Richer U-Net: Learning More Details for Road Detection in Remote Sensing Images

Yifan Zao, Zhenwei Shi, *Member, IEEE*

***Abstract*—Road detection in remote sensing images has been an important research topic in the past few decades. However, with complex backgrounds and occlusion of vehicles and trees, it is difficult for most road detection methods to obtain complete and accurate results. There will be a large number of error and omission detections in such complex scenes due to the poor utilization of detailed information. Therefore, in this paper, we propose a novel road detection method called Richer U-Net, which alleviates this problem by designing two detail enhancement strategies. Firstly, considering that convolution operation will cause the loss of detailed information in the feature map, an enhanced detail recovery structure (EDRS) is introduced to make full use of those lost information. It combines the output of each convolutional layer at the same level for the detail recovery of decoding network, leading to more accurate segmentation results. Secondly, an edge-focused loss function is proposed to guide the network to pay more attention to road edge area. By adding an enhancement factor, the pixels closer to edge will contribute more loss. Corresponding experiments are conducted on two public datasets respectively, and it can be shown that our method effectively improves final detection results.**

***Index Terms*—Remote sensing, road detection, deep learning, detailed information.**

I. INTRODUCTION

ROAD detection from remote sensing images is a very challenging research topic. It is widely used in urban transportation, disaster management and geographic information updating, and of great significance in our daily life. This task generally needs to extract the pixels of road areas, or label the pixels of road centerlines for the construction of the road network. Since it is time-consuming to manually label the road areas, accurate automatic road extraction has become an urgent problem to be solved.

Compared with natural images, remote sensing images have complex backgrounds, making it difficult to obtain accurate segmentation results. At the same time, due to the occlusion of trees, shadows and vehicles, the results may contain many error and omission detections. In the past few decades, a large number of methods [1] have been proposed to solve these problems. Among them, machine learning methods are relatively common at present, such as [2], [3]. These methods need to manually extract representative road features, and then make pixel-wise classifications using SVM or other classifiers. After obtaining road areas, morphological thinning or non-maximum suppression [4] is generally exploited to extract road centerlines.

Recently, as deep learning has achieved great success in the field of semantic segmentation [5]–[7], some deep networks begin to be applied for road detection [8]–[10]. Through learning from large amounts of data, the networks can automatically extract more effective features, thereby obtaining more accurate segmentation results. Zhang et al. [8] designed a road extraction network, which combines the strengths of residual learning and U-Net. Cheng et al. [9] proposed a cascaded convolutional network to deal with road segmentation and centerline extraction tasks simultaneously. The performance of these two methods both exceeds that of traditional road detection algorithms.

However, despite the improvements in current road detection methods, there are still many unavoidable error and omission detections, especially in the road edge areas. In the process of extracting features, the convolutional network continuously loses the details. Without these details, it is difficult to obtain accurate segmentation results, which can lead to a large number of burrs and deletions in final centerline extraction results.

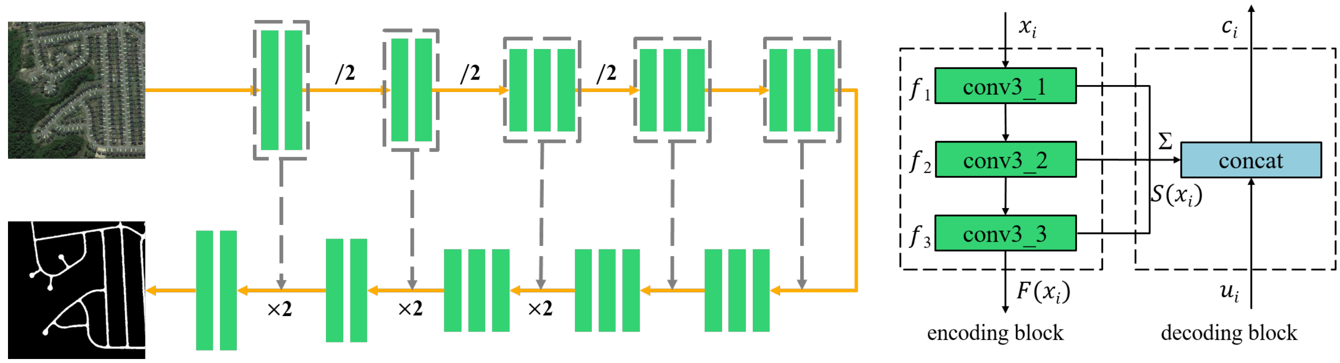
To solve the problems mentioned above, in this letter, we propose a novel method, Richer U-Net, which retains and learns richer details than other methods. Our network uses an encoder-decoder architecture similar to U-Net [6]. The decoder network preserves more details by fusing features from all convolutional layers at the same level. Meanwhile, considering the details of road edge area are more difficult to learn, we design a loss function that pays more attention to the learning of the road edge region, so that the network can get better segmentation results in the border area. Our work mainly has the following two contributions:

- An improved encoder-decoder network, Richer U-Net, is proposed, which introduces an enhanced detail recovery structure (EDRS) to preserve richer features for more accurate segmentation results.
- In order to enable the network to learn more edge detail features, we propose an edge-focused loss function to promote the segmentation results of road edge region by paying more attention to these areas.

The remainder of this letter is as follows. Section II introduces our method in detail, including the enhanced detail recovery structure (EDRS) and the edge-focused loss function. Section III shows the experimental results and related analysis, and Section IV is the conclusion.

The work was supported by the National Key R&D Program of China under the Grant 2019YFC1510905, the National Natural Science Foundation of China under the Grant 61671037 and the Beijing Natural Science Foundation under the Grant 4192034. (Corresponding author: Zhenwei Shi)

Yifan Zao (e-mail: zaoyifan@buaa.edu.cn) and Zhenwei Shi (Corresponding Author, e-mail: shizhenwei@buaa.edu.cn) are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China.



(a) Overall Architecture (b) EDRS

Fig. 1. The network structure of Richer U-Net. (a) Overall Architecture. /2 represents the pooling operation, and ×2 represents the upsampling operation. The area of gray dotted line stands for EDRS; (b) Enhanced Detail Recovery Structure (EDRS). Take the third level of convolutin layers as an example, the structure includes an encoding block and a decoding block, and the output of the three convolutional layers is fused for the detail recovery in decoding block.

II. METHODOLOGY

In this section, we explain the overall architecture of Richer U-Net first, and then introduce the two strategies of EDRS and edge-focused loss separately.

A. Richer U-Net Architecture

In convolutional neural networks, shallow features often contain more detailed information, such as edges and textures, while deep features are more informative in abstract sematic information. Therefore, most of semantic segmentation methods use strategies that integrate shallow features to obtain more refined segmentation results. For road detection from remote sensing images, this seems to be more important. Complex backgrounds and small road areas lead to the detailed information having a great influence on the segmentation results. In order to solve the problem of lack of detailed information, in this paper, we have adopted an encoder-decoder structure that combines more deep and shallow features to make the output of the network contain more detailed information. As shown in Fig.1, the decoding block merges more detailed features into the feature map during the upsampling process, so we can get

$$c_i = concat(u_i, S(x_i)) \tag{1}$$

where c_i represents the output of the ith decoding block, u_i is is the feature map after upsampling, $S(x_i)$ is the supplementary details of the corresponding encoding block, *concat* denotes the concatenation operation.

Our baseline network uses the first 13 convolutional layers of VGG16 [11] as encoder and a corresponding symmetric structure is adopted as decoder. The encoder has a downsampling multiple of 32. Considering the road area in remote sensing images is generally very narrow, if the downsampling multiple is too large, it will not only affect the segmentation accuracy, but also cause the missing of small-sized roads. By changing the stride of the latter two pooling layers to 1, the entire network downsampling multiple is changed from the original 32 times to 8 times, thus obtaining a feature map with more detailed information. As shown in Fig.1, it is the overall network structure diagram of Richer U-Net. Each convolution layer is followed by a batch normalization layer [12] and a relu activation layer [13]. These convolutional layers can be divided into five levels according to the pooling layer, each level containing 2 or 3 convolutional units, which are used to extract adjacent scale features. At the same time, we use the bilinear interpolation method to upsample feature maps, which is beneficial to obtain a denser feature map, thus ensuring the uniformity of segmentation results. After getting the road segmentation results, a morphological skeleton method is used to get the road centerline.

