction maps and loss function. Unlike segmentation net, prediction maps of the edge net represent the probabilities of whether each point is an edge point. Moreover, since there is no discrete point between neighboring sea and land edge pixels, we identify two sets of edge points, one on the land and another on the sea.

We can obtain prediction maps for the edge net only from P_1 map (probability map of the land) of the segmentation network because for each point $p_0+p_1 = 1$. Besides that, we also need the structure information of the edge. For each point i on the edge, we define the likelihood that i belongs to edge as the average p_1-distance between it and its other class neighbors

$$p^e_{1,i} = \frac{\sum_{j=1}^{8} \mathrm{SN}_i(j) \cdot |p_{1,i} - p_{1,i(j)}|}{\sum_{j=1}^{8} \mathrm{SN}_i(j)} \tag{3}$$

where the upper mark e in $p^e_{1,i}$ indicates the probability for edge and $p_{1,i(j)}$, $j \in [1, 8]$ is p_1 value of the jth neighbor of i. SN is a $w \times h \times 8$ structure that stores the structured edge information. If i is an edge point and has a different label with j, then $\mathrm{SN}_i(j) = 1$; otherwise, $\mathrm{SN}_i(j) = 0$. For land edge point i, if $p_{1,i}$ is big while $p_{1,i(j)}$ is small for its sea neighbor j, then i is more likely to be edge point.

For nonedge point, the probability that it belongs to edge is the average p_1-distance between it and its all eight neighbors

$$p^e_{1,i} = \sum_{j=1}^{8} \frac{|p_{1,i} - p_{1,i(j)}|}{8}. \tag{4}$$

To increase the likelihood for edge points on the land, we should maximize $p_{1,i}$ while minimizing $p_{1,i(j)}$ for neighbor j that has a different label with i. Let $E(L)$ denote edge set on the land and $E(S)$ denote edge set on the sea. We define loss function of the edge network as

$$\begin{aligned}\mathrm{Loss}_{\mathrm{edge}} &= -\frac{1}{N}\left\{\sum_{i\in E(L)} \frac{\sum_{j=1}^{8} \mathrm{SN}_i(j)[\log p_{1,i} + \log(1 - p_{1,i(j)})]}{\sum_{j=1}^{8} \mathrm{SN}_i(j)}\right.\\ &\quad + \sum_{i\in E(S)} \frac{\sum_{j=1}^{8} \mathrm{SN}_i(j)[\log(1 - p_{1,i})+\log p_{1,i(j)}]}{\sum_{j=1}^{8} \mathrm{SN}_i(j)}\\ &\quad \left. + \sum_{i\notin E(L)\cup E(S)} \log(1 - \sum_{j=1}^{8} \frac{|p_{1,i} - p_{1,i(j)}|}{8})\right\}\end{aligned} \tag{5}$$

where SN stores the structured edge information for ground truth edge points, as defined before.

The first two terms of $\mathrm{Loss}_{\mathrm{edge}}$ make segmentation results on the edge be in accordance with ground truth. We can also consider them as the increase in the weights of true edge points in softmax loss of the segmentation network. The third term enforces neighboring points that have the same ground truth label to have similar segmentation probabilities, which further enhances the local smooth constraint for getting spatially consistent results.

The final model of our network is

$$\mathrm{Loss}_{\mathrm{SeNet}} = \mathrm{Loss}_{\mathrm{seg}} + \mathrm{Loss}_{\mathrm{edge}}. \tag{6}$$

The integration of segmentation and edge detection helps the network to learn distinguishable features for edge points.

