SEA-LAND SEGMENTATION WITH RES-UNET AND FULLY CONNECTED CRF

Zhengquan Chu, Tian Tian, Ruyi Feng, Lizhe Wang*

Hubei Key Laboratory of Intelligent Geo-Information Processing,
School of Computer Science, China University of Geosciences, Wuhan 430074, P.R.China.

ABSTRACT

Sea-land segmentation is a key step in inshore ship detection and coast monitoring. Among the state-of-art segmentation approaches, semantic segmentation networks show great potential on this task, but there is still room for improvement. In this paper, we propose a method based on UNet for sea-land segmentation. We replace its contraction part with ResNet which specializes in handling complicated scenes, and construct a new network structure ***Res-UNet***. After preliminary segmentation results are obtained, the fully connected Conditional Random Field (CRF) model and morphological operation are then used as post-processing to obtain more precise coastlines and intact regions. We test our model on a dataset collected from Google Earth and the inspiring results validate the effectiveness of our method.

Index Terms— sea-land segmentation, semantic segmentation networks, Unet, ResNet, fully connected CRF

1. INTRODUCTION

Sea-land segmentation plays an important role in many remote sensing applications such as coast and port monitoring. It is also one of the key steps in inshore ship detection, which can eliminate the interference of land objects and enhance the accuracy and efficiency of detection. Traditional sea-land segmentation usually adopts threshold segmentation or image segmentation methods, which suffers from unsatisfactory results or dependence on hand-crafted features. Without empirical parameters and expert knowledge, effect of segmentation is hard to be guaranteed.

As the development of deep learning, Convolutional Neural Networks (CNNs) have achieved great success in various computer vision tasks, such as image classification and object detection. Its great ability of extracting hierarchical features has also been extended to image semantic segmentation. By modifying the original CNN's fully connected layers to convolutional layers and applying deconvolution for resolution restoration, a number of outstanding networks such as FCN[1] and UNet [2], have emerged on the task of segmentation. As a result, the development of deep neural networks on semantic segmentation offers us powerful tools to deal with sea-land segmentation in remote sensing fields.

Although semantic segmentation networks seem to show great potentials on sea-land segmentation tasks, there are still some challenges. Unlike natural images, where objects are inclined to appear in fixed regions of images, scenes in remote sensing images are more complex. For example, objects of different categories vary drastically in scales, even objects of a same category can be very different in shapes, colors and positions.

*Corresponding to tiantian@cug.edu.cn. This work is supported by the National Natural Science Foundation of China (Grant No. 41701417 and No. U1711266).

Focusing on this notorious problem, in this paper, we propose a new scheme to apply deep neural network of semantic segmentation on the task of sea-land segmentation. Specifically, we employ the UNet structure as our segmentation framework, and make some modification on this baseline. To deal with the complex land-cover scenes in remote sensing images, we adopt the residual modules of ResNet [3] to construct the contraction part of UNet. Therefore, the residual structure guarantees the network a better description capability when contracting images with deeper layers. However, limited by the crop size of patches, the image-level context information may somewhat be lost, which will hamper the accuracy of final segmentation result. We use two post-processing methods, fully connected conditional random field and a morphological operation, to produce a more precise segmentation result.

The rest of this paper is organized as follows: Section 2 supplements related work, Section 3 describes our proposed method, Section 4 provides experimental results and discussions, and Section 5 concludes this paper.

2. RELATED WORK

Since the appearance of CNN, many efficient networks are proposed to deal with segmentation tasks. FCN [1] is the first model which is transformed from CNN to fit the segmentation task. Different from CNNs, FCN uses "deconvolution" (also named transpose convolution) to recover its' spatial resolution, and then apply a Softmax layer to give every pixel a probability of which class it belongs to. Moreover, in order to improve the segmentation result, skip-connection architecture is used to combine former layers with later layers, and this helps FCN to make full use of both the low and high level features. Based on FCN, many other models are then proposed. SegNet [4] shares a similar idea with FCN, and the differences are the designs of encoder and decoder. It uses max pooling index in the upsampling process and obtains better resolution restoration. UNet [2], named after its U-shaped network structure, consists of a contraction and an expansion part. Compared to FCN, it uses more skip-connection layers and adds two convolutional layers between the skip-connection layers. These improvements make UNet more powerful to learn the features and recover image's original size.