In order to improve the segmentation results, most of road detection networks currently adopt an encoder-decoder structure, such as [6], [8]. However, there exist many differences between our network and other networks. On the one hand, our network pays more attention to the flow of details, and it is more efficient for the fusion of features both in space and channels and helps to eliminate semantic differences [14]. On the other hand, a modified structure is introduced, which uses a more direct way to facilitate the fusion of different features, to make the features of each convolutional layer in the encoder be fully utilized. This structure is different from the residual block, and there is no identity mapping in the module. Instead, it makes full use of more features in a reasonable way and improves the diversity of features, thereby retaining the details lost during the convolution process.

B. Enhanced Detail Recovery Structure (EDRS)

It is critical to improve the details of the deep features for road detection tasks. Liu et al. [15] pointed out that the information obtained by different convolution layers gradually becomes coarser. In each block, the useful details contained in the previous convolutional layer are lost in the last layer, while traditional networks only use the output of the last layer to recover the details. Therefore, in order to make full use of the details, an enhanced detail recovery structure (EDRS) is introduced, as shown in Fig.1. It can be seen that in the past only the final output of each block was used for the decoding network to restore the details. By adding EDRS, the outputs of all convolution layers in each block are fused and then used for the detail recovery of decoding network. The supplementary detailed information in the former case is defined as

$$S(x_i) = F(x_i) = f_3(f_2(f_1(x_i))) \tag{2}$$

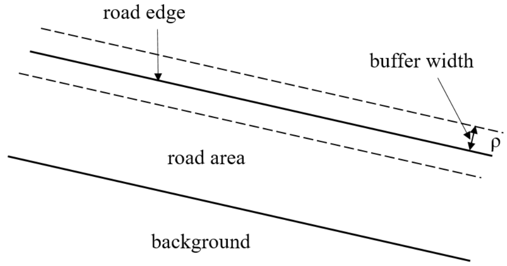


Fig. 2. The schematic diagram of edge-focused loss. The solid line represents the edge of the road, and between the two solid lines is the road area. We set the buffer width with ρ, and the area between the two dashed lines is the region where our loss focuses on learning.

where $F(x_i)$ denotes the output of the ith encoding block. After adding EDRS, it can be defined by

$$S(x_i) = f_1(x_i) + f_2(f_1(x_i)) + f_3(f_2(f_1(x_i))) \tag{3}$$

where f_1, f_2, f_3 respectively represent the equivalent functions of three convolutional layers in the encoding block. From the perspective of information flow, there is less information lost in the network, so the details of the segmentation results will be more accurate.

Unlike RCF [15] for edge detection, EDRS is designed for road detection tasks. The convolution layer before sum operation is removed. Instead, the sum operation is directly applied to the original features, which can avoid more information loss. Also, the summation result is concatenated with the previous high-level feature, and then progressive feature fusion and dimensionality reduction are performed. In fact, for road detection, this kind of progressive approach is more conducive to feature fusion. After adding EDRS, the original network only have five more sum operations, and there is no increase in other parameters.

C. Edge-Focused Loss Function

At present, most semantic segmentation methods use cross-entropy loss for training, which focuses on the classification correctness of each pixel in the optimization process and has the same weight for pixels in different regions.

However, for the binary classification problem like road detection in remote sensing images, both road and background areas are easy to learn. What is difficult to learn is their border area, which contains more useful information, and insufficient attention will lead to the edge areas of segmentation results being too rough. In response to this problem, we propose a loss function that can focus on the learning of road boundary areas. By increasing the weight of the loss in the edge region, our network can learn more details. Thus, our edge-focused loss L_{ef} can be defined as

$$L_{ef}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\Big[(1+g(d_i))\,\mathbf{1}\{y_i=j\}\log(p_j(x_i|\boldsymbol{\theta}))\Big] \tag{4}$$

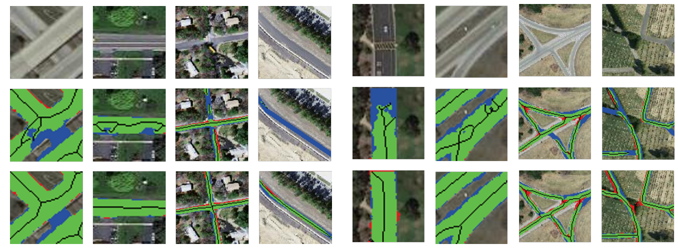
where $\boldsymbol{\theta}$ represents the network parameters, N is the number of pixels, C is the number of categories. x_i is the ith input pixel, while y_i is its corresponding label. $\mathbf{1}\{y_i = j\}$ is an indicator function, and $p_j(x_i|\boldsymbol{\theta})$ is the output of the softmax layer corresponding to the jth class. $g(d_i)$ is a reinforcement factor, it is defined by

$$g(d_i) = \begin{cases} \alpha e^{-\frac{d_i}{\rho}} & 0 \le d_i < \rho \\ 0 & d_i \ge \rho \end{cases} \tag{5}$$

where d_i represents the distance from the ith pixel to the nearest road edge pixel, α and ρ are two coefficients, of which α determines the degree of enhancement for learning effect in the boundary region, and ρ is a buffer width that determines the buffer region to be enhanced, as shown in Fig.2.

