FRF-Net: Land Cover Classification From Large-Scale VHR Optical Remote Sensing Images

Qianbo Sang, Yin Zhuang, Shan Dong, Guanqun Wang, and He Chen

***Abstract*— Deep learning (DL) technique is widely applied in remote sensing (RS) applications because of its outstanding nonlinear feature extraction ability. However, with regard to the issues of large-scale and very high-resolution (VHR) land cover classification, multi-object distributions and clear appearance with large intraclass difference become challenges for refined pixelwise land cover mapping. Focusing on these problems, the letter proposed a novel encoding-to-decoding method called the full receptive field (RF) network (FRF-Net) based on two types of attention mechanism. In the FRF-Net, ResNet-101 is used as the basic backbone. Then, the ensemble feature is generated by encoding the high-level features based on the self-attention mechanism which could achieve full RF to capture long-range semantic. Next, the encoding result is decoded by the fusion attention mechanism combined with the low-level feature to produce a fusion feature which contains a refined semantic description for accurate land cover mapping. Extensive experiments based on the GID and ISPRS data sets proved that the proposed network outperforms the state-of-the-art methods. The FRF-Net achieved 66.71% and 64.17% of the mean of classwise Intersection over Union (mIOU) with smaller computation cost on ISPRS and GID, respectively.**

***Index Terms*— Deep learning (DL), land cover classification, remote sensing (RS), semantic segmentation, very high resolution (VHR).**

I. Introduction

LAND cover classification is the popular research field to leverage remote sensing (RS) applications in urbanization planning, land resource management, and disaster monitoring [1]–[3]. With the development of RS technique, multiple sensors (i.e., optical, lidar, synthetic aperture radar) are used to provide the refined pixelwise land cover mapping results [4]–[6]. Because of very high resolution (VHR) and clear land cover appearance, optical RS becomes the main approach for refined land cover mapping. The typical land cover classification methods can be roughly divided into three categories, which include the spectral-based [7], object-oriented-based [8], [9], and pixelwise-based [10]–[13] methods.

Manuscript received April 24, 2019; revised June 21, 2019; accepted August 26, 2019. Date of publication September 25, 2019; date of current version May 21, 2020. This work was supported in part by the Chang Jiang Scholars Program under Grant T2012122 and in part by the Hundred Leading Talent Project of Beijing Science and Technology under Grant Z141101001514005. *(Corresponding author: Yin Zhuang.)*

Q. Sang, G. Wang, and H. Chen are with the Beijing Key Laboratory of Embedded Real-Time Information Processing Technology, Beijing 100081, China.

Y. Zhuang is with the School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: zhuangyin640829@163.com).

S. Dong is with the Engineering Center of Digital Audio and Video, Communication University of China, Beijing 100024, China.

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.2019.2938555

For spectral-based methods, the endmember selection and abundance estimation for the unmixing process are the challenging tasks, which struggle to get accurate land cover mapping results. The object-orientation-based methods usually use manually designed features (i.e., colors, textures, shapes, and orientations) to achieve global or local feature expression for land cover classification, but these manually designed features cannot provide a separable description of different categories in feature space. Therefore, it is difficult for the above methods to provide accurate results. With regard to the pixelwise-based methods, a lot of deep learning (DL) semantic segmentation frameworks were used because of their outstanding feature expression ability [6], [7], [10]–[13]. Lin *et al.* [10] proposed the multi-scale fully convolutional network (FCN) and conducted the maritime semantic segmentation by increasing the receptive field (RF) to maintain fine-scale details and coarse-scale high-level context feature to produce finer results. Mou *et al.* [12] proposed a novel bidirectional network called recurrent network in FCN (RiFCN) to fuse the shallow layer and deep layer feature maps and achieve more accurate semantic segmentation results. In general, these semantic segmentation methods all focus on exploring enlarged RF and multi-level feature fusion expression for refined land cover mapping.