Segmentation networks have also been improved and applied on remote sensing tasks. SeNet [5] applies the prevalent deep CNN to sea-land segmentation and uses an edge detection branch to obtain better results. DeepUNet [6] applies DownBlocks and UpBlocks based on UNet and improves the accuracy of sea-land segmentation. ShipNet [7] proposes a novel network and designs a weighted loss function for simultaneous sea-land. MS-FCN [8] proposes a multi-scale full convolutional network based sea-land segmentation method in order to solve the inshore ship detection problem.

Since the direct output of network may be rough, researchers often enhance the results by means of post-processing methods.

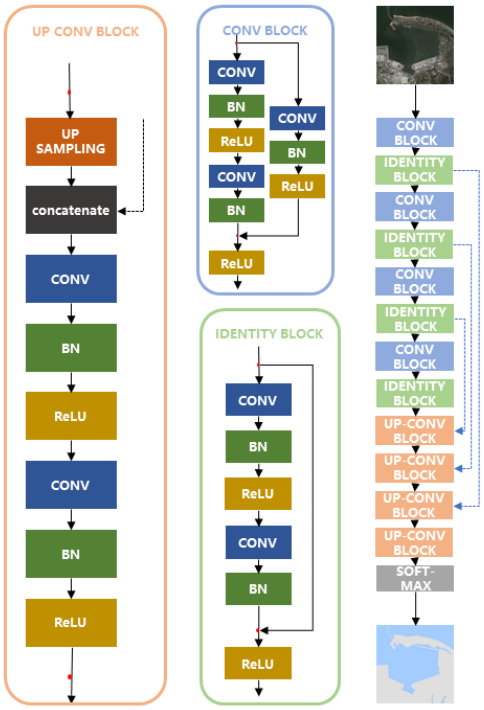


Fig. 1. ***Res-UNet*** network architecture for sea-land segmentation.

DeepLab network [9] adopts the fully connected CRF in their pipeline, which was extended from Conditional Random Field (CRF) [10]. CRF is a probabilistic model originally proposed for segmenting and labeling sequence data, and it was extended to fully connected CRF for semantic segmentation. But their applications were limited due to the complexity of inference until an efficient approximate algorithm was proposed [11]. Afterwards, fully connected CRFs have become practical tools which are commonly used for segmentation post-processing.

3. METHODOLOGY

3.1. Network design

Our network is evolved from the fully convolutional networks and UNet. Fig.1 illustrates our network architecture for sea-land segmentation. Generally, our network can be divided into two parts: the encoder and the decoder. In the encoder part, ResNet-18 is modified to accommodate sea-land segmentation task: since we need a dense prediction label, we abandon all the layers followed by the global average pooling layers, which are designed for classification tasks. Additionally, we change the kernel size of the first pooling layer from 3×3 to 2×2 in order to get the exact half size of input for the first block on account of convenient concatenation. As shown in Fig.1, the encoder part includes two kinds of blocks, which both share the same concept of "short-cut" proposed in ResNet. However, the channels of input feature maps are different for layers in these two block, so additions at the end of block would be mismatched. Therefore, the short-cut path of ***CONV_ BLOCK*** has an extra convolutional layer followed by batch normalization and ReLU activation layers. But in ***IDENTITY_ BLOCK***, the short-cut path is clean. To clarify the priority of these blocks, let's consider a complex mapping function $\mathcal{H}(x)$ between the input images(x) and its output feature maps, the mapping may consist of several stacks of convolutional and pooling layers. All theses layers can asymptotically approximate a residual function $\mathcal{F}(x)$

$$\mathcal{F}(x) := \mathcal{H}(x) - x \tag{1}$$

More precisely, the output of l^{st} layer y_l is:

$$\begin{aligned} y_l &= h(x_l) + \mathcal{F}(x_l, \mathcal{W}_l) \\ x_{l+1} &= f(y_l) \end{aligned} \tag{2}$$

