Pixel-wise segmentation of SAR imagery using encoder-decoder network and fully-connected CRF

Fei Gao[1], Yishan He[1], Jun Wang[1], Fei Ma[1], Erfu Yang[2], Amir Hussain[3]

[1] Beihang University, Beijing 100191, China
[2] University of Strathclyde, Glasgow G1 1XJ, UK
[3]Cognitive Big Data and Cyber-Informatics (CogBID) Laboratory, Edinburgh Napier University, Edinburgh EH10 5DT, UK
feigao2000@163.com, hys1009hys @126.com

Abstract. Synthetic Aperture Radar (SAR) image segmentation is an important step in SAR image interpretation. Common Patch-based methods treat all the pixels within the patch as a single category and do not take the label consistency between neighbor patches into consideration, which makes the segmentation results less accurate. In this paper, we use an encoder-decoder network to conduct pixel-wise segmentation. Then, in order to make full use of the contextual information between patches, we use fully-connected conditional random field to optimize the combined probability map output from encoder-decoder network. The testing results on our SAR data set shows that our method can effectively maintain contextual information of pixels and achieve better segmentation results.

Keywords: SAR image segmentation, encoder-decoder network, fully-connected CRF

1 Introduction

Synthetic Aperture Radar (SAR) is an advanced instrument for earth observation and is widely used in various areas of the country's economy and defense construction [1,2]. SAR image segmentation plays an important role in SAR image interpretation [3] and it is a prerequisite for many further applications. For example, in the task of assessing crop coverage, different types of crops need to be segmented first [4] and areas with oil slick need to be first segmented when detecting slick on the sea surface [5].

The segmentation of SAR image means categorizing SAR image pixel by pixel, meanwhile maintaining the spatial structure of different regions. Patch-based methods are commonly used in recent research. These methods mainly compose of three steps: (1) divide the whole SAR image into patches; (2) extract the feature of each patch for classification; (3) combine the classification results of patches into segmentation results. For example, Geng [6] proposed a deep convolutional autoencoders (DCAE) to extract and optimize patch features. DCAE consists of a feature extractor, an average pooling layer, and sparse autocoders. The feature extractor extracts gray-level cooccurrence matrix (GLCM) and Gabor features from images patches. The pooling layer conducts scale transformation. And the sparse autocoders are used for feature optimization. In [7], the author uses the pretrained Alexnet to extract patch features. K-means clustering is then applied for visual word encoding. Finally, a Naive Bayesian classifier is adopted to classify the coded patches. In [8], a hierarchically adversarial network is introduced to extract features of superpixels. Duan [9] integrates dual-tree complex wavelet transform into convolutional neural network as a hidden layer to improve the feature extraction ability of the network. Inspired by biological vision system, Gao [10] presents a hierarchical method for river detection, where biologically visual saliency modeling is used to extract superpixels' features. However, a common drawback exists in these patch-based methods. Although the patch is relatively small, it may still contain different categories of pixels. The accuracy of segmentation results will be reduced if treating the whole patch as a single category.