However, the performance of the land cover classification is severely affected by the data complexity [14]. In optical RS images, related to multiple land cover distributions and clear appearance with an intraclass drastically changing, the feature expression becomes a challenging task for the pixelwise semantic segmentation frameworks. In this letter, we not only consider effective full RF to capture the long-range semantic but also consider better feature fusion method to leverage refined land cover mapping. Then, a novel full RF network (FRF-Net) is proposed, which is composed of three parts: the ResNet-101 [15] feature extraction backbone, self-attention encoding process, and fusion attention decoding process. The encoding-to-decoding process could simultaneously ensure the accuracy of land cover classification and location. Finally, extensive experiments use the GID [11] and ISPRS [12] data sets to demonstrate that the proposed FRF-Net outperforms the state-of-the-art methods. Our method can provide 64.17% of the mean of classwise Intersection over Union (mIoU) and 76.24% of pixel accuracy (PA) on GID and also achieve 66.71% of mIoU and 85.73% of PA on ISPRS. Furthermore, the FRF-Net reduces the computation cost with fewer parameters and faster training procedure than others. In general, the main contributions of the FRF-Net this letter proposed can be summarized as two points.

1) We proposed a novel semantic segmentation network and achieved accurate land cover mapping based on two different types of attention mechanism.

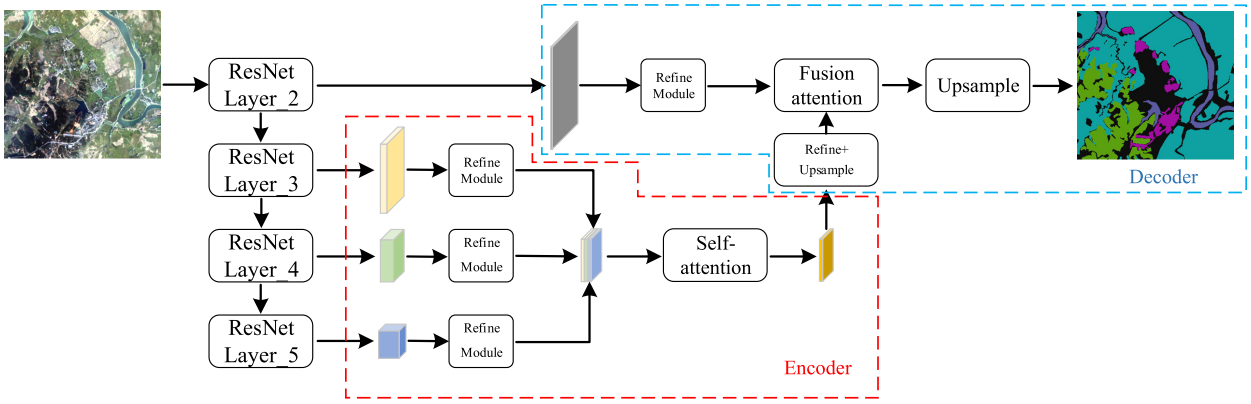

Fig. 1. General architecture of the proposed FRF-Net.

2) We proposed an encoding-to-decoding structure, which could capture the long-range semantic with effective full RF of the ensemble feature and fuse the fine-scale details of the low-level feature to generate effective feature expression.

II. Related Work and Materials

Our research field focused on addressing the problem of accurate land cover mapping by the semantic segmentation technique. In nature scene, a lot of DL-based networks such as FCN [16], U-Net [17], global convolutional network (GCN) [18], PSPNet [19], Deeplab v3+ [20], and Seg-Net [21] can provide refined results. However, when directly using these networks for land cover classification of optical RS images, they would provide a lot of incomplete and nonsequence results caused by limited RF. Therefore, we design a DL semantic segmentation network, which is very appropriate for land cover mapping of large-scale RS images.

Then, the GID [11] and ISPRS [12] data sets are used to evaluate the proposed and comparison methods. Here, the GID called "Gaofen Image Data Set" consists of 150 VHR Gaofen-2 image annotations covering areas about 50 000 km^2 in China. Furthermore, the GID contains six categories (i.e., unknown, built-up, farmland, forest, meadow, and water) with the complex RS scenes. Another data ISPRS named "International Society for Photogrammetry and Remote Sensing," is an online benchmark data set, which is composed of high-resolution airborne images of Vaihingen for semantic segmentation. The pixels in the ISPRS data set are labeled with the following six land cover classes: impervious surfaces (ISs), building, low vegetation (LV), tree, car, and background (BG).

