FRENet: Faster Road Extraction of High-Resolution Remote Sensing Images via Real-Time Semantic Segmentation

Hucheng Liu†, Wanfeng Zheng†, Zhanbei Cui, Chuang Zhang, and Ming Wu*
School of Artificial Intelligence
Beijing University of Posts and Telecommunications
{xiaocheng2333, zhengwanfeng, czb1997, zhangchuang, wuming}@bupt.edu.cn

***Abstract*—In recent years, approaches based on semantic segmentation have been proposed for road extraction of high-resolution remote sensing images. However, the computational complexity of these methods is high, which leads to low inference speed. To address this issue, we attempt to apply real-time semantic segmentation methods to road extraction. Nevertheless, due to the excessive pursuit of inference speed, their lightweight models do not have sufficient capacity to extract more comprehensive road information and the simple decoder structure has disrupted road connectivity. This paper proposes a novel network named FRENet for faster road extraction of high-resolution remote sensing images. More specifically, we propose a spatial transform fusion module (STFM) that enables the network to capture more spatial context information with fewer parameters. We also redesign a new decoder to solve the problem of road interruptions with acceptable extra computational cost. Experiment results on two public datasets demonstrate that FRENet achieves comparable performance to the state-of-the-art semantic segmentation methods while maintaining much lower computational complexity.**

I. Introduction

Road extraction is a challenging computer vision recognition task in the field of remote sensing which has been a hot research topic in the past decade. In recent years, definitions of the road extraction task are mainly divided into two types: one is graph-based methods [1], [2], [3] that iteratively predict road graphs, which only contain the centerline of each road; the other is semantic segmentation-based methods [4], [5] that predict road networks which describe whether each pixel belongs to the road. The width and outline of each road can be displayed in the segmentation result, and the road graph, which is the vectorized representation of road maps, can be obtained by post-processing the segmentation result.

In some practical application scenarios, the width and outline of the road need to be known, such as city planning, car navigations, and road damage assessment. Therefore, it is necessary to use segmentation-based methods to generate pixel-level labeling of roads. Each pixel in the remote sensing image needs to be classified as a road or background in the road segmentation task. For more accurate segmentation results, existing methods constantly design larger networks [6], [4] or use more complex modules [7], but high computational complexity results in insufficient inference speed. The Facebook team developed an app called MAP WITH AI [8] to fill more than 300,000 miles of missing roads in Thailand by using D-LinkNet [5] which is the champion of DeepGlobe Road Extraction Challenge [9]. Though parameters and MACs of D-LinkNet are much smaller compared with several classical semantic segmentation networks, it still took the Facebook team a year and a half to complete this work. When extracting roads from a large number of remote sensing images, the processing capacity of servers is limited by the inference speed of road segmentation. Although segmentation-based methods can generate not only road centerline maps but also contour maps, their computational complexity is unacceptable in practical applications. Therefore we apply real-time segmentation approaches for road segmentation to improve the inference speed of the segmentation-based method.

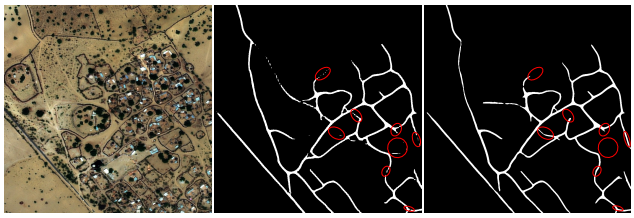

Fig. 1: The column on the left is the remote sensing image from DeepGlobe, and the other columns are inference results. The middle column is predicted by DDRNet-23-Slim with $8\times$ bilinear interpolation upsampling. The right column is predicted by our proposed FRENet. Some of the roads predicted by DDRNet-23-Slim are disrupted into dashed lines, while our results are more continuous and smooth.

Recently, some competitive real-time methods aiming at semantic segmentation of road scenes[10], [11] are proposed. Some of these methods develop complex lightweight encoders trained from scratch, one of which, DDRNet [12] hits a new peak in terms of real-time performance. However, the output resolution of DDRNet is one eighth of the input resolution. Although we can make upsampling on the segmentation results by nearest neighbor interpolation or bilinear interpolation, roads in the output will be disrupted into dashed lines, which has been shown in Fig. 1. These disruptions can have a destructive effect on downstream tasks like road centerline extraction. Thus it is necessary to eliminate this problem.

