Extracting Road Maps from High-Resolution Optical Remote Sensing Images via U-Dense Network

Jian Kang, *Student Member, IEEE*

***Abstract*—Extracting road maps from high-resolution optical remote sensing images have received much attention recently, especially with the rapid development of deep learning methods. The primary goal of this task is to distinguish between road and background pixels in the associated images, which can be defined as a binary segmentation problem. In this letter, we developed a framework for this task based on the proposed network—U-Dense—and multi-scale model ensembling. U-Dense takes advantage of the powerful encoding capability of DenseNet and the architecture of fusing multi-level feature maps to learn the binary segmentation. Compared with other state-of-the-art networks, more prominent performance can be achieved by the proposed network under several evaluation metrics, including F1 and Intersection over Union (IoU) scores. Moreover, we demonstrate that further improvement of the performance can be obtained via ensembling of the models trained at multiple scales. The superiority of the proposed approach is demonstrated by the experiments carried out on a DeepGlobe Road Extraction Dataset.**

***Index Terms*—Convolutional neural network (CNN), road extraction, high-resolution remote sensing images**

I. Introduction

AUTOMATICALLY generating road maps from images is beneficial to a wide range of application domains, such as autonomous driving, land investigation, and urban planning. One of the most appealing data sources for road map generation is high-resolution remote sensing imagery, owing to its capability of large-area coverage. However, many factors, such as shadows induced by trees and complex road network topology, mitigate the accurate extraction of road maps, which makes this task challenging.

In remote sensing, researchers have put much effort into this task for decades, and various approaches have been proposed [1]–[4]. Wegner *et al.* [5] modeled road networks via constructing higher-order conditional random fields by using superpixel segments and the connected paths among them. Similarly, in [6], by utilizing the assistant data resource—OpenStreetMap (OSM)—the authors formulated road extraction as a parameterized Markov random field problem in terms of the location of the road-segment centerlines as well as their width.

Recently, with the rapid development of deep learning techniques, convolutional neural networks (CNN) have shown great success in tackling the semantic segmentation problem. One of the most popular CNN-based methods is the fully convolutional network (FCN) [7], where an end-to-end semantic segmentation method was proposed based on the design of the fully convolutional layer. By combining low-level and high-level feature maps, Ronneberger *et al.* [8] developed the U-Net architecture, which achieved promising performance on semantic segmentation for biomedical images. In order to improve the real-time processing ability of networks, LinkNet architecture [9] was proposed to reduce the number of network parameters to learn. Taking advantage of the powerful pre-trained VGG encoders and the U-Net architecture, Iglovikov *et al.* proposed TernausNet [10], which was part of the winning solution (first out of 735) in the Carvana Image Masking Challenge [11].

Fig. 1. Examples of road extraction of the proposed approach applied to a DeepGlobe Road Extraction dataset (shown in cyan).

One of the earliest attempts for road extraction based on deep learning was carried out by [12], where restricted Boltzmann machine (RBM) was exploited to learn the mapping from input high-resolution remote sensing images to road labels. Cascaded CNN (CasNet) was introduced to automatically extract road areas and their centerlines in [13]. Zhang *et al.* [14] utilized residual blocks and proposed a deep residual U-Net for road extraction. Given the road segments from CNN, graph-based methods [15], [16] have been utilized for the final decisions in road topology.

In this letter, we introduce a framework of multi-scale model ensembling based on the proposed network—U-Dense—to learn road maps from high-resolution optical remote sensing images. Compared with other state-of-the-art networks, U-Dense demonstrates better performance on a DeepGlobe Road Extraction dataset [17]. It is worth noting that the proposed approach achieves the seventh place (participant ID: deepjoker) out of more than 350 participants during the development

phase of the DeepGlobe Road Extraction Challenge [1].