One way to overcome the above disadvantage is to adopt pixel-wise segmentation methods. In the field of optical image segmentation, pixel-based methods are mainstream. In these methods, the original image is fed as input of specially designed convolutional neural network. The feature extraction process is automatically accomplished through hidden layers. And the pixel-wise segmentation result is obtained from the output of the network. For example, Long [11] proposes a fully convolutional network (FCN) for semantic segmentation of optical images. By replacing the fully-connected layer with convolution layer, the input image size of FCN can be arbitrary. The pixel-wise segmentation results are acquired directly from the output of FCN. In [12], the author proposes a U-shaped convolutional neural network, called Unet. Unet first downsamples the input image to small feature maps and then upsamples them to pixel-wise segmentation results. It achieves good performance in biomedical image segmentation. Segnet [13] is a fully convolution network with encoder-decoder structure. It is initially used to deal with the semantic segmentation tasks of driverless vehicles or intelligent robots. The structure of its encoder and decoder is symmetrical. To make the segmentation result more accurate, the spatial information saved by the downsampling layer of encoder is then utilized in the upsampling layers of the decoder. In this way, more accurate segmentation result can be achieved by Segnet and it runs faster as well. The pixel-based method in optical image segmentation includes mainly two benefits. The first is that it is an end-to-end model that makes full use of the strong feature extraction ability of convolutional neural network. The second is that it can directly output pixel-level segmentation results, which makes the segmentation result more accurate**.** However, for high-resolution SAR images, the computational cost is unacceptable if we directly take the whole SAR image as an input. In order to address this problem, we can first divide the SAR images into patches and get the pixel-level segmentation results of patches through the network. Final segmentation results can be obtained by combining them together. But in this way, the spatial relationships between patches is neglected, which makes the segmentation results of neighbor patches not consistent enough.

Conditional random field (CRF) is a discriminative model based on undirected graph. It can naturally combine the feature information of image patches and contextual information between patches to model the posterior probability of labels. The traditional CRF model used in SAR image segmentation is based on patches. For example, in order to effectively integrate contextual prior into segmentation process, Chu [14] combines CRF with Bayesian network to optimize the over-segmented SAR images. However, patch-based CRF is not suitable for the pixel-wise segmentation in this paper. In [15], a pixel-wise CRF called fully-connected CRF with gaussian edge potentials is proposed. In addition to taking pixel-wise features into consideration, this model can integrate the spatial relationship among all pixels to optimize the segmentation result. Besides, an optimization algorithm is also proposed to make inference of pixel-wise CRF feasible.

For the above reasons, in this paper, we first divide the original SAR image into patches so as to reduce the computational complexity. Then we choose Segnet to conduct pixel-level segmentation on the patches. In order to further utilize the contextual information between patches and improve the neighborhood label consistency, we use fully-connected CRF to optimize the whole combined segmentation map.

The segmentation results of our approach show that we can obtain more accurate segmentation results within patches by using Segnet and achieve better neighborhood label consistency among different patches by adopting fully-connected CRF.

2 Proposed method

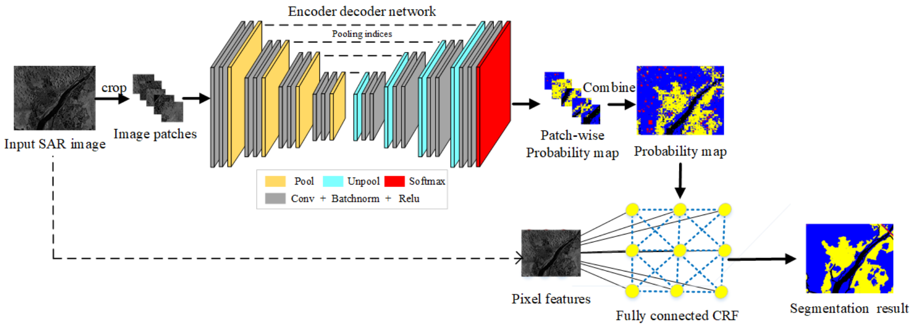


Fig. 1. The flow diagram of our method

The flow diagram of our segmentation method is shown in Fig. 1. In our approach, the original SAR image is first cropped into small image patches. Then these patches are fed into the encoder-decoder network, i.e. Segnet, to obtain patch-wise probability maps, which show the probability that each pixel belongs to different categories within the patches. The combined probability map, together with the grayscale and position features of the original SAR image pixels, are fed into fully-connected CRF for optimization. Final segmentation results can be obtained after several iteration steps using fully-connected CRF.