† These authors contributed equally to this work.
* Ming Wu is the corresponding author.

To prevent road disruption, the most challenging difficulty is to make improvements to the network architecture with minimal extra computation cost. Because of the limitation of increment on parameters and MACs, we have adopted DDRNet-23-Slim as our backbone like DDRNet and redesigned the decoder. We named our new network structure FRENet. To balance the inference speed and segmentation performance, the new decoder only has three convolutional layers. We propose a spatial transform fusion module (STFM) and add it to the new decoder structure. This module can transform the input features into different spatial conditions. Then these transformed features are sent into a shared convolutional layer. The corresponding inverse transform is performed on each output of the shared convolutional layer for further fusion. The network can capture more spatial context information through performing convolution on feature maps from different views. And the fusion after inverse transformations gathers these context information into the output features. Our proposed spatial transform fusion module has the same parameters as one convolutional layer. Its computation cost is acceptable because it receives features with lower resolution and a channel reduction layer is performed before it. After modification, our proposed FRENet has comparable inference speed to DDRNet while maintaining similar performance with larger segmentation networks.

Evaluation of our approach has been made on two public datasets. The two public datasets are DeepGlobe [9] and Massachusetts Roads Dataset [13]. Both of them consist of high-resolution remote sensing images and high-quality pixel-wise manually annotated labels. In experiments, we have evaluated the road segmentation performance and inference speed of our proposed network, while some contrast experiments made on other lightweight networks and real-time segmentation networks have proved the effectiveness of our proposed network. Evaluation metrics including mean Intersection over Union (mIoU) and inference iteration per second (FPS) have shown our approach achieves comparable road segmentation performance to the state-of-the-art while maintaining much better inference speed. And our proposed new metric named Average Path Length (APL) has shown our approach solves the problem of road interruption.

In summary, our contributions are:

- We have adopted a real-time semantic segmentation approach for road segmentation and found the road disruption problem, which can be solved by modifying the decoder with a bit of extra computation cost.
- Spatial transform fusion module (STFM) has been designed to capture contextual information from different perspectives. Fusion after inverse transforma on the output of a shared convolutional layer can reassemble these contextual information.
- We propose a new metric named Average Path Length (APL) to evaluate the road connectivity in the segmentation result. Evaluation metrics including mean Intersection-over-Union (mIoU) and inference iteration per second (FPS) have shown our approach achieves comparable road segmentation performance to the state-of-the-art while maintaining much better inference speed. The comparison results of APL have shown our approach solves the problem of road interruption.

II. Related work

A. Semantic Segmentation

Semantic segmentation is a fundamental task in which each pixel of the input image should be assigned to the corresponding label. With the rise of deep learning technologies, convolutional neural networks are applied to image segmentation and greatly outperform traditional methods based on handcrafted features. Fully Convolutional Network (FCN) [14], an end-to-end network that is almost composed of convolutional layers, is the first effective deep-learning method for the task. U-Net [15] fuses features from encoders and decoders which have the same shape to combine context information and location information. It is mainly used for biomedical image segmentation. Deeplabv2 [16] proposed an Atrous Spatial Pyramid Pooling module (ASPP) using different dilation rates to get multi-scale spatial information. Deeplabv3 [17] and Deeplabv3+ [18] improved the ASPP module and started to use the larger network as the backbone such as Xception. HRNet [19] has many high-to-low resolution subnetworks and could get rich and high-resolution representations by connecting them.

B. Road Extraction

Extracting roads from remote sensing images into binary pixels is a well-studied task. Traditional methods construct road maps by various techniques such as utilizing nearby buildings and vehicles [20], shape factors [21], simulated annealing technology [22], and distinct spectral contrast and locally linear trajectory [23]. Minimum spanning tree [24], higher-order conditional random field [25], [26], and junction process [27] are also performed to construct road graphs.