Here f is the activation function, $\mathcal{W}_i$ are weights of the networks that need to be learned. If f is an identity mapping: $x_{x+1} \equiv y_l$, then the Eq.(2) can be simplified. Moreover, if we consider all the continuous layers, then we have

$$x_L = x_l + \sum_{i=1}^{L-1} \mathcal{F}(x_i, \mathcal{W}_i) \tag{3}$$

Based on the chain rule of back-propagation, the gradient of Eq.(3) can be computed:

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\left(1 + \frac{\partial}{\partial x_L}\sum_{i=l}^{L-1}\mathcal{F}(x_i, \mathcal{W}_i)\right) \tag{4}$$

Obviously, the gradient $\frac{\partial \varepsilon}{x_l}$ will never be zero if the term $\frac{\partial}{\partial x_L}\sum_{i=l}^{L-1}$ $\mathcal{F}(x_i, \mathcal{W}_i)$ is not -1, which actually cannot be always -1 for all samples in a mini-batch. This ensures the gradient flows smoothly and never vanish during the training of a network, and this is the reason why ResNet can converge so quickly. After four combination of these two blocks, the output feature map size of encoder part will be 16 times smaller than the input images.

In the decoder part, we design an ***UP_ CONV_ BLOCK*** to recover image's resolution. This block takes the output of encoder layers as input, and then up-pooling layer is applied to up-sample the feature maps so they will have the same size as features in encoder part. Symmetrically, two more convolutional layers and batch normalization with ReLU are applied in this block to make our decoder part more smooth. As what is done in UNet to make full use of features, we stack four ***UP_ CONV_ BLOCKs*** and bridge encoder and decoder by using skip-connection between ***IDENTITY_ BLOCKs*** and ***UP_ CONV_ BLOCKs***. Finally, a Softmax layer is added to make prediction of every pixel of the image.

3.2. Post-processing

The output prediction labels of the deep network may still be rough, so a proper post-processing is adopted to adjust the prediction labels for more precise details. Fully connected Conditional Random Field (CRF) is a frequently used image processing technique to gain finer detail. And we adopt the efficient probability approximation algorithm [11] to implement fully connected CRF in our pipeline. Consider a conditional random filed $(\mathbf{I}, \mathbf{X})$, in which $\mathbf{I} = \{I_1, \ldots, I_N\}$ are input images and X are labels assigned to every pixel in an image. More specifically, the label of pixel j with color vector I_j denotes by X_j. Then the Gibbs distribution of $(\mathbf{I}, \mathbf{X})$ is:

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} exp(-\sum_{c \in C_{\mathcal{G}}} \psi_c(\mathbf{X}_c)) \tag{5}$$

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the graph on $\mathbf{X}$, $c \in C_{\mathcal{G}}$ is the cliques in $\mathcal{G}$ and its potential denotes by ψ_c. So, the Gibbs energy on a complete graph of the fully connected CRF model is:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{6}$$

The first term $\psi_u(x_i)$ is a unary potential that can be produced by a CNN in our experiments, which represents the probability distribution of labels assigned to pixels. The second term in Eq.(6) is a pairwise potential, and we use the contrast-sensitive two-kernel potentials following literature [11]:

$$\begin{aligned} k(\mathbf{f}_i, \mathbf{f}_j) = \omega^{(1)} exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right) \\ + \omega^{(2)} exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right) \end{aligned} \tag{7}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are feature vectors of pixel i and j, I_i and I_j are color vectors, p_i,p_j are their positions, and θ_α, θ_β and θ_γ are hyperparameters.

The fully connected CRF makes our prediction results even better, but we notice that it also brings in some scatter pixels in the label images. To address this problem, we use a morphological operation to find the connected domains in binary images and then remove contiguous holes smaller than the specified size. This method can be quite helpful since the subjects to be segmented are connected vast sea or land in our specific task.

4. EXPERIMENT

4.1. Dataset

The dataset we used is downloaded from Google Earth which contains 208 images, and all the images have three bands with the size about 1000×1000 and resolution of 3∼5m. We manually label every pixel to land and sea, and Fig.2 shows examples of our training images and their labels.

