

PaRK-Detect: Towards Efficient Multi-Task Satellite Imagery Road Extraction via Patch-Wise Keypoints Detection

Shenwei Xie
xieshenwei@bupt.edu.cn

Beijing University of Posts and
Telecommunications, China

Wanfeng Zheng
zhengwanfeng@bupt.edu.cn

Zhenglin Xian
2019213264@bupt.edu.cn

Junli Yang
yangjunli@bupt.edu.cn

Chuang Zhang
zhangchuang@bupt.edu.cn

Ming Wu ✉
wuming@bupt.edu.cn

Abstract

Automatically extracting roads from satellite imagery is a fundamental yet challenging computer vision task in the field of remote sensing. Pixel-wise semantic segmentation-based approaches and graph-based approaches are two prevailing schemes. However, prior works show the imperfections that semantic segmentation-based approaches yield road graphs with low connectivity, while graph-based methods with iterative exploring paradigms and smaller receptive fields focus more on local information and are also time-consuming. In this paper, we propose a new scheme for multi-task satellite imagery road extraction, Patch-wise Road Keypoints Detection (PaRK-Detect). Building on top of D-LinkNet architecture and adopting the structure of keypoint detection, our framework predicts the position of patch-wise road keypoints and the adjacent relationships between them to construct road graphs in a single pass. Meanwhile, the multi-task framework also performs pixel-wise semantic segmentation and generates road segmentation masks. We evaluate our approach against the existing state-of-the-art methods on DeepGlobe, Massachusetts Roads, and RoadTracer datasets and achieve competitive or better results. We also demonstrate a considerable outperformance in terms of inference speed.

1 Introduction

Road extraction from satellite imagery is a long-standing computer vision task and a hot research topic in the field of remote sensing. Extracting roads is challenging due to the complexity of road networks along with the diversity of geographical environments. With

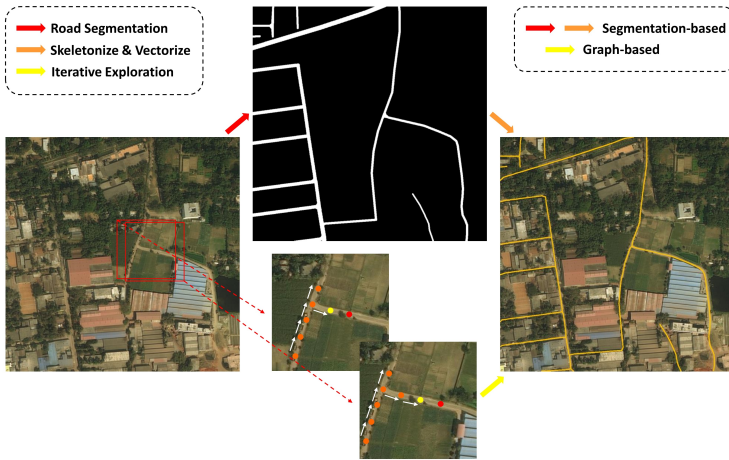


Figure 1: Illustration of segmentation-based approaches and graph-based approaches for road extraction from satellite imagery. Red and orange arrows present segmentation-based approaches that perform road segmentation with skeletonization and vectorization. The yellow arrow stands for the iterative exploration utilized in graph-based approaches.

the rapid development of image interpretation and the cutting-edge architecture of Convolutional Neural Networks (CNNs) designed for computer vision tasks, automatically extracting wide-coverage road graphs from high-resolution aerial images becomes more accurate and efficient. As shown in Fig. 1, prior prevailing works fall into two main categories: pixel-wise semantic segmentation-based approaches and graph-based approaches. Semantic segmentation-based approaches perform binary classification to distinguish road pixels from background pixels and form segmentation maps as an intermediate representation of road graphs. Road networks are constructed by road centerlines and edges through skeletonizing segmentation masks and edge detection. In contrast, graph-based approaches extract road graphs directly by yielding vertices and edges of road networks. The iterative exploration paradigm has been utilized to search for the next move in correspondence to the roads in satellite imagery.