Recent works apply deep learning to generate road maps with higher performance. Zhang *et al.* [4] apply residual connections [28] to the U-Net [15] to learn more delicate features for road segmentation. In the DeepGlobe Road Extraction Challenge, D-LinkNet got IoU score of 0.6453 on the validation set. The D-LinkNet combines dilation convolutions [29] and LinkNet [30] to enlarge the receptive field for road extraction from high-resolution satellite imagery. Compared with several classical semantic segmentation networks [31], [18], parameters and MACs of D-LinkNet are much smaller. Moreover, its performance on road segmentation is much better than small networks like U-Net [15].

C. Real-time Semantic Segmentation

Real-time Semantic Segmentation is a task demanding for a short time in the inference stage. Usually, these networks designed for real-time semantic segmentation need a trade-off between accuracy and speed. BiSeNetV1 [32] has two branches (Spatial Path and Context Path) whose features are merged at the end of the network. BiSeNetV2 [33] improves it by using global average pooling for context embedding and proposes attention-based feature fusion. The two pathways in BiSeNetV1&V2 are initially separate while the two branches in Fast-SCNN [34] share the learning to downsample module. CABiNet [35] adopts the overall architecture of Fast-SCNN but uses the MobileNetV3 [36] as the context branch.

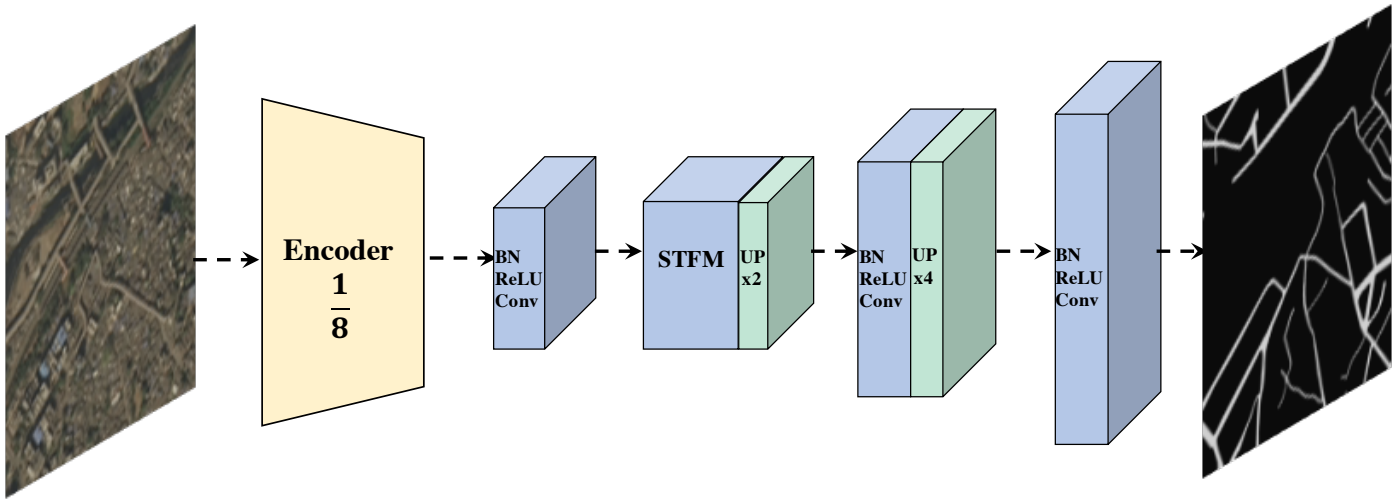

Fig. 2: The overview of FRENet on road extraction. We adopt DDRNet-23-Slim as our encoder whose output resolution is one eighth of the input resolution. The blue block in the right consists of the batch normalization layer, ReLU layer and convolutional layer. STFM is our proposed spatial transform fusion module. "UP" denote upsampling layer with bilinear interpolation and the number is the ratio of upsampling.

SFNet [37] proposed a Flow Alignment Module (FAM) to get better upsampled features than bilinear interpolation. DDRNet [12] now is the state-of-art method for real-time semantic segmentation of Cityscapes. It has two branches, one of which aims to generate high-resolution features. At the same time, the other captures semantic information by downsampling and constructs a Deep Aggregation Pyramid Pooling Module (DAPPM) to adapt to various receptive fields.

III. Method

In this section, our description of our approach consists of three parts: 1) overall architecture of our proposed network; 2) spatial transform fusion module; 3) objective function of our method.