II. Methodology

A. U-Dense

One of the most prevalent architectures for binary semantic segmentation is U-Net[8], which fuses multi-level feature maps to hierarchically increase the spatial resolution of the output probability map. The encoder exploited in the original U-Net is built by plain neural units, while the feature representation capability of plain neural unit is overtaken by residual neural unit proposed in [18]. More recently, a multi-layer dense block was proposed in DenseNet architecture [19], where each layer is connected to every other layer. Taking advantage of this type of connection, the forward feature propagation ability is strengthened, and the number of parameters can be reduced as the network going deeper. Inspired by this, we seek to investigate whether the U-Net encoders can be made of dense blocks and if an improvement of semantic segmentation can be achieved. To this end, as illustrated in Fig. 2, a U-Dense network is proposed in this section. Its encoder part is mainly composed of four dense blocks and three transition layers, and the decoder part is constructed based on five plain neural units. To be specific, the dense block and transition layer are explained as follows.

1) Dense block: A dense block is mainly composed of multiple layers that are built through their connections between each layer and subsequent layers, as presented in Fig. 3 (Left). Specifically, the output of lth layer can be represented by:

$$\mathbf{x}_l = H_l([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}]), \tag{1}$$

where $[\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}]$ denotes the concatenation of the feature maps provided by all the preceding layers, and $H_l(\cdot)$ refers to the nonlinear mapping function of the layers. Different from the residual block utilized in [18], the dense block exploits the concatenation operator to combine the learned feature maps, which can increase the variation in the input of the following layers and improve efficiency.

2) Transition layer: Due to the concatenation operator utilized inside dense blocks, spatial resolution of feature maps cannot be down-sampled through dense blocks. Therefore, as illustrated in Fig. 3 (Right), transition layers consisting of batch normalization, convolutional and pooling layers are introduced between separate dense blocks for the spatial down-sampling.

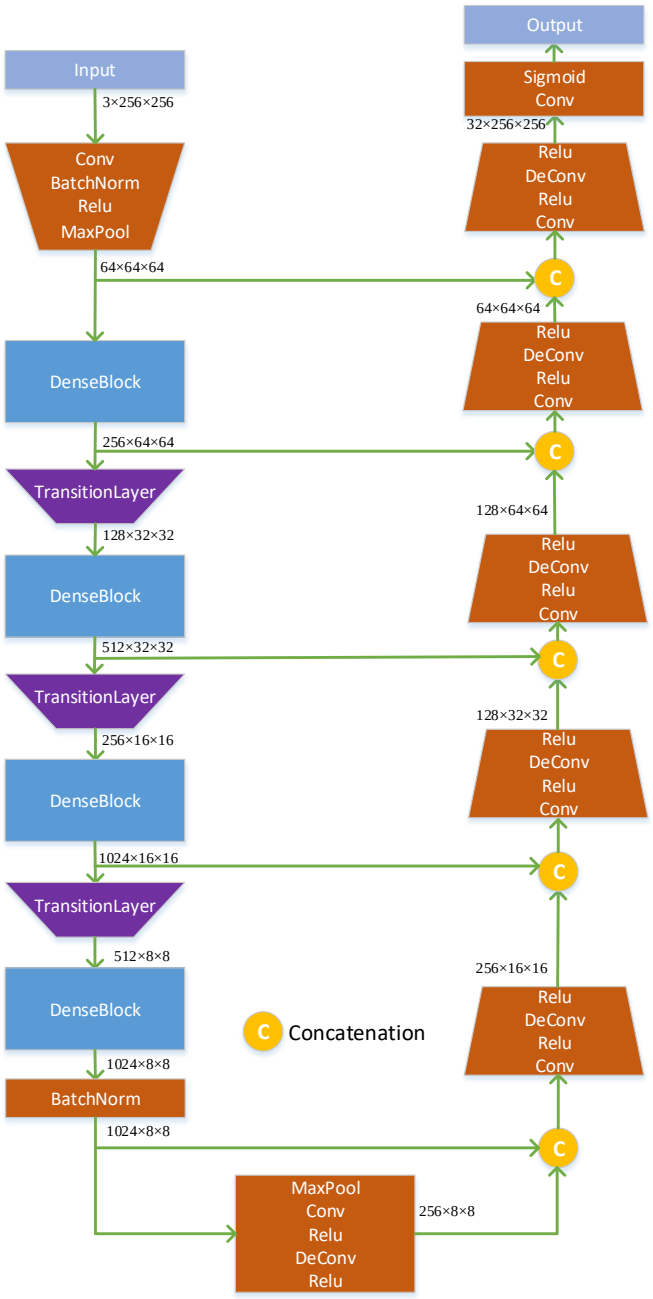


Fig. 2. The proposed architecture of U-Dense. Its encoder part is mainly composed of four dense blocks and three transition layers, and the decoder part is constructed based on five plain neural units.