However, extensive experiments reveal that these two approaches both have some drawbacks when it comes to certain evaluation perspectives. Under the circumstances of severe occlusion and geographical environments with high complexity, segmentation-based approaches usually yield road graphs with low connectivity. On the other hand, graph-based approaches benefit from iterative exploring paradigms that overcome connectivity issues, nevertheless, leading to smaller receptive fields and focusing more on local features without a global perspective. Moreover, the iterative exploring paradigm only searches one step at a time which is time-consuming for road graph inference.

According to the above discussions, we propose a new scheme for satellite imagery road extraction, patch-wise road keypoints detection (termed as PaRK-Detect), which is inspired by keypoint detection and combines the inherent superiority in the global receptive field of semantic segmentation-based approaches while ravel out its drawbacks that the absence of connectivity supervision affects road network topology. More specifically, by dividing high-resolution satellite imagery into small patches, the road graphs are partitioned section by section. The framework built on top of D-LinkNet architecture determines whether each

patch contains a road section and if so, then detects the road keypoint in that patch. Establishing the adjacent relationships which represent the connection status between the road patches is also conducted by the framework. In this way, our proposed approaches enable road graph construction in a single pass rather than iterative exploration, in other words, a much faster inference speed becomes reality. Meanwhile, our proposed multi-task network also performs pixel-wise road semantic segmentation to enrich contextual information and enhance information capture capabilities.

Through evaluating our approach against the existing state-of-the-art methods on the publicly available DeepGlobe dataset, Massachusetts Roads dataset, and RoadTracer dataset, we show that our new scheme achieves competitive or better results based on several metrics for road extraction and yields road networks with high accuracy and topological integrity.

Our main contributions can be concluded as follows:

- We propose a patch-wise road keypoints detection scheme that combines road extraction with keypoint detection to construct road graphs in a single pass.
- We propose an efficient multi-task framework that performs road segmentation and road graphs extraction simultaneously.
- We conduct extensive experiments over widely-used datasets and achieve competitive or better results using several road extraction metrics.

2 Related Work

Pixel-wise Semantic Segmentation-based Approaches. With the increasing popularity and dominance of convolutional neural network (CNN) architecture, utilizing the symmetrical [8, 16] or heterogeneous [6, 12, 25] encoder-decoder framework becomes the favorite paradigm for semantic segmentation tasks. Road segmentation is a binary classification task that splits pixels of remote sensing images into road and non-road. Many refined and improved segmentation networks [8, 12, 24, 26] have been proposed for this scene. For example, Buslaev *et al.* [8] proposed a fully convolutional network (FCN) that consists of pretrained ResNet [10] and a decoder adapted from vanilla U-Net [16]. Zhang *et al.* [24] proposed ResUnet, an elegant architecture with better performance that combines the strengths of residual learning and U-Net. Zhou *et al.* [26] proposed D-LinkNet that built with LinkNet [5] architecture and adopts dilated convolution layers in its central part to enlarge the receptive field and ensemble multi-scale features. DeepRoadMapper [12] improves the loss function and the post-processing strategy that reasons about missing connections in the extracted road topology as the shortest-path problem. Although these powerful networks provide a global perspective of road networks and detailed road features, segmentation-based approaches suffer from the absence of connectivity supervision and yield segmentation masks with redundant road information which is meaningless for road graphs construction. Our proposed patch-wise road keypoints detection scheme fully considers the connectivity status and focuses on the road graph while retaining a global perspective.