A. Overview of Network Architecture

This proposed FRENet is designed for faster road segmentation. Modifications have been made to improve the inference speed and segmentation performance.

To improve the inference speed, we adopt DDRNet-23-Slim as our backbone, which evolved from HRNet [19] but with only two parallel branches. After comparing DDRNet-23-Slim with other real-time segmentation networks, we have found that DDRNet-23-Slim is the most suitable backbone for our task. But its disadvantage is also destructive: disrupting roads into small broken dashed lines and making the edges of the road unsmooth because its output resolution is one-eighth of the input resolution. To address this issue, we have removed the original segmentation head of DDRNet and redesigned a new decoder. The whole network architecture has been shown in Fig. 2.

Our redesigned decoder consists of three convolutional layers with batch-normalization layers and activation layers between them and a spatial transform fusion module. We have not employed more convolutional layers because too many layers will dramatically decrease the inference speed on GPU and these extra MACs can also decrease the inference speed on CPU. Using features extracted by the encoder as its input, the decoder first uses a 3×3 convolutional layer to make channel dimension reduction. It is more popular to use point-wise convolution with kernel size 1 to reduce the number of channels. But we have found the inference speed has not been influenced by whether we use 3×3 or 1×1 convolutional layer thanks to GPU acceleration. Rather than 1×1 convolutional layer, 3×3 convolutional layer can make better performance.

Following the channel reduction layer, we use spatial transform fusion module (STFM) to make inferences from different spatial views. Features fused by STFM can capture more spatial context information, which can be utilized in subsequent layers. After STFM, the resolution of the feature maps is upsampled by a factor of two by bilinear interpolation. Then a 3×3 convolutional layer is adopted in series, followed by a bilinear interpolation upsampling layer with a ratio of 4. Finally, the number of channels is reduced to 1 by the last 3×3 convolutional layer, followed by a sigmoid activation layer.

The redesigned decoder can dramatically eliminate road disruption with acceptable extra computational cost. Some comparisons of results between DDRNet-23-Slim and FRENet has been shown in Fig. 1.

B. Spatial Transform Fusion Module

Because of the limitation of extra parameters, we have designed the spatial transform fusion module (SFTM), which has shown its ability to capture multi-view contextual information.

The STFM receives features produced by the previous convolutional layer with a reduction in the number of channels. It transforms the input to three other different conditions by clockwise rotation, vertical flip and horizontal flip. Therefore the transformed feature is four times the input. Different from those methods which expand the size of feature maps on the "channel" dimension, we expand the size of feature maps on the "batch" dimension.

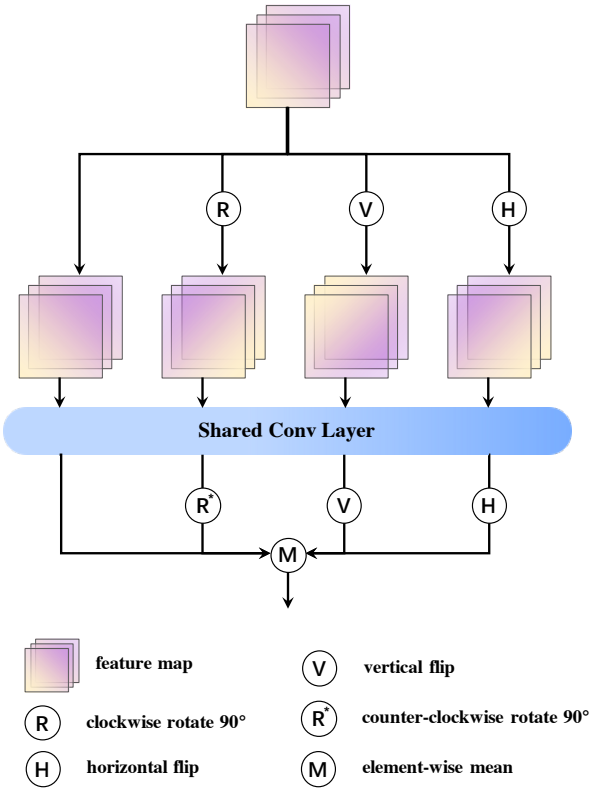

Fig. 3: Structure of our proposed spatial transform fusion module. The input feature maps are transformed to three other conditions by rotation and flip. The transformed feature maps and the original feature maps are fed into a shared convolutional layer. After the convolution operation and inverse transformation, the element-wise mean operation produces the output feature maps.