III. Methodology

In this section, the general overview and analysis of the network are introduced in Section III-A. Then, the encoding and decoding processes are individually described in detail to show their roles and principles in Sections III-B and III-C, respectively.

A. General Overview and Analysis

The semantic segmentation problem could be separated as two aspects: the pixelwise classification and location tasks. The classification mainly relies on the larger RF in the high-level features, where the RF should be large enough to capture long-range semantic and improve the classification accuracy. On the other hand, the fine-scale detail information of the low-level feature plays a major role in the location task.

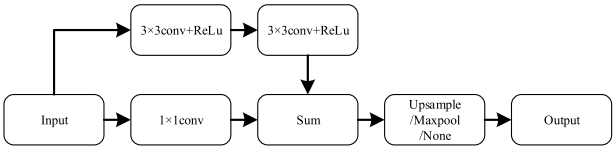

Fig. 2. Refinement module in the proposed FRF-Net.

For the pixelwise classification, as applied by almost all semantic segmentation networks, the gradual convolutional operations could truly enlarge the RF to capture the global dependencies, but the process is computation-consuming and the RF is also inefficient. Regarding the pixelwise location, previous works just roughly connect the low-level features with the high-level features, which cannot adaptively merge features for extracting valid information. Therefore, aiming at addressing the aforementioned problems, we propose the FRF-Net based on the self-attention and fusion attention mechanisms to prevent the quantities of the convolutional operations, obtain effective full RF, and generate the valid fusion feature expression.

The framework is illustrated in Fig. 1. The pretrained ResNet-101 [15] proved to be helpful for avoiding overfitting [22], which is applied as the basic backbone to generate multi-level feature expressions. Next, the encoding module can give effective full RF for each pixel of the ensemble feature to capture the long-range semantic based on the self-attention mechanism. Then, the decoding module merges the ensemble feature with fine-scale information from the low-level feature based on the fusion attention mechanism. Before the encoding and decoding processes, the refinement module in Fig. 2 is used to enhance the multi-level features and regulate them into the same size. The encoding and decoding processes are described in detail as follows.

B. FRF-Net Encoding Process

The land cover pixelwise mapping results depend on the classification capability to a large extent. In general, the encoding can further abstract the high-level features and easily obtain larger RF. We go even further in our encoding module by introducing the self-attention mechanism to capture the long-range semantic information.

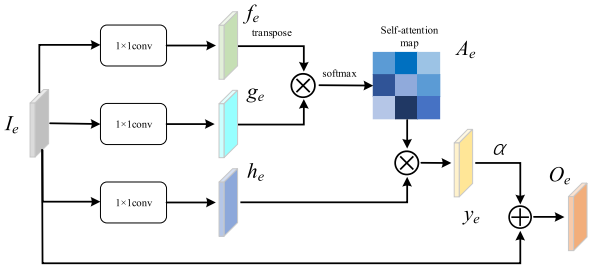

Fig. 3. Self-attention module for encoding high-level feature.

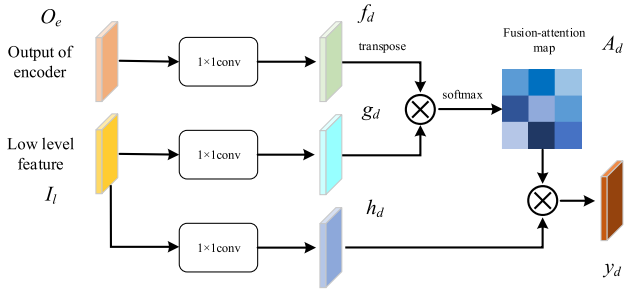

Fig. 4. Fusion attention module for fine-scale feature fusion in the decoding process.