Graph-based Approaches. Unlike semantic segmentation approaches, graph-based approaches directly extract vertices and edges to construct road graphs. Through exploiting the iterative exploring paradigm, the next moves searched along the road are successively added to the existed road graph. For instance, RoadTracer [2] utilizes the iterative searching

guided by a CNN-based decision function that decides the stop probability and angle vector of the next move, while VecRoad [18] proposes a point-based iterative exploration scheme with segmentation-cues guidance and flexible steps. Apart from these iterative methods, He *et al.* [19] proposed Sat2Graph with a graph-tensor encoding (GTE) scheme which encodes the road graph into a tensor representation. Bahl *et al.* [20] combined an FCN in charge of locating road intersections, dead ends, and turns, and a Graph Neural Network (GNN) that predicts links between these points. Our scheme constructs road graphs in a single pass through detecting patch-wise road keypoints without the iterative exploring paradigm.

Multi-Task Road Extraction Approaches. Road extraction can have different purposes and tasks in different scenarios. City planning and road network updating focus on topology information in road graphs, while emergency response may focus on road damage status that can be evaluated through the segmentation mask. Cheng *et al.* [21] proposed CasNet, a cascaded end-to-end CNN, to simultaneously cope with road detection and centerline extraction. Yang *et al.* [22] proposed a recurrent convolution neural network U-Net (RCNN-UNet) with the multitask learning scheme. Wei *et al.* [23] extracted road surface and centerline concurrently and proposed a multistage framework that consists of boosting segmentation, multiple starting points tracing, and fusion. Our proposed framework conducts road segmentation and road graph construction simultaneously to improve accuracy, robustness, and efficiency.

Keypoint Detection. In recent years, keypoint detection has enjoyed substantial attention and has a promising application prospect in automatic driving and human tracking. Human pose estimation, facial landmark detection, and detection of specific types of object parts are collectively known as keypoint detection. Since the pioneering work DeepPose [24], many DNN-based methods for keypoint detection have been proposed. Basically, constructing ground truth falls into three categories: directly using coordinates of keypoints [25], constructing heatmap [26, 27], and jointly utilizing heatmap and offsets [28]. The stacked hourglass network (SHN) [29] has been widely used as its repeated bottom-up, top-down processing with successive pooling and upsampling ensures different features are extracted from feature maps of different scales to achieve the best performance. Cai *et al.* [30] proposed Residual Steps Network (RSN) that aggregates intra-level features to retain rich low-level spatial information. Additionally, Pose Refine Machine (PRM), an efficient attention mechanism, has been proposed to make a trade-off between the local and global representation for further refinement. Shi *et al.* [31] proposed a novel improved SHN with the multi-scale feature aggregation module designed for accurate and robust facial landmark detection. Then, an offset learning sub-network is adopted to refine the inferred keypoints. Inspired by these works, our proposed approach performs patch-wise offset learning for road keypoints.

3 Methods

This section presents our proposed patch-wise road keypoints detection scheme. Section 3.1 defines fundamental elements of road status used in patch-wise road representation and illustrates the patch-wise road keypoints detection scheme in detail. Once the definitions of road representation and the procedures of the scheme are learned, the architecture of the proposed multi-task framework is explained in section 3.2. We also describe novel learning objectives designed for our scheme in section 3.3. Finally, an effective graph optimization

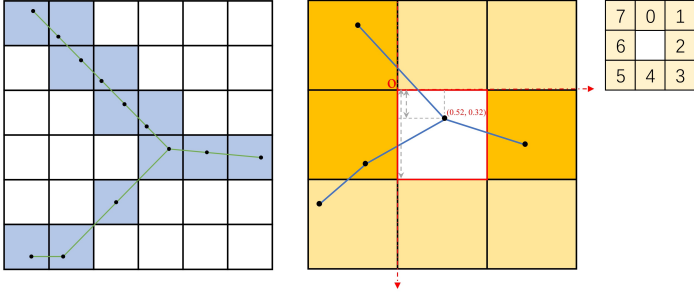


Figure 2: **Illustration of PaRK-Detect scheme.** **Left:** blue patches contain road while white patches are non-road, black dots are road keypoints, and green lines represent links. **Right:** The reference point of relative offset is the upper left corner of a patch. Dark yellow patches are linked with the center patch while light yellow ones are not. We order the eight adjacent patches into numbers 0-7. Here the linked patches are 2, 6, and 7.

strategy is briefly illustrated in section 3.4.