2.1 Encoder-decoder network for pixel-wise segmentation

For pixel-wise segmentation network, we use Segnet proposed in [13], which achieves pixel-wise segmentation through end-to-end training. The network can be divided into two parts: encoder and decoder. The specific structure of these two parts is shown in Fig. 2. The encoder structure includes the first 13 convolution layers in VGG16. Five pooling layers is adopted to reduce the size of the feature maps. The structure of the decoder is symmetrical to the encoder, which consists of five upsampling layers and 13 convolution layers. The upsampling layer utilizes the position information stored during pooling. In this way, the spatial relationship can be maintained when upsamping the feature maps to the original size. Besides, the number of the training parameters are greatly reduced. The feature maps are sparse after upsampling, so trainable convolution layers are adopted to generate dense feature maps. The output of the decoder is fed into the softmax classifier to obtain the category probability of each pixel.

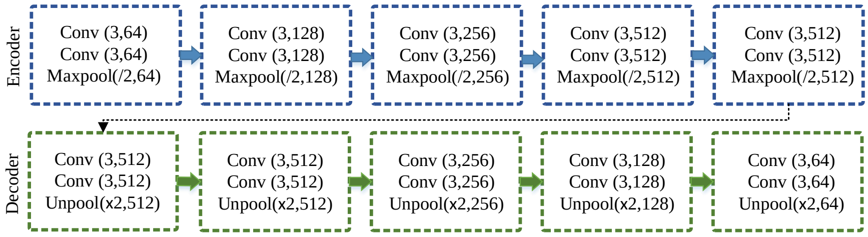


Fig. 2. The structure of encoder and decoder used in Segnet, the first item in bracket means kernel size, the second item in bracket means output dimension.

2.2 Fully-connected CRF for posteriori probability optimization

Random field $\mathbf{I}$ is a set of random variables $\{I_1, I_2, \cdots, I_N\}$, which represents a high resolution the SAR image. I_i represents the feature vector of pixel i. Random field $\mathbf{X}$ is a set of random variables $\{X_1, X_2, \cdots, X_N\}$, X_i representing the label of pixel i. Conditional random field $(\mathbf{I}, \mathbf{X})$ can be defined by Gibbs distribution as:

$$P(X = x \mid I) = \frac{1}{Z(I)} \exp(-E(X \mid I)) \tag{1}$$

where $E(X \mid I)$ denotes energy function, $Z(I)$ is a normalization term.

It can be seen from Eq. (1) that solving the maximum posteriori probability can be simplified to minimizing the energy function $E(X \mid I)$, as is shown in (2).

$$x^* = \arg\max_x P(x \mid I) = \arg\min_x E(x \mid I) \tag{2}$$

The energy function is composed of unary term ψ_u and pairwise term ψ_p:

$$E(x)=\sum\nolimits_{i}\psi_{u}(x)+\sum\nolimits_{i\neq j}\psi_{p}(x_{i},x_{j}) \tag{3}$$

For fully-connected CRF, the output probability map by Segnet can be used for unary term ψ_u . And the pairwise term ψ_p considers the relationship between each pixel and all other pixels, which can be describe by (4).

$$\psi_{p}(x_{i},x_{j})=\mu(x_{i},x_{j})\underbrace{\sum\nolimits_{\text{m}=1}^{K}\omega^{(m)}k^{(m)}(\mathbf{f}_{i},\mathbf{f}_{j})}_{k(\mathbf{f}_{i},\mathbf{f}_{j})} \tag{4}$$

where μ denotes label compatibility function， $\omega^{(m)}$ is the weight coefficient and $k(\mathbf{f}_i,\mathbf{f}_j)$ indicates two-kernel potentials, which is shown in (5).

$$k(\mathbf{f}_{i},\mathbf{f}_{j})=\omega^{(1)}\exp\left(-\frac{|\,p_{i}-p_{j}\,|^{2}}{2\theta_{\alpha}^{2}}-\frac{|\,g_{i}-g_{j}\,|^{2}}{2\theta_{\beta}^{2}}\right)+\omega^{(m)}\exp\left(-\frac{|\,p_{i}-p_{j}\,|^{2}}{2\theta_{\gamma}^{2}}\right) \tag{5}$$

where p_i and p_j denotes the position of pixel i and pixel j , g_i and g_j denotes the grayscale value of the pixels in SAR image. θ_α and θ_β control the degree of nearness and similarity, θ_γ controls the degree of smoothness. These parameters are not trainable in our experiment.

