Supplemental Materials for Traffic Scene Parsing through the TSP6K dataset

Peng-Tao Jiang[1,2*] Yuqi Yang[1*] Yang Cao[3] Qibin Hou[1,4†] Ming-Ming Cheng[1,4] Chunhua Shen[2]

[1]VCIP, CS, Nankai University [2]Zhejiang University [3]HKUST [4]NKIARI, Shenzhen Futian

pt.jiang@mail.nankai.edu.cn, yangyq2000@mail.nankai.edu.cn, andrewhoux@gmail.com

1. More Experiments Results

1.1. Ablation Study

The number of region tokens and heads. First, we study the impact of the number of tokens and heads on the performance. As shown in Tab. 1, using 5 region tokens instead of 1 region token brings 0.6% mIoU scores and 1.2% iIoU scores improvement. This fact demonstrates that the number of region tokens largely affects the parsing of traffic participants, especially for small objects. When further improving the number of region tokens, we observe nearly no performance gain, which indicates 5 region tokens are enough for semantic region refining. Besides, we also attempt to increase the number of attention heads. It can be seen that adding more heads brings no performance gain. For readers to better understand the region tokens, we have visualized the attention maps of different tokens, as shown in Fig. 2. It can be seen that different tokens are responsible for different semantic regions.

Table 1. Ablation on the number of tokens and attention heads.

Settings	#Tokens	Attention Heads	mIoU$_{val}$	iIoU$_{val}$
1	1	12	75.2	57.2
2	5	12	**75.8**	58.4$_{(+1.2)}$
3	20	12	75.7	58.3$_{(+1.1)}$
4	20	24	75.5	**58.6**$_{(+1.4)}$

Region tokens *vs.* Class Tokens. In the design of the detail refining decoder, we utilize the region tokens to refine a specific semantic region. Here, one may raise a question: "How would the performance go when we utilize class tokens as done in the original Transformers instead of the region tokens"? We perform an experiment that learns 21 class tokens, each of which corresponds to a class. The final concatenated features are sent to a depth-wise convolutional layer with 21 groups. When using the class tokens, we can obtain 75.2% mIoU scores and 57.1% iIoU scores on the validation set. Compared with class tokens, the decoder with 5 region tokens can obtain 75.8% mIoU and 58.4% iIoU scores, which works better than using class tokens. Moreover, when the number of classes in the dataset is large, the class tokens will consume high computational costs. In contrast, using the region tokens is more flexible in that there is no need to adjust the number of region tokens when the number of classes rises.

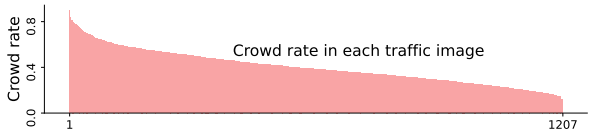

Figure 1. Crowd rate analysis of TSP6K validation set.

The importance of the encoder-decoder structure. In Sec. 4.2 of the main paper, we have analyzed that the encoder-decoder structure is vital for small object parsing. Thus, we apply the encoder-decoder structure to our segmentation network for utilizing high-resolution low-level features. Without the encoder-decoder structure, i.e., we directly connect the region refining module to the encoder, the mIoU and iIoU scores decrease by 0.7% and 1.3%, respectively. This experiment indicates that the high-resolution low-level features can benefit the parsing of the traffic participants. Thus, the encoder-decoder structure is vital for scene parsing.

1.2. Traffic Flow Analysis

One underlying application scenario of the monitoring scene parsing models is analyzing the traffic flow. Once we obtain the scene parsing results from well-trained models on the TSP6K dataset, we attempt to utilize these results to compute the traffic flow. Our solution is very simple. We first compute the area of the traffic participants (humans and vehicles) S_t and the area of the road S_r. Then, the crowd rate can be approximately calculated by S_t / (S_t + S_r). Fig. 1 shows the crowd rate of different traffic images on the TSP6K validation set. The higher the crowd rate, the more significant the traffic flow. The forthcoming traffic participants can arrange their travel plans based on the current crowd rate.