3.1 PaRK-Detect Scheme

As introduced above, our PaRK-Detect scheme detects patch-wise road keypoints and constructs road graphs in a single pass. A high-resolution satellite image (e.g. of size 1024×1024 pixels) can be divided into non-overlapping patches (e.g. 64^2 patches of size 16×16 pixels) as the minimum unit for road representation. Fig. 2 shows the illustration of the PaRK-Detect scheme. Three fundamental elements are defined for road representation in each patch: P , S , and L .

After dividing entire large satellite imagery into N^2 non-overlapping small patches, the road networks are also partitioned section by section by the grids. P stands for the probability of whether a patch contains a road section. If a patch does not cover road, then we should not consider it any further.

For patches that cover road, S stands for the position of patch-wise road keypoint, which is also defined as the most crucial point in a patch. The priority of road intersection is higher than the road endpoint. If there is no intersection or endpoint in a patch that contains a road section, then the keypoint is defined at the geometric center of the road segment in that patch, which is also called the midpoint. We use relative offset to represent the position of the patch-wise road keypoint.

Furthermore, L stands for the link status or connectivity status between these patches. Each patch can be linked to eight patches around itself. If two adjacent patches are linked, then it means there exists a road segment between the road keypoints of them. We use eight probability values to represent link status. A road graph consists of vertexes and edges which are formed by these keypoints and links.

3.2 Framework Architecture

As illustrated by Fig. 3, we designed a multi-task framework for our PaRK-Detect scheme that performs road segmentation and road graph construction simultaneously. Following the encoder-decoder frameworks paradigm, we use ResNet34 [14] pretrained on ImageNet [9]

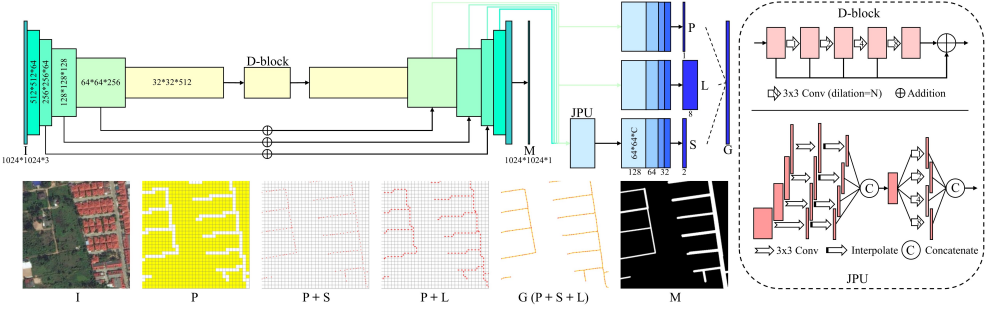


Figure 3: Overview of our proposed multi-task framework architecture. The rectangles are feature maps of different scales. **I**: input satellite image, **P**: patch-wise road probability, yellow patches represent non-road while white patches represent road, **S**: patch-wise road keypoint position, **L**: patch-wise link status, **G**: road graph, **M**: road segmentation mask. Here we just show 32^2 patches out of 64^2 for better presentation.

dataset as encoders to extract high-level semantic information such as road features (e.g. the encoder part downsamples satellite images of size 1024×1024 and outputs feature maps of size 32×32). We also adopt dilated convolution with skip connections same as the bottleneck of D-LinkNet [26] after encoders to increase the receptive field of feature points while keeping detailed information.

For road segmentation, the framework retains the decoder that utilizes the transposed convolution and the skip connection same as LinkNet [5], restoring the resolution of feature map (e.g. from 32×32 to 1024×1024).