As can be seen from the equation above, our proposed edge-focused loss does not change the loss of pixels outside the road edge buffer area, but by adding a variable coefficient, the loss of edge area is enhanced, and the closer the pixel is to the edge, the more enhancement is obtained.

In order to reduce the amount of computation in the process of calculating loss, we use the method of morphological dilation to get the distance to the nearest road edge. For the input binary image, we first perform the contour extraction operation to get the road edge, and then continuously use the disk with radius of 1 for the dilation operation. Meanwhile, we set distance value of the added pixel after the kth dilation to k, and the edge pixel value is set to 0, thereby obtaining a distance map. It can be seen that the distance from the pixel to the edge obtained by the dilation is the sum of the lateral distance and the longitudinal distance, so d_i is actually the manhattan distance.



(a) EDRS (U-Net) (b) edge-focused loss (SegNet)

Fig. 3. Visual strategy ablation analysis. The first row is the original image, the second row is the road detection results without EDRS or edge-focused loss, the third row is the results with EDRS or edge-focused loss.

III. EXPERIMENTS

In order to verify the effectiveness of our method, in this section, we conduct experiments on two public datasets respectively, and compare the results with other methods.

A. Datasets

1) Google Earth Dataset: The dataset is constructed by Cheng et al. [9]. It contains a total of 224 remote sensing images, of which the training set contains 180 images, the validation set contains 14, and the test set contains 30. The resolution of images is 1.2m, and the average width of road is about 15 pixels.

TABLE I
HYPERPARAMETRIC EXPERIMENT RESULTS ON TWO VALIDATION SETS, WHERE THE METRICS IN BOLD ARE THE BEST.

Google Earth	$\alpha = 1$	ρ	1	2	3	4	5	6
		IOU	0.8422	0.8503	0.8508	0.8566	**0.8621**	0.8612
	$\rho = 5$	α	1	2	3	4	5	6
		IOU	0.8621	0.8625	0.8627	**0.8632**	0.8625	0.8614
Massachusetts	$\alpha = 1$	ρ	1	2	3	4	5	6
		IOU	0.5734	0.5786	**0.5812**	0.5765	0.5742	0.5724
	$\rho = 3$	α	1	2	3	4	5	6
		IOU	0.5812	0.5825	0.5850	**0.5863**	0.5835	0.5812

2) Massachusetts Dataset: The dataset is published by Volodymyr Mnih [16] with the resolution of 1m. The training set contains 1108 images, the validation set is 14 and the test set is 49. Since the road in this dataset is narrow with the average width of 7 pixels and there are many occlusions, the difficulty of segmentation is relatively large.

B. Experiment Setting

The deep learning framework Caffe [17] is used to implement our network. The encoder uses the pre-training parameters of VGG16 [11] on the ImageNet for initialization, and the decoder adopts the msra [18] for initialization. In order to train the network better, some data augmentation operations are adopted. For an original image, we first randomly crop 25 images of 320×320 size, and then the cropped images are rotated and mirrored. Also, we set the initial learning rate to 0.01, and it is gradually reduced by a factor of 0.1 every 10000 iterations, and there are a total of 60,000 training iterations. With SGD method for training, the batch size is set to be 4. The momentum is 0.95, and the weight decay is 0.0005. In the test phase, we divide the image into small blocks of 512*512 for processing, and there exist overlapping areas of 20 pixels between these blocks to avoid the border effect. All compared methods have the same experiment settings.

C. Results and Comparisons

1) Evaluation Metric: Similar to most of semantic segmentation methods, IOU is adopted to measure the performance of road detection, which can be defined as

$$IOU = \frac{TP}{TP + FP + FN} \tag{6}$$

where TP represents the true positive, FP represents the false positive, and FN represents the false negative.