According to mean field approximation theory, the problem of estimating the maximum posteriori probability can be transformed into minimizing the K-L divergence of a distribution function $Q(x)$ and the probability function $P(x)$ by iteration. The iteration process is as follows:

$$Q_{i}(x_{i}=l)=\frac{1}{Z_{i}}\exp\left\{-\psi_{u}(x_{i})-\sum_{l'\in L}\mu(l,l')\sum_{m=1}^{K}\omega^{(m)}\sum_{j\neq i}k^{(m)}(\mathbf{f}_{i},\mathbf{f}_{j})Q_{j}(l')\right\} \tag{6}$$

3 Experiment

3.1 Experiment data

The SAR data set for experiment includes 36 airborne SAR images, which were acquired in Fangchenggang, Guangxi, China. These images cover an area of about 30 km × 30 km in total. The size of each image is 1419 × 1122 and the resolution is 2m. For training and testing, we labeled all the images manually according to Google Map, categorizing the area into four classes: urban, farmland, river and background.

For the purpose of comparison, 7 of all 36 SAR images were selected as training set and the rest 29 SAR images belong to test set in all experiments. Several SAR images and corresponding ground truth in the training set are shown in Fig. 3 It can be seen that although the distribution of different regions in high resolution images is not uniform, each category of regions is relatively concentrated. It reflects the neighborhood consistency of SAR images.

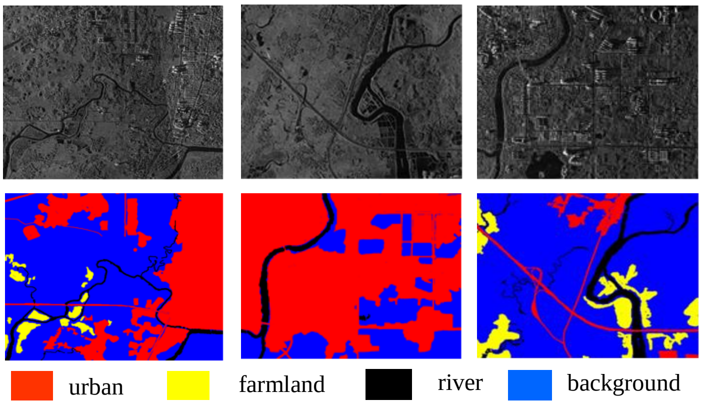


Fig. 3. Several SAR images and corresponding ground truth from our dataset

3.2 Evaluation metrics

In our experiment, we choose four pixel-wise metrics, including overall accuracy (OA), overall precision (OP), f1-score and kappa coefficient to quantitatively evaluate the performance of segmentation methods. These metrics are calculated according to (7)-(10). Our experiment platform is configured with 32G memory, Intel (R) Xeon (R) CPU L5639 @ 2.13GHz * 1, and one Tesla K20c GPU.

$$OA = \sum_{i=1}^{c} x_{ii} / N \tag{7}$$

where x_{ii} denotes the diagonal elements of the confusion matrix, N stands for the total number of pixels of SAR image and c represents the number of the categories.

$$OP = (\sum_{i=1}^{c} \alpha_i x_{ii}) / N \tag{8}$$

where $\alpha_i = \sum_{j=1}^{c} x_{ji} / \sum_{j=1}^{c} x_{ij}$, x_{ij} is the element in the i_{th} row and j_{th} column of the confusion matrix.

$$\text{f1-score} = \frac{2 \cdot \text{precision} \times \text{recall}}{\text{precision+recall}} \tag{9}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{10}$$

where $p_0 = OA$, $p_e = \sum_{i=1}^{c} \gamma_i / N, \gamma_i = \sum_{j=1}^{c} x_{ji} \times \sum_{j=1}^{c} x_{ij}$.