For road graph construction, three networks are designed for predicting P , L , and S respectively. These three fundamental elements of patch-wise road representation are indispensable yet sufficient to construct road graphs. Patch-wise probability predicting network performs classification of whether a patch covers road. Patch-wise link predicting network determines the connection status between a patch and its eight adjacent patches. Patch-wise position predicting network locates the patch-wise keypoint in each patch. Both probability and link predicting networks only consist of convolutional layers and batch normalization with feature maps of the same size (e.g. 64×64) and decreasing channels. In order to keep low-level detailed information, position predicting network combines joint pyramid upsampling (JPU) [22] additionally for multi-scale feature fusion. The outputs of these three networks are feature maps (e.g. of size 64×64) with 1, 8, and 2 channels respectively.

3.3 Learning Objectives

To train our multi-task framework, we define a joint learning objective for road segmentation and road graph construction. The joint learning objective can be formulated as follow:

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{graph}$$

Road segmentation loss remains the same as segmentation-based methods [26] that evaluate the difference between labels and predicted masks with binary cross entropy (BCE) and dice coefficient loss. For road graph construction, we define the training loss functions \mathcal{L}_P ,

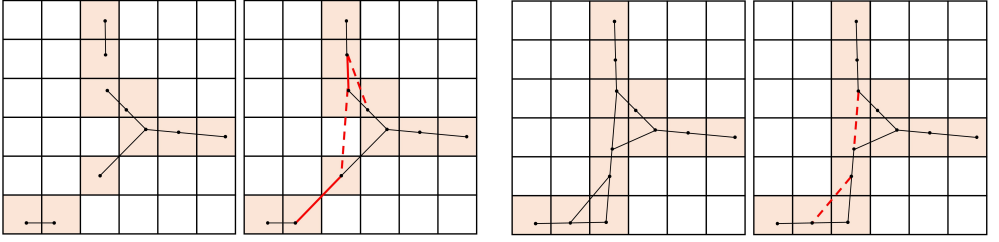


Figure 4: **Illustration of graph optimization strategies.** **Left:** connecting adjacent but unconnected endpoints. Red solid lines are links added while red dotted lines are links that should not be added. **Right:** removing triangle and quadrilateral. Red dotted lines are links removed.

\mathcal{L}_L , and \mathcal{L}_S for P , L , and S respectively as follows:

$$\begin{aligned}\mathcal{L}_{graph} &= \alpha \mathcal{L}_P + \beta \mathcal{L}_S + \gamma \mathcal{L}_L \\ \mathcal{L}_P &= - \sum_{i=1}^{N^2} [P_{gt}^i \log P_{pre}^i + (1 - P_{gt}^i) \log (1 - P_{pre}^i)] \\ \mathcal{L}_L &= - \frac{1}{\Omega_P} \sum_{\Omega_P} \sum_{j=1}^8 [L_{gt}^{ij} \log L_{pre}^{ij} + (1 - L_{gt}^{ij}) \log (1 - L_{pre}^{ij})] \\ \mathcal{L}_S &= - \frac{1}{\Omega_P} \sum_{\Omega_P} \sum_{j=1}^2 |S_{gt}^{ij} - S_{pre}^{ij}|\end{aligned}$$

where $\Omega_P = \{i | P_{gt}^i = 1, i \in [1, N^2]\}$, which is a set of patches that contain road, and j stands for the dimension of each patch-wise element. Patch-wise road probability loss and patch-wise link status loss use binary cross entropy (BCE) as the loss function while patch-wise road keypoint position loss uses mean absolute error (MAE). On account that we only consider patches that cover road and ignore other patches when it comes to locating road key-points and evaluating connection status, we define partial loss functions for L and S . α , β , and γ are the weights of the loss functions for three patch-wise elements.

3.4 Graph Optimization

As shown in Fig. 4, we develop several graph optimization strategies to refine the patch-wise link status to improve the performance of road network construction.

Connecting adjacent but unconnected endpoints. Due to the occlusion of buildings and trees, the connectivity status of adjacent patches not always appears realistic. Through extensive observation, we find out that adjacent but unconnected endpoints predicted by the framework are mostly connected actually.