The self-attention mechanism in the encoding module is illustrated in Fig. 3. First, the refined high-level features from the ResNet-101 layer_3, layer_4, and layer_5 are concatenated as the input $I_e \in R^{C_{I_e} \times H_e \times W_e}$. Then, I_e is fed into two 1×1 conv layers separately to generate f_e and g_e with the same number of channels C_e. Then, they are reshaped to $R^{C_e \times N_e}$, where $N_e = H_e \times W_e$. After that, the correlation of the matrix multiplication between transposed f_e and g_e is used to generate the trainable self-attention map $A \in R^{N_e \times N_e}$. After the softmax layer, the variance correlation coefficients in A can be expressed as follows:

$$a_{j,i} = \frac{\exp(p_{ij})}{\sum_{i=1}^{N} \exp(p_{ij})} \tag{1}$$

where $p_{i,j} = (f_e^T \otimes g_e)_{i,j} = \sum_{n=1}^{C} f_{e\,i,n}^T \otimes g_{en,j}$, $a_{j,i}$ is the variance correlation coefficient, which indicates the contribution of the ith pixel making to the jth pixel inference. Through another 1×1 conv, h_e is generated and reshaped to $R^{C_e \times N_e}$. The encoding feature y_e is generated by multiplying h_e with the self-attention map A_e and reshaping to $y_e \in R^{C_e \times H_e \times W_e}$. y_e obtains full RF to model the long-range semantic dependence. Finally, the input is added back to generate the output O_e, which can be expressed as follows:

$$O_e = \alpha y_e + I_e \tag{2}$$

where α is a trainable coefficient initialized as "0." As shown in (2), the output ensemble feature O_e is composed of input high-level feature I_e and adaptive y_e.

C. FRF-Net Decoding Process

Generally, the output of the encoder with decreased resolution would lose the detail information. Therefore, the decoding process is an essential part to recover the details and resolution. The bilinear up-sampling or deconvolution could also be considered as a simple decoding module which recovers details not depending on the low-level feature. However, with regard to the complicate and irregularity boundaries or shapes of land cover, recovering by the naive up-sampling or deconvolution module will affect the accuracy of the pixelwise location. In some previous works, the low-level feature is directly combined with the high-level feature, and this combination is not an effective way to merge two-level features. To efficiently fuse the ensemble feature O_e in (2) into the low-level feature, we proposed the decoding module based on the fusion attention mechanism which is illustrated in Fig. 4.

The ensemble feature O_e and low-level feature I_l are individually fed into the decoding attention module, in which the trainable fusion attention map A_d is used to supervise the ensemble feature fused into the low-level feature to generate y_d.

Finally, different from the encoding module (2), we do not add the back input feature conv which is used to adjust the channel number of the low-level feature I_l and the ensemble feature O_e separately. Empirically, the channels' number of the low-level feature I_l should be less than the output of decoding y_d. Otherwise, the decoding module perhaps overweights the low-level feature and ignores the dependence of the ensemble feature. Similarly, the channel number of the ensemble feature should be more relatively to promise the accuracy of the classification. These three parts of I_l, y_d, and O_e are concatenated, and a few convolutions are applied to reduce the channels according to classes and refine the details of the concatenated feature maps. Finally, the fusion feature maps are bilinear up-sampled to increase the spatial resolution to the original images.

IV. Experiments

To demonstrate the effectiveness of the proposed FRF-Net, the GID and ISPRS data sets, which are described in Section II, are used, and several state-of-the-art powerful networks such as Deeplab v3+ [20], GCN [18], PSPNet [19], U-Net [17], and Seg-Net [21] are chosen as comparisons.

Our method is implemented on Pytorch. The experiments are performed on NVIDIA TITAN Xp GPU. We use standard stochastic gradient descent (SGD) optimizer with momentum 0.9 and weight decay 0.0005. Our learning rate scheduled by poly started with 1×10^{-2}. All models are trained for 500 epochs with a batch size of 16. We also apply data augmentations to avoid overfitting, including random flipping horizontally, random scaling, random Gaussian blur, and random cropping. To unify the metrics with the state-of-the-art methods, the PA [16] and mIoU indexes [16] are used to evaluate all comparison methods.