This operation setup permits the convolutional layer to be shared in forward propagation without extra convolution kernels. During training, the shared convolutional layer is forced to capture contextual information from four different views in total. Its convolution kernels are trained to extract spatial information from rotated and flipped features, which means the kernels have obtained rotation invariance and flip invariance. Besides, these transformations can also expand the reception field for the convolutional kernel and help them receive point-wise features from various locations in the whole feature map.

After feature extraction by the shared convolutional layer, features are inverse transformed to the original condition. Then features are fused by element-wise mean. The overall process is similar to model averaging, but we have not employed other network architectures but fused features extracted from different views. Although the expansion on the "batch" dimension can cause four times the computational cost of the shared convolutional layer, the inference speed is not decreased much thanks to the GPU acceleration. The architecture of STFM is shown in Fig. 3.

C. Objective Function

Our overall objective function consists of three parts. Here, Y denotes the target mask with pixel values in $\{0, 1\}$. X denotes the output of our network. y and x represent the pixel values in Y and X.

$$\mathcal{L}_{BCE} = \mathbb{E}_{x,y}[-y \cdot logx - (1-y) \cdot log(1-x)] \quad (1)$$

BCE (binary cross entropy) loss is widely used in binary segmentation. The output of our network has only one channel. After being activated by the sigmoid layer, values of each pixel are mapped between 0 and 1 indicating the probability of the pixel belonging to the road category. The BCE loss is able to tell the model each pixel should be classified into road category or background.

$$\mathcal{L}_{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \quad (2)$$

We also use dice loss to improve the segmentation performance. Dice loss can calculate the similarity between the segmentation result and ground truth and give the network feedback to optimize the network. In Equation 2, $|X|$ is the reduction result of summing X.

$$\mathcal{L}_{Dis} = \mathbb{E}_{x,x_t}[(x - x_t)^2] \quad (3)$$

In addition to the objective function between segmentation results and ground truth, there is also a distillation loss between segmentation results produced by our network and a pretrained teacher network. In Equation 3, x_t denotes pixel value in the soft label produced by the pretrained teacher network. By employing the pretrained teacher network, the student network is forced to produce similar segmentation results to the teacher network. Though information from ground truth is more accurate, soft labels from the teacher network can guide the student network to converge towards better performance.

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \mathcal{L}_{Dice} + \lambda_{dis}\mathcal{L}_{Dis} \quad (4)$$

Our overall objective can be described in Equation 4. Here, λ_{dis} is a hyper-parameter, which has been set to 20 during training.

IV. Experiment

A. Experiment Setup

Datasets. The DeepGlobe [9] dataset consists of 6226 densely annotated images and the resolution of each image is 1024×1024. We have adopted its validation set as our test set because the ground truth of its validation set and test set are not released. We upload the inference of our network to the online test inference for evaluation. The Massachusetts Roads Dataset [13] contains 1108 finely annotated images for training, 14 images for validation, and 49 images for testing. The resolution of images is 1500×1500. We fill the image to 1504×1504 for training and testing.

Implementation Details. To make fair comparisons, experiments using different networks have the same training strategy. We have used Adam optimizer and the objective function described in III-C. When λ_{Dis} in our objective is not zero,

Method	DeepGlobe				Massachusetts				Params(M)
	mIoU↑	APL↑	FPS↑	MACs(G)↓	mIoU↑	APL↑	FPS↑	MACs(G)↓	
D-LinkNet34 [5]	**64.58**	588.59	37.31	134.23	**65.88**	911.16	16.52	289.56	31.10
D-LinkNet18* [5]	64.07	539.88	48.11	95.52	65.21	959.17	21.14	206.0	20.99
U-Net* [15]	63.09	522.76	86.13	35.03	63.98	866.66	39.71	78.82	39.50
BiSeNetV2* [33]	62.36	229.21	85.46	49.12	64.11	308.98	38.79	105.96	3.34
CABiNet* [35]	60.32	202.83	99.93	**6.78**	60.02	223.08	50.12	**14.62**	**2.58**
DDRNet-23-Slim* [12]	64.22	281.18	**127.24**	18.23	64.30	445.25	**78.16**	39.37	5.69
Ours*	64.40	**645.10**	111.02	21.68	65.84	**1273.86**	52.02	46.81	5.74