3.3 Experiment settings

We conduct three experiments for comparison, finetune Alexnet [7], deep convolutional autoencoder (DCAE) [6], and Segnet without fully-connected CRF.

In the Alexnet model, the convolution and pooling layers of Alexnet pretrained on Imagenet dataset are used as feature extractor, which outputs 256-dimensional features. These features are then fed into a fully-connected network with a hidden layer for classification. During training, the parameters of convolution and pooling layers are fixed and the parameters of the full connection layer are updated by backward propagation algorithm. The loss function used in this model is mean square error function; learning rate is set to 1e-4; batch size is set to 330; the number of epochs is 40; the weight decay coefficient is 0.0005. As suggested in [7], the patches are cropped with the size of 21*21 and the step of 10.

In DCAE model, GLCM and Gabor features are first extracted using convolutional kernels. Average pooling is then employed for scale transformation and Principal Component Analysis (PCA) is used to reduce the computational cost. Two cascaded autoencoders are then utilized to optimize these features. The training of the autoencoders is conduct through greedy layer-wise strategy. Finally, a fully-connected layer with the softmax activation function is used to classify the optimized features. Mean square error function is used as loss function; the learning rate is set to 1e-4; batch size is set to 330; the number of epochs is set to 40; the weight decay coefficient is 1e-8 and the patch size is 32*32 with a step of 16.

The Segnet model is introduced in Section 2. For hyperparameters, we adopt mean square error function as the loss function; the learning rate is set to 0.01; batch size is set to 300; the training epochs is set to 50; the weight decay coefficient is 1e-8 and patch size is 32*32.

3.4 Experiment result

In this part, we will quantitatively compare our method with the aforementioned methods, i.e. Segnet, DCAE and Alexnet. Table 1 lists the segmentation results of these methods on test set. As is shown in Table 1, the segmentation results of Segnet combined with fully-connected CRF is better than other methods in OA, OP, f1-score and kappa, reaching 86.36, 86.54, 85.69, 0.7187 respectively. And the segmentation result of Segnet is better than DCAE and Alexnet. The results suggest that Segnet is capable of making more accurate prediction within patches. Fully-connected CRF makes further improvement thanks to its ability of utilizing contextual information.

In order to further compare the segmentation performance of different methods in different categories, Table 2 gives the f1-score in different categories achieved by each method. We can see from the table that fully-connected CRF greatly improves the segmentation result in different categories, especially in farmland and urban areas.

Table 1. The overall performance metrics for each method

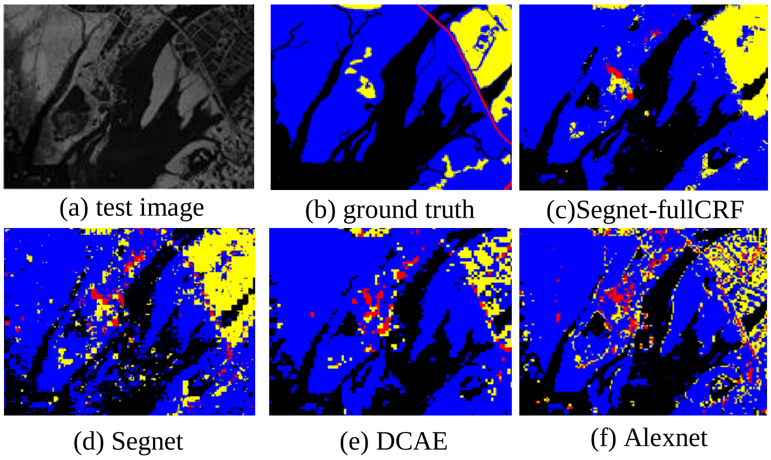
Metrics	Segnet-fullCRF	Segnet	DCAE	Alexnet
OA	**86.36**	81.86	81.09	78.32
OP	**86.54**	83.02	82.66	78.09
f1-score	**85.69**	81.98	80.99	76.56
κ	**0.7187**	0.6233	0.6076	0.5289