Removing triangle and quadrilateral. Several intersections are so large that cover several patches, this will leads to unnecessary links between these patches and inaccurate road graphs. We remove the diagonal link if a triangular connection between adjacent patches appears. In addition, we remove the longest link if a quadrilateral connection between adjacent patches appears.

4 Experiments

4.1 Experimental Datasets

DeepGlobe Road Extraction Dataset [8] is collected from the DigitalGlobe platform and other datasets, and consists of high-resolution satellite images of size 1024×1024 pixels. It covers images captured over Thailand, Indonesia, and India, and the ground resolution is 0.5m/pixel. We utilize the original training set which consists of 6226 images and randomly split it into 3984, 997, and 1245 images for training, validation, and testing respectively.

Massachusetts Roads Dataset [14] is collected from publicly available imagery and metadata and consists of satellite images of size 1500×1500 pixels. It covers images captured over a wide variety of urban, suburban, and rural regions of the state of Massachusetts, and the ground resolution is 1m/pixel. We crop the 1171 original images into 3168 images of size 1024×1024 pixels and randomly split them into 2027, 507, and 634 images.

RoadTracer Dataset [2] is collected from Google and consists of high-resolution satellite images of size 4096×4096 pixels. It covers the urban core of forty cities across six countries, and the ground resolution is 0.6m/pixel. We crop and resize the training set which consists of 300 images into 1200 images of size 1024×1024 pixels and randomly split it into 960 and 240 images for training and validation. 15 8192 \times 8192 images are used for testing.

The split of the datasets aims for approximate distribution of 64%, 16%, and 20%.

4.2 Implementation Details

All the experiments are conducted with the PyTorch framework on NVIDIA 3090Ti GPUs. We use DeepGlobe Road Extraction Dataset [8], Massachusetts Roads Dataset [14], and RoadTracer Dataset [2] for experiments. The input satellite images are with a dimension of 1024×1024 and the patch size is set to 16×16 . During training, the batch size is fixed as 4 and the total epoch is set to 300. We use Adam optimizer with the learning rate originally set to $2e-4$ and reduced by 5 times once the loss keeps relatively higher for several epochs. The weights of loss functions α , β , and γ are set to 0.5, 1, and 1 respectively. To avoid overfitting, we utilize a variety of data augmentation ways, including horizontal flip, vertical flip, random rotation, and color jittering. We use the pixel-based F1 score [20] and APLS [20] metric to evaluate road graphs while using the IoU score to evaluate segmentation performance.

4.3 Comparison with Other Methods

For segmentation-based approaches, we compare the PaRK-Detect scheme with D-LinkNet [26] on DeepGlobe and Massachusetts Roads Datasets. The comparison results are shown in Tab. 1. Note that road graphs can be constructed through skeletonizing segmentation masks generated by segmentation-based approaches for evaluating APLS metric. Only focusing on pixel-wise semantic features, D-LinkNet acquires more road morphological information as well as contour features, nevertheless, this is redundant for constructing road graphs. On the contrary, our approach provides connectivity supervision while retaining the global receptive field and achieves outstanding performance with pixel-based F1 score and APLS metric. Besides, as shown in Tab. 3, our method outperforms D-LinkNet by 23.1% in terms of inference speed as segmentation-based approaches require skeletonization and vectorization in post-processing procedures.

Method	DeepGlobe			Massachusetts Roads		
	P-F1	APLS	IoU	P-F1	APLS	IoU
D-LinkNet	74.87	62.74	65.27	71.10	59.08	56.41
Ours	78.04	69.83	65.53	74.80	63.37	56.80

Table 1: Comparison with segmentation-based approach on DeepGlobe and Massachusetts Roads Dataset.

Method	P-F1	APLS
VecRoad	72.56	64.59
RoadTracer	55.81	45.09
Ours	74.50	63.10

Table 2: Comparison with graph-based approaches on RoadTracer Dataset.