TABLE I: Quantitative results on two datasets. The first group is the state-of-the-art method for road extraction and the second group is several lightweight networks and real-time semantic segmentation networks. ↑ indicates the larger the better for this column, and ↓ indicates the smaller the better. The method is trained with a distillation loss if the method is marked with *. The pretrained teacher network is D-LinkNet34. Due to the different resolutions of two datasets, the results of FPS and MACs are different.

a pretrained D-LinkNet34 is adopted as the teacher network. We have done experiments with and without the distillation objective function. The training lasts for 200 epochs in total. The learning rate is set to 2e-4 in the first 100 epochs and decreases linearly in the second 100 epochs. Data augmentations including rotating, flipping, scaling and color jittering scaling and color jittering are utilized. Test time augmentation (TTA) is used during the testing phase.

Metrics. Mean Intersection-over-Union (mIoU) is used to evaluate the segmentation performance. For a binary semantic segmentation task, it can be described in Equation 5. Inference iteration per second (FPS) is used to evaluate the inference speed.

$$mIoU = \frac{TP}{TP + FP + FN} \tag{5}$$

We propose a new metric named Average Path Length (APL) to evaluate the road connectivity in segmentation results. We first use a skeleton algorithm to generate masks containing the centerlines of roads. Then we use Breadth First Search (BFS) to calculate the number of centerlines in the mask, S_l. The sum of the number of pixels of all centerlines is S_p. If the number of images in the test set is **n**, the APL value is described in Equation 6. This APL metric is a variant of average path length similarity [38].

$$APL = \frac{\sum_{i=1}^{n} S_{pi}}{\sum_{i=1}^{n} S_{li}} \tag{6}$$

B. Comparision with State-of-the-art Methods

As can be observed from Table I, our method achieves a new state-of-the-art trade-off between segmentation performance and inference speed on two public datasets. FRENet achieves 64.40% mIoU on the test set of the DeepGlobe dataset at 111 FPS and 65.84% mIoU on the test set of the Massachusetts Roads Dataset at 52 FPS.

Compared with state-of-the-art methods for road extraction, FRENet reasons approximately three times as fast as D-LinkNet34 at the cost of a little mIoU loss. FRENet applies the real-time semantic segmentation method to road extraction and uses the real-time semantic segmentation method DDRNet-23-Slim which has the fastest inference speed as the backbone. Compared with several common real-time semantic segmentation methods, FRENet has the most outstanding performance on mIoU and reasons only slightly slower than DDRNet-23-Slim. With the help of our proposed spatial transform fusion module, FRENet is trained to be able to capture contextual information from different views so that it can have sufficient ability to extract road information and achieves better segmentation performance.

According to the results of APL in Table I and Fig. 4, FRENet has the best APL scores on both DeepGlobe dataset and Massachusetts Roads Dataset. The Average Path Length of FRENet is more than two times as long as DDRNet-23-Slim, which demonstrates that our method dramatically eliminates road disruption and gets more accurate road edges. Our redesigned decoder only consists of three convolutional layers and a spatial transform fusion module. The extra computational cost of the new decoder is acceptable, which can be seen from the comparison of inference speed. Compared with state-of-the-art methods for road extraction, FRENet has a slight mIoU loss which means it could not detect roads in some areas, but it has the best road connectivity performance on the roads it could detect.

C. Ablation

In this section, we conduct experiments on several variants of our approach to evaluate each component. As illustrated in Table II, the basic DDRNet-23-Slim without distillation can achieve 63.81% mIOU and 290.72 APL on the test set of DeepGlobe dataset at 127 FPS and 63.54% mIOU and 367.33 APL on the test set of Massachusetts Roads Dataset 78 FPS. With the distillation loss, DDRNet-23-Slim can achieve a gain of 0.41% and 0.76% mIoU over the baseline. This demonstrates soft labels from the teacher network can guide the student network to converge towards better performance.

By redesigning the decoder with three convolutional layers, FRENet without STFM can achieve more than two times APL compared with DDRNet-23-Slim with the distillation loss.