B. Joint Loss Function

Given the input images $\mathbf{Y}_i$ and the associated ground truth maps $\mathbf{G}_i$, the joint loss function is exploited for learning networks, which combines pixel-wise binary cross entropy (BCE) loss and the Dice coefficient. In particular, the Dice coefficient is determined by the true positive (TP), false positive (FP), and false negative (FN) based on the prediction and ground truth, which can be written as :

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{2}$$

[1]http://deepglobe.org/challenge.html

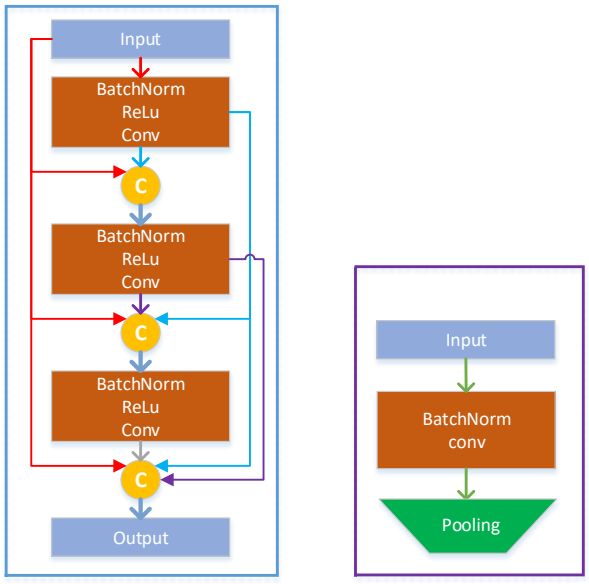
Correspondingly, the joint loss function is formulated as:

$$L = \frac{1}{N}\sum_{i=1}^{N}(\text{BCE}(F(\mathbf{Y}_i), \mathbf{G}_i) + 1 - \text{Dice}(F(\mathbf{Y}_i), \mathbf{G}_i)), \tag{3}$$

where N is the number of images in a batch, and $F(\mathbf{Y}_i)$ represents the output probability map of the trained network, given the input image $\mathbf{Y}_i$.

C. Multi-scale Model Ensembling

Different from other kinds of images, remote sensing images always have large-scale spatial sizes; e.g., 2048×2048 pixels. It



Dense block Transition layer

Fig. 3. Dense block (Left) and Transition layer (Right) utilized in U-Dense. A dense block is mainly composed of multiple layers that are built through their connections between each layer and subsequent layers, and the learned feature maps inside the block are combined by the concatenation operator. The transition layer is exploited for down-sampling the spatial size of feature maps.

may be difficult to directly feed the batches of such images into networks due to the memory limitation of GPU. Therefore, image tiles with smaller spatial sizes are usually sampled from the original dataset to train networks. However, one may not know which parameter of the spatial size is the best for the final decision produced by networks. Furthermore, image tiles with different spatial sizes may capture contextual information at multiple scales. Therefore, a multi-scale model ensembling procedure is conducted based on the proposed U-Dense in this letter. In particular, U-Dense models are trained on image tiles with 256×256 and 512×512 pixels. The decision fusion is carried out by averaging the probability maps produced by the two models, and the binary segmentation can be obtained at a given threshold. In this letter, the threshold is set at 0.5.

III. Experiments

A. Dataset

In this letter, a DeepGlobe Road Extraction dataset [2] is exploited for the experiments. It covers images captured over Thailand, Indonesia, and India. The ground resolution of the image pixels is 50 cm/pixel. During the development phase, the dataset totally contains 1243 high-resolution remote sensing images with a spatial size of 1024×1024 pixels. For the following analysis, we randomly split the whole data into training, validation, and test samples with the percentages of 80%, 10%, and 10%, respectively.