Fig. 4 and Fig. 5 visually illustrate the segmentation results of the methods on two SAR images in the test set. It is shown in Fig. 4 that the segmentation result of Segnet is better than other patch-based methods. After further optimization by fully-connected CRF, the label consistency of farmland and river areas is strengthened and the edges between different categories of areas are preserved. It proves that the contextual information between patches is effectively utilized by fully-connected CRF.

Table 2. f1-score in different categories

Category	Segnet-fullCRF	Segnet	DCAE	Alexnet
urban	**66.62**	56.69	61.63	45.79
farmland	**80.71**	61.45	43.25	25.12
river	**79.84**	75.00	76.20	70.15
background	**90.50**	87.80	86.98	80.69

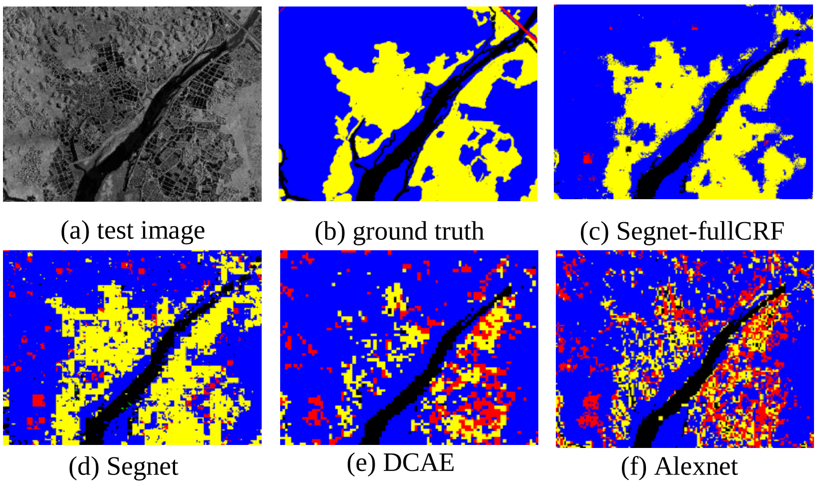
The segmentation results shown in Fig. 5 clearly indicate that the label consistency of farmland area is enhanced and the edges of river and farmland become smoother after optimization by fully-connected CRF. In addition, many misclassified pixels in the river and background area are corrected after optimization. It verifies that fully-connected CRF model is capable of jointly utilizing the gray feature of pixels and the contextual information so that better segmentation can be achieved.



(a) test image (b) ground truth (c)Segnet-fullCRF

(d) Segnet (e) DCAE (f) Alexnet

Fig. 4. The segmentation results of each model(**a**) Input SAR image (**b**) Ground truth. (**c**) Segnet-fullCRF(OA=86.36) (**d**)Segnet(OA=81.86) (**e**)DCAE(OA=81.09) (**f**) Alexnet (OA=78.32)



(a) test image (b) ground truth (c) Segnet-fullCRF

(d) Segnet (e) DCAE (f) Alexnet

Fig. 5. The segmentation results of each model (**a**) Input SAR image. (**b**) Ground truth. (**c**) Segnet-full CRF. (**d**) Segnet. (**e**) DCAE. (**f**) Alexnet

4 Conclusion

In this paper, we combine the encoder-decoder network used in optical image segmentation with fully-connected CRF for high resolution SAR image segmentation. To improve the segmentation accuracy within the patch, we use Segnet to obtain pixel-wise segmentation result of the patches. In order to make full use of contextual information and strengthen neighborhood label consistency between patches, we adopt fully-connected CRF to optimize the probability maps output by Segnet. This method is compared with several other patch-based segmentation methods in our experiment. The experiment result demonstrates that the encoder-decoder network has superior performance in SAR image segmentation, and that fully-connected CRF effectively utilizes contextual information and greatly optimizes the segmentation results.