The training set is composed of 150 images we choose randomly, and the remaining are used for evaluation. In order to fit our network, firstly, we crop patches of size 224×224 from the 150 original images randomly to obtain 3750 patches. Then each of them is rotated and flipped, and further processed by changing the gamma curve or adding some noises. The evaluation set is also cropped randomly, but without augmentation. Finally, we get 22500 images for training and 1250 images for test.

4.2. Experiment details

For all the experiments, we initialize our models with he_normal initializer [12] and train them using Adam optimizer[13]. The initial learning rate is set to be 1e–3 and multiplied 0.5 every 10 epochs. We use validation set to earlystop the training. We monitor evaluation losses and use patience of 15. Due to the limited physical memory on GPUs, we set the batch size to 4 for all experiments.

We train our model on a 1050Ti GPU with Keras and TensorFlow backbend. The evaluation losses of our ***Res-UNet*** model converge after 43 epochs for a total of 193K iterations.

After generating test images, we use fully connected CRFs to detail our result. We use both unary and pairwise potential as described in Sec.3.2, and set the hyperparameters as follows: $\theta_\alpha = 80, \theta_\beta = 13, and \theta_\gamma = 3$. To eliminate tiny scattered holes or isolated pixels, we use morphological methods to get final results, where the minimum size of threshold is set to 200 pixels.

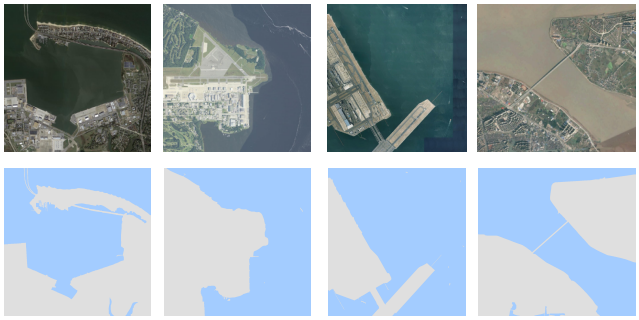


Fig. 2. Examples of training data. The first row shows four RGB images, and corresponding label images are on the second row.

4.3. Results and analysis

Our method is tested and experimental results are shown in Table.1. ***UNet*** [2] is the baseline model, and the model we extend with ResNet is named as ***Res-UNet***. To evaluate the effectiveness of post-processing methods, we also show the test results after applying fully connected CRF and morphological operation followed behind, which are denoted by ***Res-UNet_CRF*** and ***Res-UNet_mor*** respectively. The framework ***Res-UNet_mor*** achieves the best result of 98.25% overall accuracy and 98.15% f1-score on both sea and land segmentation average from 58 test images.