*The first two authors contributed equally to this work. Part of this work was done when P.-T. Jiang was a postdoc researcher at Zhejiang University.

†Q. Hou is the corresponding author.

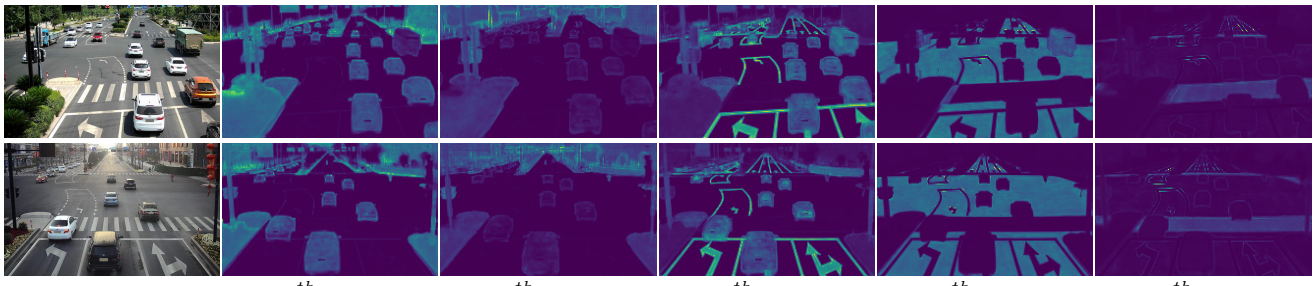

Figure 2. Visualizations of the attention map corresponding to each token. We randomly select several tokens for visualization. One can see that the visualizations associated with different region tokens focus on different semantic regions. These region tokens can help our method better process the region details.

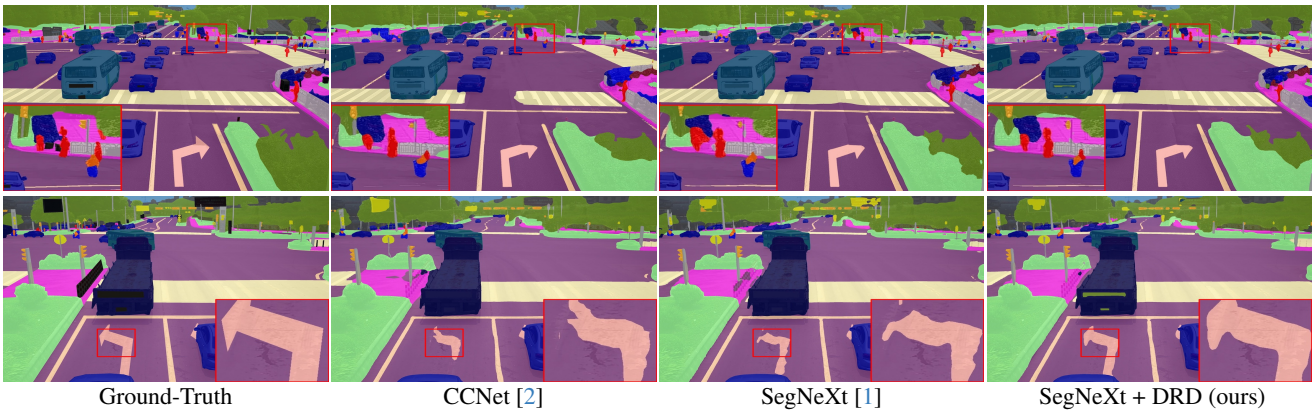

Figure 3. Visualization of the scene parsing results from different methods. One can see that our method can well process the region details. When taking the bottom scene as an example, our method can generate a more accurate mask for the arrow while other methods fail. Zoom in for the best view.

1.3. Visualization Comparison

We provide some qualitative results in Fig. 3 for visual comparison. We can see that our method achieves sharper results than CCNet and SegNeXt.

2. More Examples in TSP6K

We provide more examples picked from the TSP6K dataset, which is shown in Fig. 4. The picked examples are from different weather conditions and times in a day.