[2]https://competitions.codalab.org/competitions/18467#participate

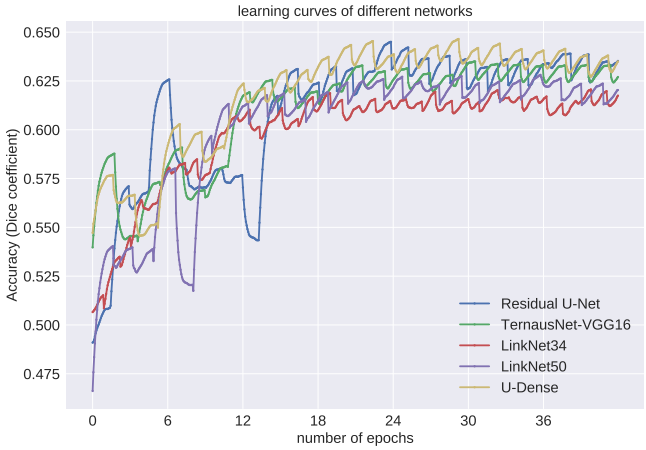


Fig. 4. The learning curves of U-Dense, compared to those of state-of-the-art networks, based on the validation dataset. The accuracy is measured by the Dice coefficient. It is apparent that the proposed network performs better than the others during the training.

B. Training

We trained the proposed network together with several state-of-the-art networks, including LinkNet [9], Residual U-Net [14], and TernausNet [10], on the same dataset. For data augmentation, random crops (256 out of 300 and 512 out of 600), vertical/horizontal flips, and random rotations were adopted for the training samples. The learning rate for all the methods was set at 5×10^{-4} and decayed by a factor of 0.1 for every 12 epochs. The total number of epochs was 40. The Adam optimizer [20] was exploited for minimizing the joint loss. Online hard negative mining procedure was also adopted for training the networks. For monitoring the training status, we calculated the Dice coefficient of the predictions based on the validation dataset. As illustrated in Fig. 4, learning curves of different networks trained on image tiles, with a size of 256×256 were demonstrated. Results indicate that U-Dense behaves better than the other networks during training based on the validation data.

C. Evaluation

Based on the trained networks, we first created a visual comparison based on their inferences on the test data with a full spatial size. Since the models were trained on image tiles (256×256 and 512×512), inference on the whole image was conducted in a sliding window manner. To mitigate the boundary effects, overlapping patches were cropped and fed into the networks. Moreover, to avoid inaccurate inference on the areas near boundaries, the central parts of the probability maps obtained from the networks were further cropped. By stitching all the tiles together and averaging the overlapping areas, the final prediction on the whole image could be obtained. As illustrated in Fig. 5, we demonstrate the produced road maps based on the comparison of networks. Most road areas can be classified correctly by all methods. However, for the areas indicated by red rectangles, the proposed method shows robustness. For example, as present in the third row, the paved areas between small buildings are wrongly recognized