III. Results and Evaluation

A. Data Description and Augmentation

The remote sensing images used in this letter are natural-colored and collected from GE with a spatial resolution of 3.0–5.0 m. There are 140 training images, 6 validation images, and 60 testing images, each with size larger than 800 × 800. We label ground truth of all the images by hands. The ground

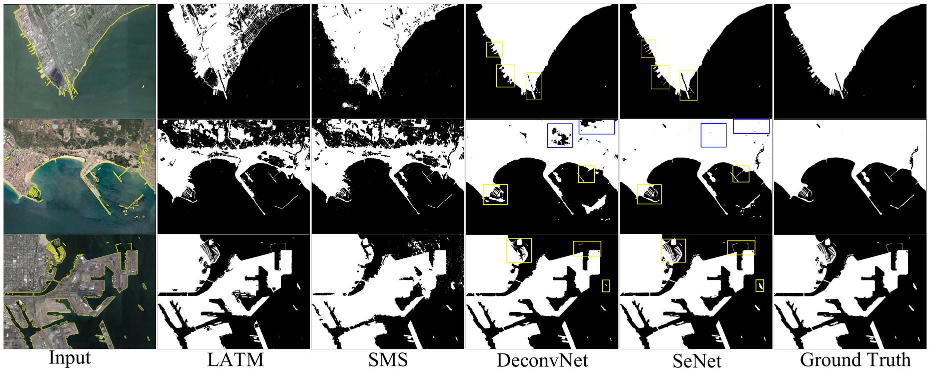

Fig. 4. Visual comparison of segmentation results of four methods. No morphological operations are implemented for DeconvNet and our method (SeNet). We consider ships as part of the land. The edge results of SeNet are overlapped on the input. By comparing the yellow rectangle regions marked in results of DeconvNet and ours, we can see the improvement of edge accuracy that our network has brought. The comparison of blue rectangle regions shows that our results are more spatially consistent.

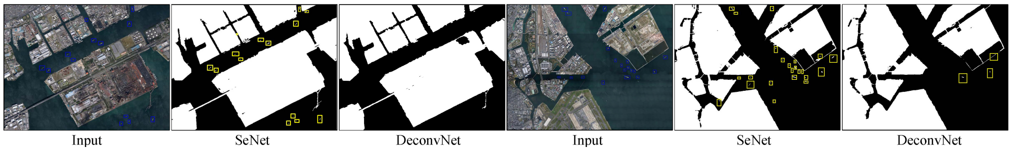

Fig. 5. Segmentation of ships of our method compared with DeconvNet. Blue rectangles are ground truth ships, and yellow rectangles are ship candidates.

truth of the edge is automatically calculated in the training phase for each input.

To augment the training and validation sets, we first crop four corners and center of the original images with size 300×300. Second, the original images are random cropped to generate another 20 samples. Then we rotate and mirror each of the above 25 samples to further extend one sample to eight. In this way, the training and validation sets are augmented 200 times. There are totally 28 000 training samples and 1200 validation samples, each with size 300×300.

B. Strategy: Training, Fine Tuning, and Inference

Due to the influence of waves and shallow waters, the edges of sea and land in some training images are obscure and even cannot be precisely recognized by humans' eye. Therefore, we pick 110 original training images that have clear edge distribution to form fine-tuning set. The fine-tuning set is augmented in the same way as before.

We train the network in two phases. First, local regularized DeconvNet is trained with all the training samples. Then we fine tune the overall SeNet with the fine-tuning set. We set the batch size as four, initial learning rate (lr) as 0.01, and decrease lr to one-tenth after 10 000 iterations. Both for training and fine tuning, the loss of the network is converged after about 50 000 iterations in the experiments.

At testing phase, only segmentation net is used to calculate the accuracy. The input size of the net is adjusted to 800×800 to increase the segmentation accuracy. Input size for each layer is also adjusted accordingly. We process each of the testing images with bilinear interpolation, and reinterpolate outputs of the net to the original size. The negative effects that brought by interpolation are negligible since our testing images have size similar to 800×800.

C. Compared Methods and Evaluation Indexes

Three approaches are compared with our method: LATM: by fitting Gaussian curves for local regions [9]; SMS: global thresholding method by building statistical model for the sea [10]; DeconvNet [11].