References

[1] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[2] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 603–612, 2019. 2

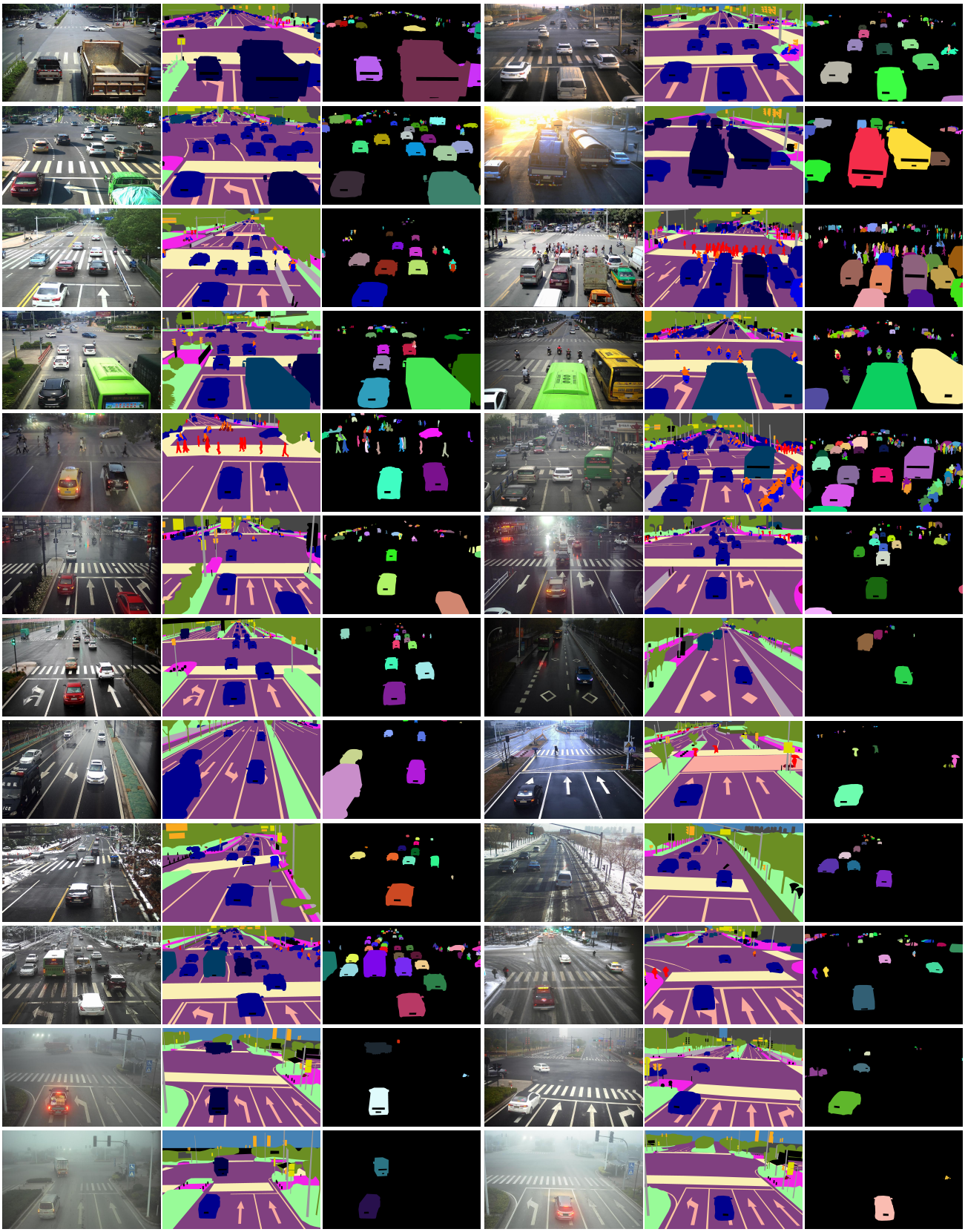

Figure 4. Examples are randomly picked from the TSP6K dataset. Each image is associated with its corresponding semantic label and instance label. We have masked the vehicle plates for privacy protection.