2) Hyperparameter Selection: In edge-focused loss, α determines the degree of enhancement and ρ is a buffer width. To some extent, α is insensitive to dataset and ρ should be proportional to the average road width in the dataset. There is no strong correlation between them, so they can be decoupled to get the suboptimal values separately. As shown in Table I, we set $\alpha = 1$ and then explore an optimal value for ρ. For Google Earth, this ρ value is 5, and for Massachusetts, this value is 3. Then we fix the optimal ρ value and explore an optimal value for α. It can be found that for the two datasets, $\alpha = 4$ seems to be a reasonable choice. Also, the average road width of the Google Earth is 15, and the Massachusetts is 7. So ρ is recommended to take about 1/3 to 1/2 of the average road width.

TABLE II
ABLATION EXPERIMENTS ON TWO TEST SETS, WHERE THE HORIZONTAL LINE INDICATES THAT THE EXPERIMENT CANNOT BE PERFORMED. *EDRS* REPRESENTS THE NETWORK WITH THE ENHANCED DETAIL RECOVERY STRUCTURE, *EF* STANDS FOR TRAINING WITH THE EDGE-FOCUSED LOSS.

Datasets	EDRS	EF	SegNet [5]	CasNet [9]	U-Net [6]	Baseline
Google Earth			0.8955	0.8955	0.9104	0.9128
	✓		—	—	0.9122	0.9146
		✓	0.8960	0.8976	0.9125	0.9134
Massachusetts			0.5542	0.5481	0.5842	0.5955
	✓		—	—	0.6074	0.6139
		✓	0.5752	0.5621	0.6125	0.6098

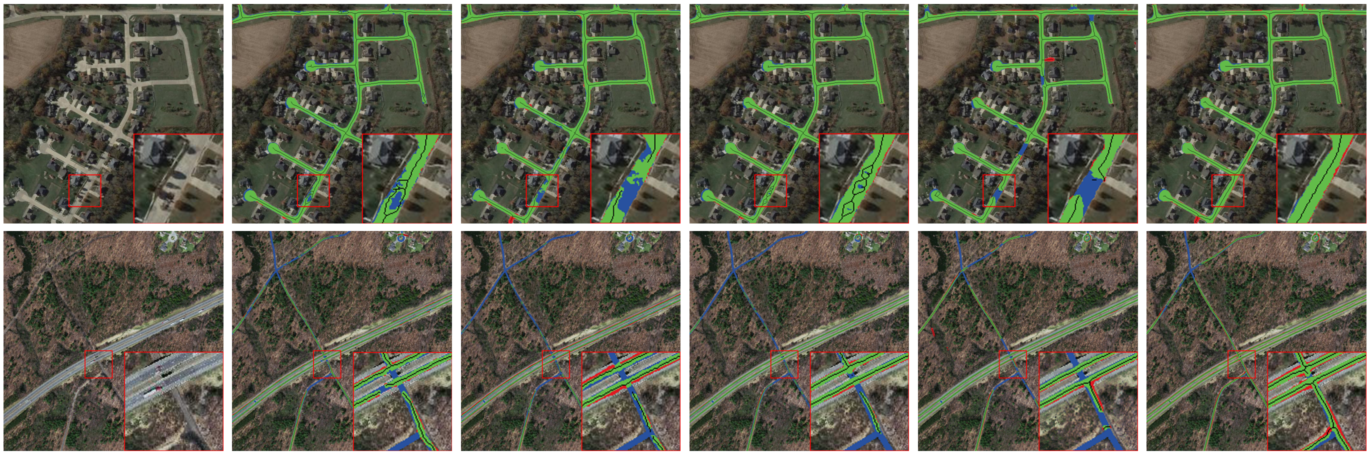
TABLE III
COMPARISON EXPERIMENTS ON TWO TEST SETS, WHERE THE METRICS IN BOLD ARE THE BEST. OURS REPRESENTS OUR BASELINE WITH *EDRS* AND EDGE-FOCUSED LOSS.