We implement LATM and SMS by ourselves with C++ programming, which have achieved the same performance on images presented in original letters. Connected mislabeled regions within land with size smaller than 200 pixels are rectified to improve performance of LATM and SMS. The architecture of DeconvNet at http://cvlab.postech.ac.kr/research/deconvnet/ is used both for comparative method and for our fundamental network. We impose no morphological operation on DeconvNet and our method to better assess their performance.

Evaluation indexes are defined to compare and analyze the results. For segmentation results, we define land precision (LP), land recall (LR), overall precision (OP), and overall recall (OR) as

$$\mathrm{LP} = \frac{\mathrm{TP}_L}{\mathrm{TP}_L + \mathrm{FP}_L}, \quad \mathrm{LR} = \frac{\mathrm{TP}_L}{\mathrm{TP}_L + \mathrm{FN}_L}$$
$$\mathrm{OP} = \frac{\mathrm{TP}_L + \mathrm{TP}_S}{\mathrm{TP}_L + \mathrm{FP}_L + \mathrm{TP}_S + \mathrm{FP}_S}$$
$$\mathrm{OR} = \frac{\mathrm{TP}_L + \mathrm{TP}_S}{\mathrm{TP}_L + \mathrm{FN}_L + \mathrm{TP}_S + \mathrm{FN}_S} \tag{7}$$

where TP_L, FP_L, and FN_L are true positive, false positive, and false negative of land. TP_S, FP_S, and FN_S are true positive, false positive, and false negative of sea. OP combines precision of land and sea. OR combines recall of land and sea.

As for the edge accuracy, we first define edge precision (EP) and edge recall (ER) as

$$\mathrm{EP} = \frac{\sharp \text{ of true edge points on segmentation edge}}{\sharp \text{ of edge points in segmentation results}}$$
$$\mathrm{ER} = \frac{\sharp \text{ of true edge points on segmentation edge}}{\sharp \text{ of true edge points}}. \tag{8}$$

One edge point in segmentation results is considered as correct if the distance between it and its closest ground truth edge point is smaller than N, and one ground truth edge point

TABLE I

SEGMENTATION RESULTS ON FOUR EVALUATION INDEXES

Method	LP(%)	LR(%)	OP(%)	OR(%)
LATM	98.22	79.18	87.05	87.04
SMS	95.93	93.93	94.07	96.05
DeconvNet	98.53	97.56	97.29	97.27
SeNet	**99.69**	**98.15**	**98.12**	**98.11**

TABLE II

F1-MEASURE (%) OF THE EDGE RESULTS

Method	N=1	N=2	N=3	N=4	N=5
LATM	38.68	42.14	44.52	46.30	47.75
SMS	53.52	58.50	61.92	64.43	66.39
DeconvNet	76.04	79.84	81.93	83.35	84.44
SeNet	**91.07**	**92.19**	**92.98**	**93.59**	**94.08**

is regarded as on the edge of segmentation results if it falls within N pixels of it. N will be set with different values. Then the F1-measure [21] of the edge accuracy is $\text{F1-measure} = 2\text{EP} \cdot \text{ER}/\text{EP} + \text{ER}$.

D. Visual and Quantitative Analysis

We compare the performances of four methods and show the results in Fig. 4. Traditional methods (LATM and SMS) are more prone to misclassify land pixels for images with complicated texture and intensity change, such as the top three images in Fig. 4. DeconvNet gets better results, but still with mislabels in land. Compared with DeconvNet, our method (SeNet) can obtain more spatially consistent results. Moreover, we achieve better edge results due to the discriminative ability for edge pixels that our edge net has brought.

We calculate average LP, LR, OP, and OR on 60 testing images and present them in Table I. Our method has best results on all of the four indexes. Since most of the testing images are harbor images that have relative peace sea surface, the LP values of the four methods are high. On the other hand, the misclassifications in land, which are revealed by LR, are different for different methods. SeNet has almost 0.83% improvement in OP and 0.84% improvement in OR with respect to DeconvNet.