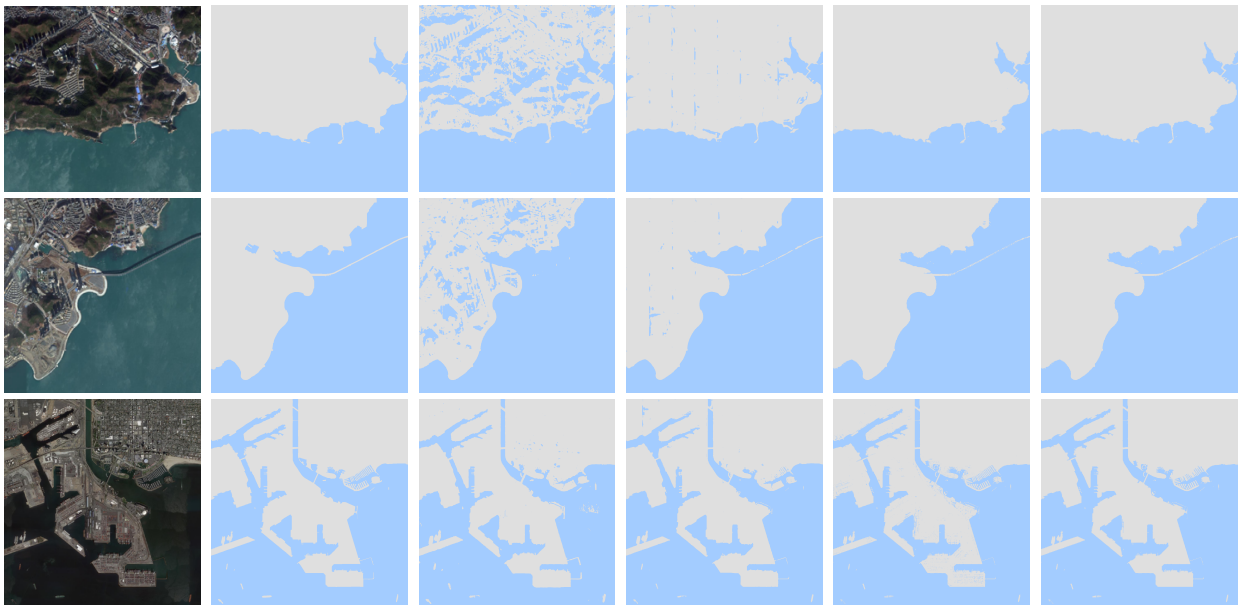
Visual results are shown in Fig.3. Obviously, ***UNet*** roughly distinguish sea and land, but fail to correctly classify the whole land due to its complicated ground objects. For example, some shadows of high mountains or tall buildings and dark green vegetation are easily mistaken for sea by ***UNet***. In comparison, the remarkable feature-extraction ability of ***Res-UNet*** makes it can understand the images in a global perspective and achieves higher segmentation accuracy. Moreover, fully connected CRF takes the original images into account and brings in more details for the coast and better integrity for the land. As the third row in Fig.3 shows, ***Res-UNet_CRF*** accurately distinguish the inshore ships from sea even their intervals are very narrow. However, the fully connected CRF may also cause some "noise" points, and this drawback can be mitigated by morphology operation shown as ***Res-UNet_mor***.

Table 1. Segmentation overall accuracy and F1-score of sea and land on test set.

Models	Sea	Land	F1-score	OA
UNet	0.9229	0.9348	0.9289	0.9343
Res-UNet	0.9713	0.9743	0.9728	0.9753
Res-UNet_CRF	0.9792	0.9792	0.9792	0.9804
Res-UNet_mor	**0.9815**	**0.9815**	**0.9815**	**0.9825**

5. CONCLUSION

In this paper, we exploit the semantic segmentation network to tackle the problem of sea-land segmentation. In consideration of the feature extraction capability of ResNet, we implement residual modules on



(a) Original images (b) Ground truth (c) UNet (d) Res-UNet (e) Res-UNet_CRF (f) Res-UNet_mor

Fig. 3. Visual results of tested models.

the basis of UNet to design a novel network ***Res-UNet*** for our segmentation task. In order to fine-tune our result, fully connected CRF is selected to provide more details. To further eliminate the isolated pixels, a morphological operation is finally taken to produce more reasonable segmentation results.

6. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, Cham, 2015, pp. 234–241.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint:1511.00561*, 2015.

[5] Dongcai Cheng, Gaofeng Meng, Guangliang Cheng, and Chunhong Pan, "SeNet: structured edge network for sea–land segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 2, pp. 247–251, 2017.

[6] Ruirui Li, Wenjie Liu, Lei Yang, Shihao Sun, Wei Hu, Fan Zhang, and Wei Li, "Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018.

[7] Shihao Sun, Zexin Lu, Wenjie Liu, Wei Hu, and Ruirui Li, "Shipnet for semantic segmentation on vhr maritime imagery," in *2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 6911–6914.

[8] Lei Liu, Guowei Chen, Zongxu Pan, Bin Lei, and Quanzhi An, "Inshore ship detection in sar images based on deep neural networks," in *2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 25–28.

[9] L. C. Chen, G Papandreou, I Kokkinos, K Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[10] J. Lafferty, "Conditional random fields : Probabilistic models for segmenting and labeling sequence data," *Proceedings of ICML*, 2001.

[11] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," in *Advances in Neural Information Processing systems*, 2011, pp. 109–117.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, vol. 00, pp. 1026–1034.

[13] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint: 1412.6980*, 2014.