Datasets	SegNet [5]	CasNet [9]	U-Net [6]	ResUnet [8]	Ours
Google Earth	0.8955	0.8955	0.9104	0.8982	**0.9154**
Massachusetts	0.5542	0.5481	0.5842	0.6181	**0.6277**

3) Ablation Analysis: For a fairer comparison, we combine VGG16 with a symmetrical decoding structure as our baseline. Since SegNet and CasNet use the Max-pooling Indices for upsampling, EDRS cannot be applied to these two networks, but edge-focused loss can. Thus, we separately studied the impact of these two strategies on different networks. As shown in Table II, EDRS is applied to U-Net and our baseline respectively and has stable improvements for both backbones. Also, experiments about edge-focused loss are conducted on SegNet, CasNet, U-Net and our baseline respectively. It can be seen that for different networks, the results using edge-focused loss are better than those using cross entropy loss. The above results fully demonstrate the validity of our strategies and hyperparameters for different backbones. Combined with the visual results in Fig.3, it can be found that EDRS and edge-focused loss can indeed improve the details of the segmentation and avoid many omission detections and discontinuities, getting more uniform road areas and smoother centerlines.

4) Comparison Experiments: As shown in Table III, it can be seen that our method has achieved the state-of-the-art results. Since Google Earth dataset is relatively simple, our improvement seems to be small. But in Massachusetts dataset, our approach has greatly improved the results. In fact, the roads in Massachusetts dataset are much smaller, which also reflects the effectiveness of our method to improve the road detection results by retaining and learning more details. Also, the roads in Massachusetts are regular and the MSE loss helps to learn this regular structural information, which lead to ResUnet performing well on Massachusetts but getting poor results on Google Earth.

As shown in Fig.4, we adopt a more intuitive way to show the ground truth and segmentation results. For Google Earth dataset, due to the occlusion of the shadows and trees,



(a) Original Image (b) SegNet (c) CasNet (d) U-Net (e) ResUnet (f) Ours

Fig. 4. Visual comparision of different road detection algorithms. The first row is on Google Earth dataset, the second row is on Massachusetts dataset. The area inside the red box is partially enlarged. (a) Original Image; (b) SegNet [5]; (c) CasNet [9]; (d) U-Net [6]; (e) ResUnet [8]; (f) Ours. In the result images, green represents true positive (TP), red represents false positive (FP), blue represents false negative (FN), black represents the centerline.

there is an obvious omission detection at the lower left for SegNet and CasNet, as is shown in red box, and their edge areas of the segmentation are too rough. The omission detection is slightly avoided for U-Net. ResUnet has many obvious omission detections, which is likely due to the MSE loss. Our method completely avoids the bottom left miss, meanwhile with a finer edge and a smoother centerline. For Massachusetts dataset, there are large amounts of omission detections in SegNet, CasNet, U-Net and ResUnet. In our method, more omissions are avoided, and a relatively complete road centerline is obtained.

IV. CONCLUSION

In this letter, Richer U-Net is proposed for road detection in remote sensing images. By learning and retaining more details with EDRS and edge-focused loss, our network finally obtains more accurate segmentation results. The experiment results show that the detailed information contributes to the identification of the road pixels and plays an important role in the road segmentation task. However, for images with more complex backgrounds, the road detection results are still not ideal. Meanwhile, the improvement of details can not avoid all error and omission detections. In the future work, we hope to propose a more effective road detection method to solve these problems by combining tracking strategies.

REFERENCES

[1] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, and P. Eklund, "A review of road extraction from remote sensing images," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 3, no. 3, pp. 271–282, 2016.

[2] S. Das, T. T. Mirnalinee, and K. Varghese, "Use of salient features for the design of a multistage framework to extract roads from high-resolution multispectral satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3906–3931, 2011.

[3] Q. Zhang and I. Couloigner, "Benefit of the angular texture signature for the separation of parking lots and roads on high resolution multi-spectral imagery," *Pattern Recognition Letters*, vol. 27, no. 9, pp. 937–946, 2006.

[4] G. Cheng, F. Zhu, S. Xiang, and C. Pan, "Road centerline extraction via semisupervised segmentation and multidirection nonmaximum suppression," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 4, pp. 545–549, 2017.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[7] B. Pan, Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu, "Coinnet: Copy initialization network for multispectral imagery semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 816–820, 2019.

[8] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[9] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3322–3337, 2017.

[10] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "Roadnet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2043–2056, 2018.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[12] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on International Conference on Machine Learning*, 2015, pp. 448–456.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[14] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 78–95, 2018.

[15] Y. Liu, M. M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5872–5881.

[16] V. Mnih, "Machine learning for aerial image labeling," *Doctoral*, 2013.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Acm International Conference on Multimedia*, 2014, pp. 675–678.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.