F1-measures of edge results are presented in Table II. The distance N is set with different values. We can see that SeNet achieves significant promotion of the edge accuracy compared with traditional methods and DeconvNet.

We also calculate the running time of different methods. For an 800 × 800 image, the times of SMS and LATM are 3.72 and 36.85 s, respectively, on 2.8 GHz Intel, Microsoft visual studio 2010 platform. On NVIDIA K20 GPU with 4-GB memory, both DeconvNet and SeNet need only 650 ms.

E. Segmentation for Ships

The segmentation accuracy of ships counts for the subsequent region of interest (ROI) extraction and detection. Our method can better segment some thin and long structures such as wharfs, as well as small targets such as ships, which will benefit the ROI extraction and ship detection. The results in Fig. 5 demonstrate that SeNet has good effects on ship segmentation. By calculating the area as well as shape characteristics, we can obtain ship candidates, i.e., the yellow rectangle regions.

IV. CONCLUSION

In this letter, we have applied the prevalent CNNs to sea–land segmentation and proposed a SeNet. The proposed local smooth regularization makes segmentation results spatially consistent. By integrating edge network with Deconvnet, we can get more accurate edge results compared with traditional methods and DeconvNet.

REFERENCES

[1] J. E. Pardo-Pascual, J. Almonacid-Caballer, L. A. Ruiz, and J. Palomar-Vázquez, "Automatic extraction of shorelines from Landsat TM and ETM+ multi-temporal images with subpixel precision," *Remote Sens. Environ.*, vol. 123, pp. 1–11, Aug. 2012.

[2] G. Liu, Y. Zhang, X. Zheng, X. Sun, K. Fu, and H. Wang, "A new method on inshore ship detection in high-resolution satellite images using shape and context information," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 3, pp. 617–621, Mar. 2014.

[3] C. Zhu, H. Zhou, R. Wang, and J. Guo, "A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[4] T. Kuleli, A. Guneroglu, F. Karsli, and M. Dihkan, "Automatic detection of shoreline change on coastal Ramsar wetlands of turkey," *Ocean Eng.*, vol. 38, no. 10, pp. 1141–1149, Jul. 2011.

[5] T. Zhang, X. Yang, S. Hu, and F. Su, "Extraction of coastline in aquaculture coast from multispectral remote sensing images: Object-based region growing integrating edge detection," *Remote Sens.*, vol. 5, no. 9, pp. 4470–4487, 2013.

[6] S. K. McFEETERS, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *Int. J. Remote Sens.*, vol. 17, no. 7, pp. 1425–1432, 1996.

[7] C. Xie, X. Huang, W. Zeng, and X. Fang, "A novel water index for urban high-resolution eight-band WorldView-2 imagery," *Int. J. Digit. Earth*, vol. 9, no. 10, pp. 925–941, 2016.

[8] X. Huang, C. Xie, X. Fang, and L. Zhang, "Combining pixel- and object-based machine learning for identification of water-body types from urban high-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2097–2110, May 2015.

[9] H. Liu and K. C. Jezek, "Automated extraction of coastline from satellite imagery by integrating canny edge detection and locally adaptive thresholding methods," *Int. J. Remote Sens.*, vol. 25, no. 5, pp. 937–958, Mar. 2004.

[10] X. You and W. Li, "A sea-land segmentation scheme based on statistical model of sea," in *Proc. 4th Int. Congr. Image Signal Process. (CISP)*, vol. 3. Oct. 2011, pp. 1155–1159.

[11] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. ICCV*, Dec. 2015, pp. 1520–1528.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.

[13] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[16] I. Ševo and A. Avramović, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 740–744, May 2016.

[17] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 157–161, Feb. 2016.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[20] Y. Jia *et al.* (Jun. 2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[21] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vo. 2, no. 1, pp. 37–63, 2011.