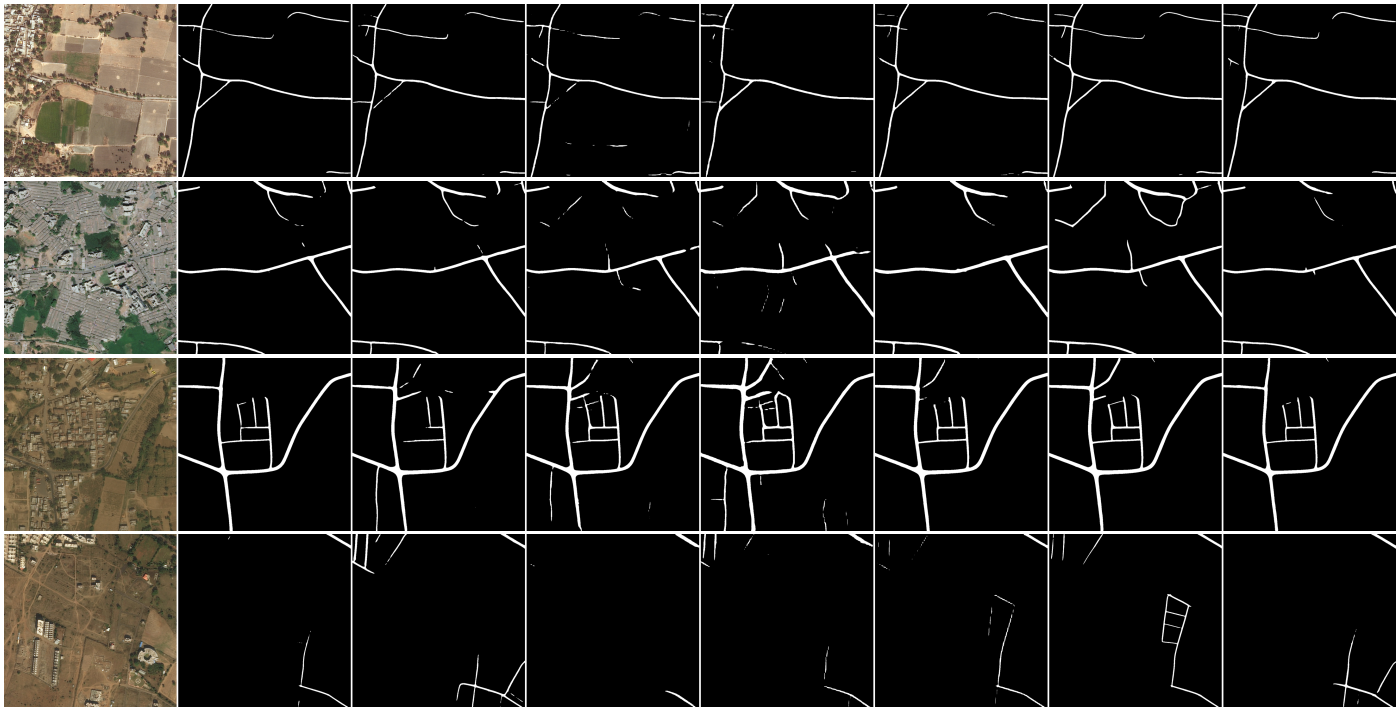

Fig. 4: Qualitative results on two datasets. Method names are annotated at the bottom. The first column on the left is the input remote sensing image. The first column on the right is the result of the teacher.

	Configuration	DeepGlobe			Massachusetts		
		mIoU	**APL**	**FPS**	**mIoU**	**APL**	**FPS**
A	Baseline ($\lambda_{Dis} = 0$)	63.81	290.72	**127.24**	63.54	367.33	**78.16**
B	(A) + $\lambda_{Dis} = 20$	64.22	281.18	**127.24**	64.30	445.25	**78.16**
C	(B) + new decoder	64.15	610.93	122.82	65.67	1152.73	69.47
D	(C) + Conv	64.23	624.47	120.13	65.73	1241.07	63.28
E	(C) + STFM (Ours)	**64.40**	**645.10**	111.02	**65.84**	**1273.86**	52.02

TABLE II: Ablation Study of FRENet on two datasets. The baseline is DDRNet without distillation. "B" represents training DDRNet-23-Slim with the distillation loss. "C" represents redesigning the decoder with three convolutional layers without STFM. "D" represents replacing STFM with a single convolutional layer.. "E" represents our model, FRENet.