A. Comparisons on GID Data Set

For the GID data set, all data are four times down-sampled and clipped with a size of 360×340. Where 3750 images, which have a large intraclass difference and small interclass diversities, are split into 2000 training and 1750 validation images with annotation. For six land cover categories, the performances of the comparison methods are shown in Table I. The FRF-Net can achieve 64.17% of mIoU and 76.24% of PA which outperforms the other DL networks on the GID data set.

The results also prove that the proposed FRF-Net can provide better land cover mapping results. The visualized

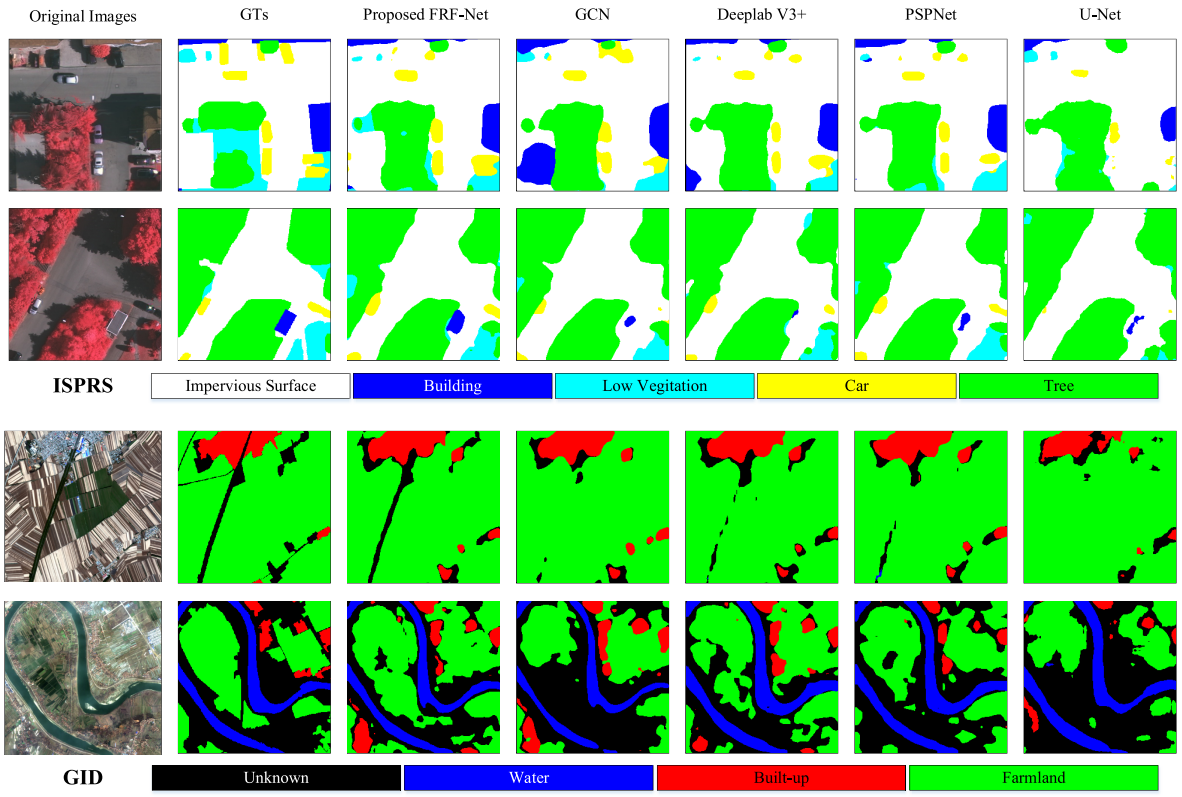

Fig. 5. Land cover mapping results of the method proposed and comparisons of the ISPRS and GID data sets: the first two lines are from ISPRS and the last two lines are from GID.

results are illustrated in the last two lines in Fig. 5. The results generated by our method could possess more refined detail which is especially obvious in the third line.

B. Comparisons on ISPRS Data Set

The FRF-Net is also evaluated and compared with others on the ISPRS data set, described in Section II. Here, the ground truths (GTs) for training and validation are available for 16 of 33 tiles in the data set, and the remaining GTs are used for testing. Then, we split 16 tiles into 11 training tiles and 5 validation tiles. Another 17 tiles are used as the testing data. Specifically, only three channels of RGB are used in all experiments, and the IR channel and digital surface models (DSM) channel are not applied.