Acknowledgements

This research was funded by the National Natural Science Foundation of China, grant nos. 61771027, 61071139, 61671035. Professor A. Hussain was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/M026981/1. The SAR images used in the experiments are the courtesy of Beijing Institute of Radio Measurement, the authors would like to thank them for their support in this work.

References

1. Gao, F., Huang, T., Wang, J., Sun, J., Hussain, A., & Yang, E.: Dual-Branch Deep Convolution Neural Network for Polarimetric SAR Image Classification. Applied Sciences 7(5), 447-460 (2017).
2. Fei, G. , Zhenyu, Y. , Jun, W. , Jinping, S. , Erfu, Y. , & Huiyu, Z.: A novel active semi-supervised convolutional neural network algorithm for sar image recognition. Computational Intelligence and Neuroscience 14(7), 1-8 (2017).
3. Oliver C, Quegan S.: Understanding Synthetic Aperture Radar Images, (2004)
4. Shao, Y. , Fan, X. , Liu, H. , Xiao, J. , Ross, S. , & Brisco, B. , et al.: Rice monitoring and production estimation using multitemporal radarsat. Remote Sensing of Environment 76(3), 310-325 (2001).
5. Frédéric Galland, Philippe Réfrégier, & Germain, O. .: Synthetic aperture radar oil spill segmentation by stochastic complexity minimization. Geoscience & Remote Sensing Letters IEEE 1(4), 295-299 (2004).
6. Geng, J. , Fan, J. , Wang, H. , Ma, X. , Li, B. , & Chen, F. .: High-resolution sar image classification via deep convolutional autoencoders. IEEE Geoscience and Remote Sensing Letters 12(11), 1-5 (2015).
7. Tian, T. , Chang, L. , Jinkang, X. , & Jiayi, M. .: Urban area detection in very high resolution remote sensing images using deep convolutional neural networks. Sensors 18(3), 904-910 (2018).
8. Ma, F. & Gao, F. & Sun, J. & Zhou, H. & Hussain, A.: Weakly Supervised Segmentation of SAR Imagery Using Superpixel and Hierarchically Adversarial CRF. Remote Sensing 11. 512. 10.3390/rs11050512, (2019).
9. Duan, Y., Liu, F., Jiao, L., Zhao, P., & Zhang, L.: SAR Image segmentation based on convolutional-wavelet neural network and markov random field. Pattern Recognition 64, 255-267. doi:https://doi.org/10.1016/j.patcog.2016.11.015, (2017).
10. Gao, F. , Ma, F. , Wang, J. , Sun, J. , & Zhou, H. .: Visual saliency modeling for river detection in high-resolution sar imagery. IEEE Access PP(99), 1-8 (2017).
11. Long, J. , Shelhamer, E. , & Darrell, T. .: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence 39(4), 640-651 (2014).
12. Ronneberger, O., Fischer, P., & Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. Paper presented at the International Conference on Medical Image Computing & Computer-assisted Intervention, (2015).
13. Badrinarayanan, V. , Kendall, A. , & Cipolla, R. .: Segnet: a deep convolutional encoder-decoder architecture for image segmentation, (2015).
14. Chu, H. , Xinlong, L. , Di, F. , Bo, S. , Bin, L. , & Mingsheng, L. .: Hierarchical terrain classification based on multilayer bayesian network and conditional random field. Remote Sensing 9(1), 96-108 (2017).
15. Krähenbühl, P., & Koltun, V.: Efficient inference in fully-connected crfs with gaussian edge potentials. Paper presented at the Advances in neural information processing systems, (2011).