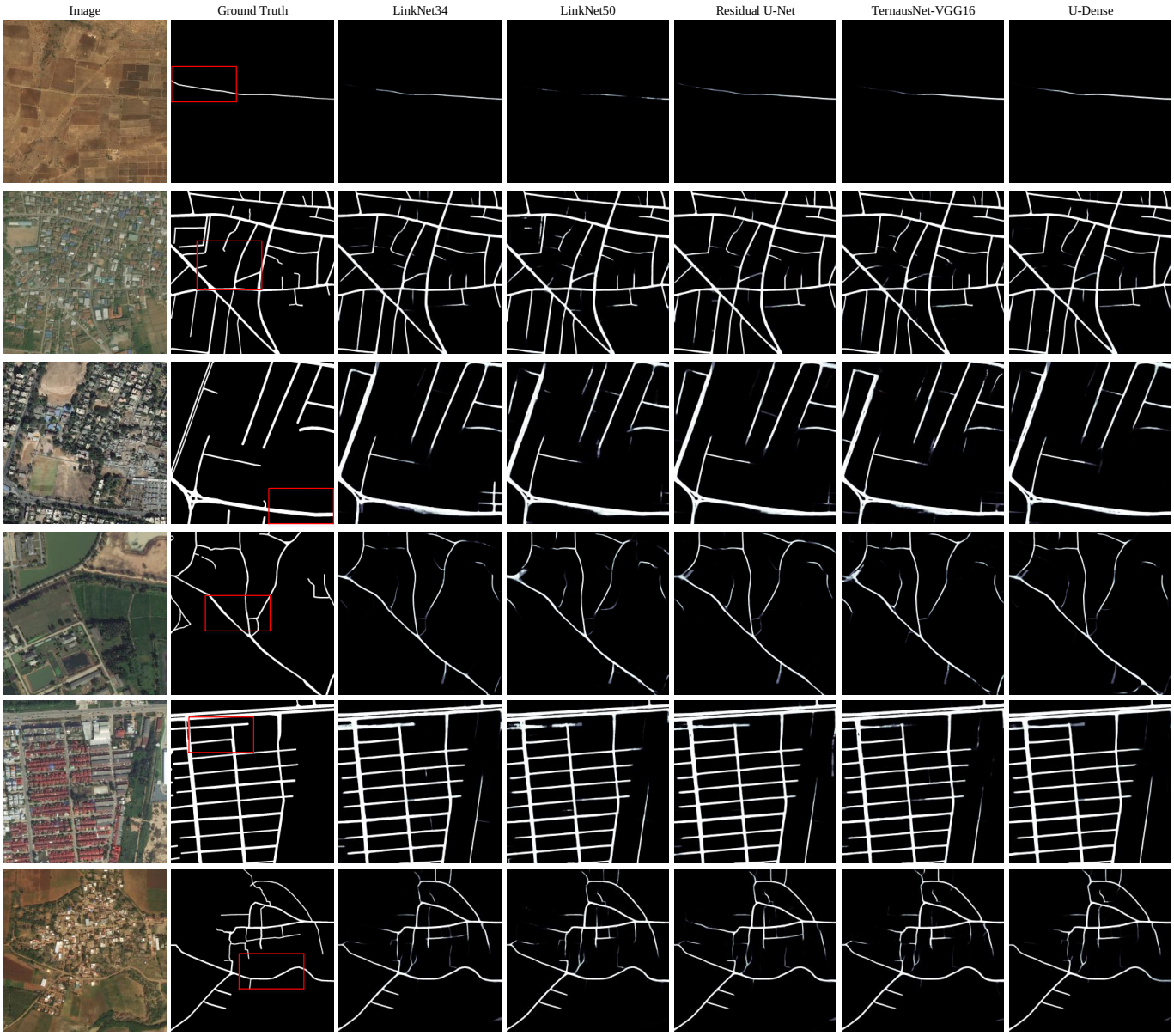


Fig. 5. Visualized comparison of the networks' inference on the test images with a full spatial size. Most road areas can be correctly categorized by all the methods. However, as indicated by the areas within red rectangles, U-Dense demonstrates better robustness than the others. For example, in the third row, the paved areas between small buildings are wrongly recognized by the other networks. In a comparison, our network can correctly classify them as background.

by the other networks; e.g., LinkNet34. In a comparison, U-Dense can accurately categorize them as background. Moreover, for a quantitative comparison, we evaluated the inference performances using two metrics, F1 and IoU scores, and presented them in Table I. Specifically, the metrics are defined as follows:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{5}$$

Consistent with the previous analysis, higher scores of F1 and IoU can be achieved by the proposed network than by the other networks. Besides, based on the ensembling of models trained on different scales, the associated performance can be further improved, as displayed in Fig. 6. The plausible reasons may be that the robustness, with respect to the patch size for the training, can be enhanced, and multi-scale contextual information can be considered in the final decision.

TABLE I
F1 AND IOU SCORES OF THE COMPARING NETWORKS ON THE TEST DATASET.