The visualized results are illustrated in the first two lines in Fig. 5. The detail per-class results of the proposed FRF-Net and the comparisons are shown in Table II. The normalized confusion matrixes of all methods are illustrated in Fig. 6. Here, related to six land cover categories, they are called IS, BG, tree (T), building (B), car (C), and LV. The method this letter proposed achieved 65.74% of mIoU and 86.04% of PA. As shown in the last two lines in Table II, we can see that due to the imbalance of the per-class pixel number, the single class indexes of BG and C perform badly. Therefore, we calculate the frequency of the per-class pixels to weight the loss function. The less pixel classes of BG and C would be improved by about 8% of PA. Furthermore, from Fig. 5 and Table II performance comparisons, we can also see that the proposed FRF-Net could produce more accurate classification results of the land cover. It is worthy to be mentioned that the FRF-Net surpasses others, such as GCN, with the same backbone of ResNet101. As shown in Table III, compared with PSPNet or

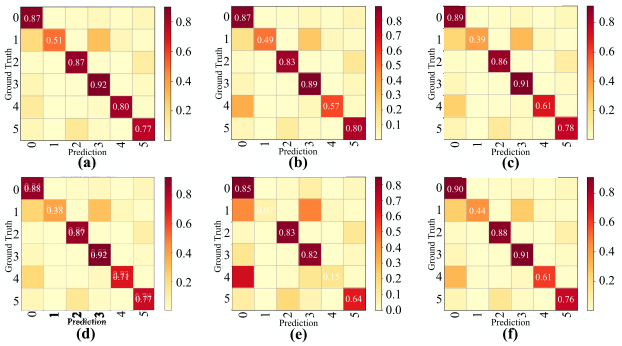

Fig. 6. Normalized confusion matrix of (a) FRF-Net (with loss balance). (b) GCN. (c) PSPNet. (d) FRF-Net (without loss balance). (e) U-Net. (f) DeepLab. Class labels are as follows: 0 = ISs, 1 = BG, 2 = Tree, 3 = Building, 4 = Car, and 5 = LV.

TABLE I
PERFORMANCES OF DIFFERENT MODELS ON GID DATA SET

Methods	PA(%)	mIOU(%)
Deeplab V3+ [20]	75.33	62.85
GCN [18]	75.34	62.46
PSPNet [19]	75.85	63.48
U-Net [17]	72.07	57.01
Seg-Net [21]	69.91	52.54
Our	**76.24**	**64.17**

Deeplab V3+, whose accuracy is close to our model, the FRF-Net takes a smaller computation cost, which involves time, number of Float-point operations (FLOPs), and parameters' evaluation.

TABLE II
COMPARISON PERFORMANCES OF SIX CATEGORIES OF LAND USES IN ISPRS DATA SET

Methods	IS(%)	BG(%)	T(%)	B(%)	C(%)	LV(%)	PA (%)	mIOU (%)
Deeplab V3+ [20]	**89.77**	43.66	**88.27**	91.10	61.01	76.31	**86.14**	***65.86***
GCN [18]	87.08	***49.14***	82.81	89.27	57.41	**80.44**	84.52	63.02
PSPNet [19]	***89.22***	38.50	86.30	90.99	61.20	***78.04***	85.81	65.07
U-Net [17]	84.83	7.22	82.53	82.15	14.84	63.86	77.58	46.04
FRF-Net	88.19	38.32	***87.43***	**92.44**	***71.24***	76.88	***86.04***	65.74
FRF-Net + loss _balance	86.79	**50.97**	86.90	***92.24***	**80.11**	76.95	85.73	**66.71**

(The per-class results on ISPRS testing set. Bold numbers represent the best score for a class, italic numbers the second best score.)