Method	Type	8192×8192
VecRoad	Graph	1327.7
RoadTracer	Graph	343.4
D-LinkNet	Seg	10.1
Ours	PaRK-Detect	7.7

Table 3: Run-time in seconds of different approaches on one 8192×8192 test image.

For graph-based approaches, we compare our scheme with RoadTracer [14] and VecRoad [15] on RoadTracer Dataset. The comparison results are shown in Tab. 2. Our approach achieves competitive results compared with RoadTracer and VecRoad while the inference speed of the PaRK-detect scheme far exceeds that of these approaches.

4.4 Ablation Studies

Fig. 5 presents the results of constructed road graphs based on our proposed PaRK-Detect scheme. To verify the effectiveness of our approach and the framework architecture, we conduct ablations studies for multi-scale feature fusion and graph optimization strategies. As shown in Tab. 4, introducing JPU to perform multi-scale feature fusion improves the overall performance of our method. Graph optimization strategies improve the pixel-based F1 score and APLS of constructed road graphs by 1.7% and 9.0% respectively on the DeepGlobe dataset and by 0.6% and 10.8% respectively on the Massachusetts Roads dataset. Performing graph optimization will not influence the IoU metric of road segmentation.

Besides, in order to explore the impact of the multi-task paradigm for road extraction, we remove the decoder part of our framework and compare the road graph construction results with the original framework. As shown in Table 4, simply performing road graph construction without road segmentation diminishes pixel-based F1 score and APLS, which implies road segmentation can be served as additional supervision and provides road morphological information and contour features for achieving better alignment of roads.

Patch size is a crucial hyperparameter in our PaRK-Detect scheme and different patch sizes lead to different framework architectures and discrepant road alignment accuracy. We use 16×16 as the default setting and conduct extensive experiments to explore the impact of patch size. We observe that the appropriate patch size is determined by the width of the road in satellite imagery. If the road width is mostly concentrated between 12-24 pixels (e.g. DeepGlobe and Massachusetts Roads dataset), 16×16 is the suitable patch size. If the road width is mostly 24-48 pixels (e.g. RoadTracer dataset), we should choose 32×32. However, we utilize interpolation to resize the RoadTracer dataset and the patch size remains 16×16.

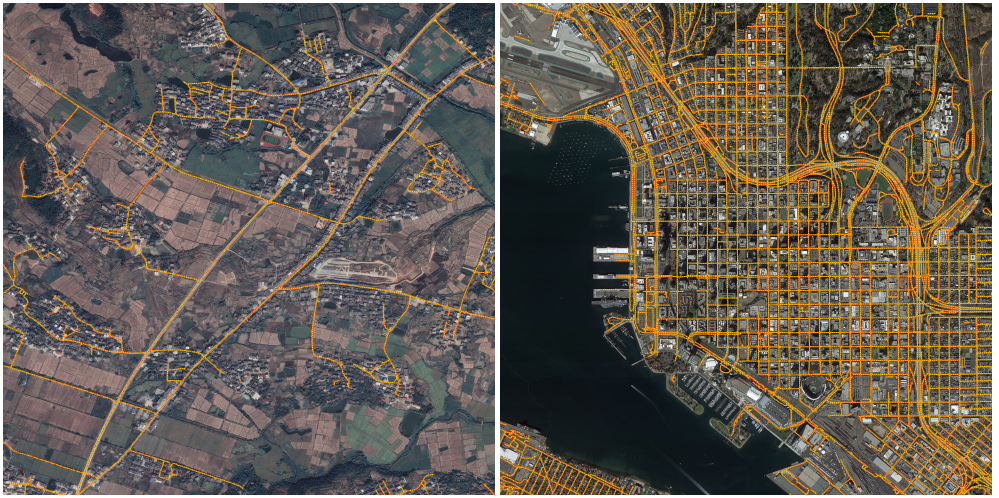


Figure 5: Results of constructed road graphs with PaRK-Detect scheme.