Network	F1	IoU
LinkNet34	0.779	0.651
LinkNet50	0.772	0.643
Residual U-Net	0.779	0.652
TernausNet-VGG16	0.780	0.653
U-Dense	**0.786**	**0.660**

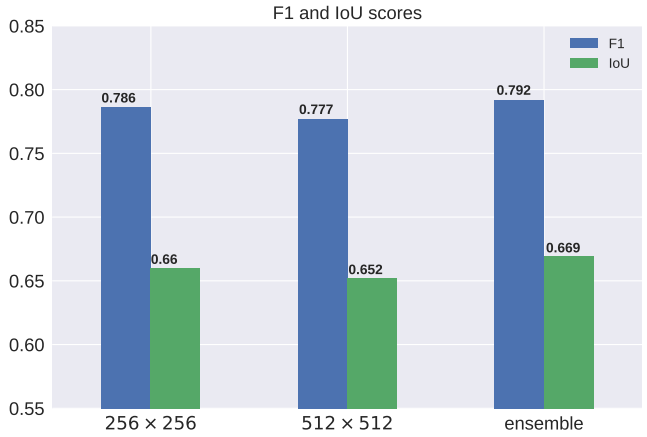


Fig. 6. F1 and IoU scores produced by each single-scale U-Dense network (256×256 and 512×512) and their ensembling. It can be observed that based on ensembling of the models trained on different scales, the associated performance can be improved. The plausible reasons may be that the robustness, with respect to the patch size for the training, can be enhanced, and multi-scale contextual information can be considered in the final decision.

IV. Conclusion

In this letter, a novel network—U-Dense—is proposed for road extraction in high-resolution remote sensing images, which takes advantage of the prominent ability of feature encoding based on dense blocks and the fusion of multi-level feature maps to learn the high-resolution road masks. Based on a DeepGlobe Road Extraction dataset, U-Dense achieves better performance than other state-of-the-art networks according to the metrics of F1 and IoU scores. Besides, we also introduce a multi-scale model ensembling procedure and analyze its performance against each single model. We observe from the experiments that by fusing multi-scale models, more accurate road masks can be obtained.

For the feature work, we would like to investigate whether multi-task learning can be adopted for road extraction. Besides the consideration of pixel-wise category, we seek to combine other information, such as road topology, into the training of networks.

Acknowledgment

The authors would like to thank the Leibniz Supercomputing Center (LRZ) for providing the GPU computing time as well as NVIDIA Corporation for its donation of the Titan X Pascal GPU, which was used in this research.

References

[1] S. Hinz and A. Baumgartner, "Automatic extraction of urban road networks from multi-view aerial imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 1-2, pp. 83–98, 2003.

[2] J. Hu, A. Razdan, J. C. Femiani, M. Cui, and P. Wonka, "Road network extraction and intersection detection from aerial images by tracking road footprints," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4144–4157, 2007.

[3] X. Huang and L. Zhang, "Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines," *International Journal of Remote Sensing*, vol. 30, no. 8, pp. 1977–1987, 2009.

[4] C. Unsalan and B. Sirmacek, "Road network detection using probabilistic and graph theoretical methods," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4441–4453, 2012.

[5] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "Road networks as collections of minimum cost paths," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 108, pp. 128–137, 2015.

[6] G. Mattyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1689–1697.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[8] 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*. Springer, 2015, pp. 234–241.

[9] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," *arXiv preprint arXiv:1707.03718*, 2017.

[10] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[11] Kaggle: Carvana Image Masking Challenge, https://www.kaggle.com/c/carvana-image-masking-challenge, 2017.

[12] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *European Conference on Computer Vision*. Springer, 2010, pp. 210–223.

[13] 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.

[14] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, 2018.

[15] I. Demir, F. Hughes, A. Raj, K. Tsourides, D. Ravichandran, S. Murthy, K. Dhruv, S. Garg, J. Malhotra, B. Doo *et al.*, "Robocodes: Towards generative street addresses from satellite imagery," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1486–1495.

[16] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *International Conference on Computer Vision*, vol. 2, no. 4, 2017.

[17] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," *arXiv:1805.06561*, 2018.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.