TABLE III
COMPARISON OF COMPUTATIONAL CONSUMPTIONS

Methods	Time(ms)	Frame(fps)	FLOPs(G)	Params(M)
Deeplab V3+[20]	36.76	27.20	22.17	59.34
GCN[18]	7.35	136.05	15.50	58.94
PSPNet[19]	147.88	6.76	63.81	65.58
U-Net[17]	6.94	144.09	15.23	31.04
Our	17.97	55.65	19.98	48.33

V. CONCLUSION

In this letter, we presented an FRF-Net for land use classification, which depended on an encoding-to-decoding process based on the two different attention mechanisms. The self-attention was set up to build the encoding module to capture the long-range semantic for improving the accuracy of classification. A decoding module based on fusion attention was designed to efficiently fuse the low-level feature with the ensemble feature. Then, compared with the existing methods, our method achieved more outstanding performance on the GID and ISPRS data sets with lower computational complexity and less computation cost.

REFERENCES

[1] W. Chen, X. Li, H. He, and L. Wang, "A review of fine-scale land use and land cover classification in open-pit mining areas by remote sensing techniques," *Remote Sens.*, vol. 10, no. 1, p. 15, Dec. 2018. doi: 10.3390/rs10010015.

[2] J. R. B. Fisher, E. A. Acosta, P. J. Dennedy-Frank, T. Kroeger, and T. M. Boucher, "Impact of satellite imagery spatial resolution on land use classification accuracy and modeled water quality," *Remote Sens. Ecol. Conservation*, vol. 4, no. 2, pp. 137–149, Jun. 2018. doi: 10.1002/rse2.61.

[3] N. Joshi *et al.*, "A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring," *Remote Sens.*, vol. 8, no. 1, p. 70, Jan. 2016. doi: 10.3390/rs8010070.

[4] C. Sukawattanavijit, J. Chen, and H. Zhang, "GA-SVM algorithm for improving land-cover classification using SAR and optical remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 284–288, Mar. 2017. doi: 10.1109/LGRS.2016.2628406.

[5] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, "A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 7, no. 2, pp. 491–502, Feb. 2014. doi: 10.1109/JSTARS.2013.2292032.

[6] L. Wan, H. Zhang, G. Lin, and H. Lin, "A small-patched convolutional neural network for mangrove mapping at species level using high-resolution remote-sensing image," *Ann. GIS*, vol. 25, no. 1, pp. 45–55, Jan. 2019. doi: 10.1080/19475683.2018.1564791.

[7] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–70, Nov. 2018.

[8] Y. Liu, Q. Ren, J. Geng, M. Ding, and J. Li, "Efficient patch-wise semantic segmentation for large-scale remote sensing images," *Sensors*, vol. 18, no. 10, p. 3232, Sep. 2018.

[9] M. Gholoobi and L. Kumar, "Using object-based hierarchical classification to extract land use land cover classes from high-resolution satellite imagery in a complex urban area," *J. Appl. Remote Sens.*, vol. 9, no. 1, May 2015, Art. no. 096052. doi: 10.1117/1.JRS.9.096052.

[10] H. Lin, Z. Shi, and Z. Zou, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," *Remote Sens.*, vol. 9, no. 5, p. 480, 2017. doi: 10.3390/rs9050480.

[11] X.-Y. Tong *et al.*, "Learning transferable deep models for land-use classification with high-resolution remote sensing images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018.

[12] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," May 2018, *arXiv:1805.02091*. [Online]. Available: https://arxiv.org/abs/1805.02091

[13] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017. doi: 10.1109/LGRS.2017.2681128.

[14] J. Li, X. Huang, and J. Gong, "Deep neural network for remote sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, to be published. doi: 10.1093/nsr/nwz058.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[18] P. Chao, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large Kernel matters–improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4353–4361.

[19] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[20] 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.*, Sep. 2018, pp. 801–818.

[21] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2015, pp. 1–14.

[22] X. Huang, X. Han, S. Ma, T. Lin, and J. Gong, "Monitoring ecosystem service change in the City of Shenzhen by the use of high-resolution remotely sensed imagery and deep learning," *Land Degradation Develop.*, vol. 30, no. 12, pp. 1490–1501, Jul. 2019. doi: 10.1002/ldr.3337.