Although there is a small drop in inference speed, the results demonstrate that the redesigned decoder structure can better decode the road information extracted by the encoder and eliminates road disruption.

To evaluate the effectiveness of our proposed spatial transform fusion module (STFM), we replaced STFM with a single convolutional layer. The results show that FRENet with an additional convolutional layer improves a bit on both mIoU and APL. Compared to a single convolutional layer, the spatial transform fusion module transforms the input feature maps to different spatial conditions and the fusion after inverse transformations gathers these context information into the output features. Results demonstrate the spatial transform fusion module brings an gratifying improvement on both mIoU and APL. The comparison results with baseline in inference speed show that the computation cost is acceptable.

V. Conclusion

In this paper, we are devoted to accelerate the inference speed of road extraction, while maintaining comparable segmentation performance. We have applied real-time segmentation methods to road extraction and designed a FRENet with a spatial transform fusion module for faster road extraction of high-resolution remote sensing images. Experiments on two public datasets have proved the effectiveness of our method, which achieves comparable performance to the state-of-the-art with much faster inference speed.

REFERENCES

[1] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "Roadtracer: Automatic extraction of road networks from aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4720–4728.

[2] S. He, F. Bastani, S. Jagwani, M. Alizadeh, H. Balakrishnan, S. Chawla, M. M. Elshrif, S. Madden, and M. A. Sadeghi, "Sat2graph: road graph extraction through graph-tensor encoding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 51–67.

[3] Y.-Q. Tan, S.-H. Gao, X.-Y. Li, M.-M. Cheng, and B. Ren, "Vecroad: Point-based iterative graph exploration for road graphs extraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8910–8918.

[4] 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.

[5] L. Zhou, C. Zhang, and M. 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*, 2018, pp. 182–186.

[6] M. Cheng, K. Zhao, X. Guo, Y. Xu, and J. Guo, "Joint topology-preserving and feature-refinement network for curvilinear structure segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7147–7156.

[7] Y. Xu, H. Chen, C. Du, and J. Li, "Msacon: Mining spatial attention-based contextual information for road extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[8] Facebook, "Map with ai," https://mapwith.ai/#13/24.02269/-104.68274/0/55.

[9] 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," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 172–181.

[10] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 607–12 616.

[11] P. Hu, F. Perazzi, F. C. Heilbron, O. Wang, Z. Lin, K. Saenko, and S. Sclaroff, "Real-time semantic segmentation with fast attention," *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 263–270, 2020.

[12] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *arXiv preprint arXiv:2101.06085*, 2021.

[13] V. Mnih, *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013.

[14] 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.

[15] 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.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[19] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[20] 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.

[21] M. Song and D. Civco, "Road extraction using svm and image segmentation," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 12, pp. 1365–1371, 2004.

[22] R. Stoica, X. Descombes, and J. Zerubia, "A gibbs point process for road extraction from remotely sensed images," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 121–136, 2004.

[23] S. Das, 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.

[24] E. Türetken, F. Benmansour, and P. Fua, "Automated reconstruction of tree structures using path classifiers and mixed integer programming," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 566–573.

[25] J. D. Wegner, J. A. Montoya-Zegarra, and K. Schindler, "A higher-order crf model for road network extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1698–1705.

[26] ——, "Road networks as collections of minimum cost paths," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 108, pp. 128–137, 2015.

[27] D. Chai, W. Forstner, and F. Lafarge, "Recovering line-networks in images by junction-point processes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1894–1901.

[28] 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.

[29] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[30] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[32] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[33] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3051–3068, 2021.

[34] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[35] S. Saksena, "Cabinet: Efficient context aggregation network for low-latency semantic segmentation," Master's thesis, University of Twente, 2020.

[36] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[37] J. Lee, D. Kim, J. Ponce, and B. Ham, "Sfnet: Learning object-aware semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2278–2287.

[38] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv preprint arXiv:1807.01232*, 2018.