multi-scale feature fusion	graph optimization	segmentation	patch size	DeepGlobe			Massachusetts Roads		
				P-F1	APLS	IoU	P-F1	APLS	IoU
✓	✓	✓	16×16	77.44	69.07	65.24	74.36	62.80	56.43
✓	✓	✓	16×16	76.70	64.06	65.53	74.35	57.17	56.80
✓	✓	✓	16×16	76.97	69.64	-	72.98	58.07	-
✓	✓	✓	16×16	78.04	69.83	65.53	74.80	63.37	56.80
✓	✓	✓	8×8	66.48	58.12	65.49	65.52	49.68	56.72
✓	✓	✓	32×32	76.41	67.91	65.48	73.57	60.90	56.74

Table 4: Ablation studies of our proposed scheme. No mask will be predicted for evaluating IoU if we remove the road segmentation part of the proposed multi-task framework.

5 Conclusion

In this paper, we propose a patch-wise road keypoints detection scheme that combines road extraction with keypoint detection to construct road graphs in a single pass. We also propose an efficient multi-task framework that performs road segmentation and road graph construction simultaneously. We conduct extensive experiments over widely-used datasets and achieve competitive or better road extraction results with superior inference speed against existing works. Through ablation experiments, we verify the effectiveness of our approach and demonstrate the impact of the multi-task paradigm and patch size. We believe our work presents an important step forward and we hope this paper can inspire future works on road extraction from satellite imagery.

6 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62076093 and the High-Performance Computing Platform of BUPT.

References

- [1] Gaetan Bahl, Mehdi Bahri, and Florent Lafarge. Road extraction from overhead images with graph neural networks. *arXiv preprint arXiv:2112.05215*, 2021.
- [2] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David DeWitt. Roadtracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4720–4728, 2018.
- [3] Alexander Buslaev, Selim Seferbekov, Vladimir Iglovikov, and Alexey Shvets. Fully convolutional network for automatic road extraction from satellite imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 207–210, 2018.
- [4] Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xiangyu Zhang, Xinyu Zhou, Erjin Zhou, and Jian Sun. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, pages 455–472. Springer, 2020.
- [5] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.
- [6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Re-thinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [7] Guangliang Cheng, Ying Wang, Shibiao Xu, Hongzhen Wang, Shiming Xiang, and Chunhong Pan. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 55(6):3322–3337, 2017.
- [8] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Songtao He, Favyen Bastani, Satvat Jagwani, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Mohamed M Elshrif, Samuel Madden, and Mohammad Amin Sadeghi. Sat2graph: Road graph extraction through graph-tensor encoding. In *European Conference on Computer Vision*, pages 51–67. Springer, 2020.

- [12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [13] Gellért Mátyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE international conference on computer vision*, pages 3438–3446, 2017.
- [14] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.
- [15] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [17] Husen Shi and Zengfu Wang. Improved stacked hourglass network with offset learning for robust facial landmark detection. In *2019 9th International Conference on Information Science and Technology (ICIST)*, pages 58–64. IEEE, 2019.
- [18] Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. Vecroad: Point-based iterative graph exploration for road graphs extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8910–8918, 2020.
- [19] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.
- [20] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.
- [21] Yao Wei, Kai Zhang, and Shunping Ji. Simultaneous road surface and centerline extraction from large-scale remote sensing images using cnn-based segmentation and tracing. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8919–8931, 2020.
- [22] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yizhou Yu. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. *arXiv preprint arXiv:1903.11816*, 2019.
- [23] Xiaofei Yang, Xutao Li, Yunming Ye, Raymond YK Lau, Xiaofeng Zhang, and Xiaohui Huang. Road detection and centerline extraction via deep recurrent convolutional neural network u-net. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9): 7209–7220, 2019.
- [24] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

- [25] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [26] Lichen Zhou, Chuang Zhang, and Ming Wu. D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 182–186, 2018.