AttaNet: Attention-Augmented Network for Fast and Accurate Scene Parsing

Song Qi[†] Kangfu Mei[†‡] Rui Huang[†‡]

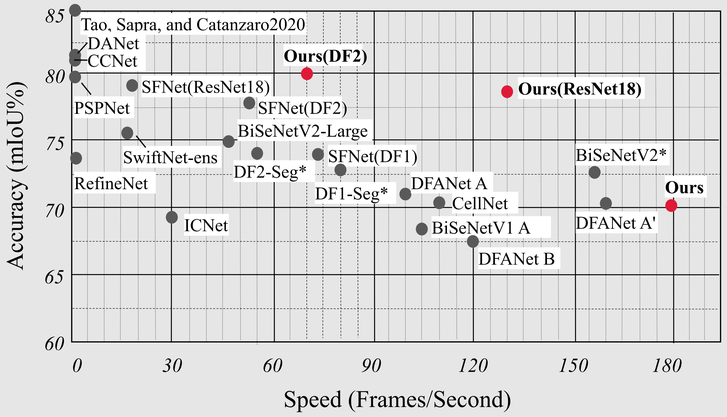
[†]Shenzhen Institute of Artificial Intelligence and Robotics for Society

[‡]The Chinese University of Hong Kong, Shenzhen

https://github.com/songqi-github/AttaNet

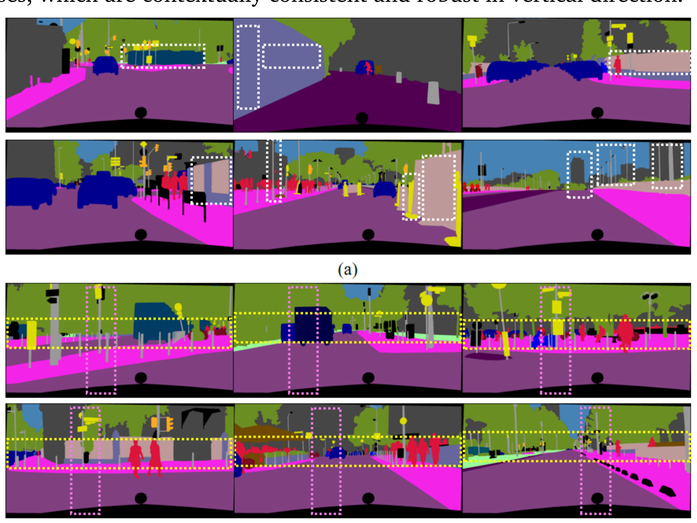
Summary

- Generating features that capture both global context and multi-level semantics leads to high computational complexity, which is problematic in real-time scene parsing.
- We propose Attention-Augmented Network (AttaNet) that consists of two primary modules: Strip Attention Module (SAM) and Attention Fusion Module (AFM) for powerful feature extraction in high efficiency.
- Our AttaNet achieves the leading performance on Cityscapes and ADE20K with great trade off on efficiency compared with other state-of-the-art methods.



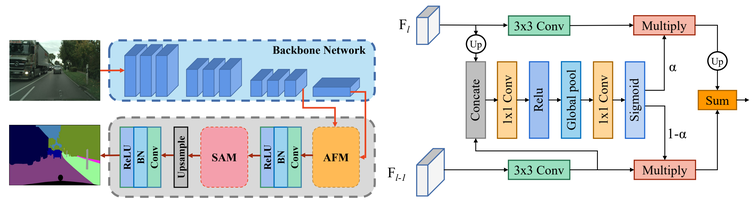
Motivation of SAM

We oberserved that various networks all achieved the lowest accuracies in some classes, which are contextually consistent and robust in vertical direction.



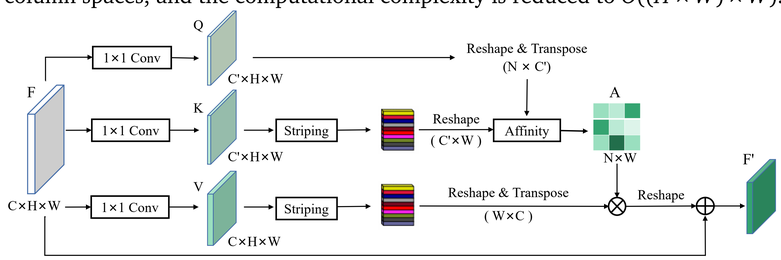
Pipeline and AFM Architecture

AFM adopts an attention strategy that learns to weight multi-level features at each pixel location with fewer computation. In AFM, we use predicted attention mask α and $1-\alpha$ to make sure that the complementary features are optimally integrated.



SAM Architecture

SAM utilizes a striping operation to encode the global context in the vertical direction and then harvests long-range relations along the horizontal axis. By applying SAM, each position in the feature map is connected with pixels in different column spaces, and the computational complexity is reduced to $O((H \times W) \times W)$.

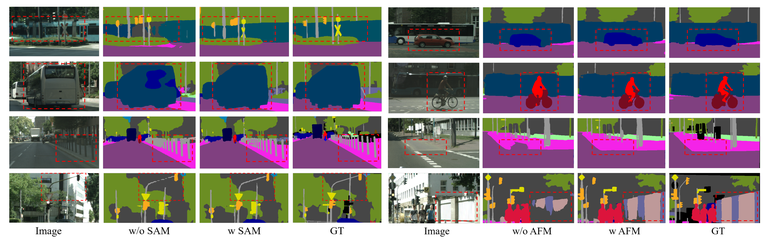


Abalation Study on SAM and AFM

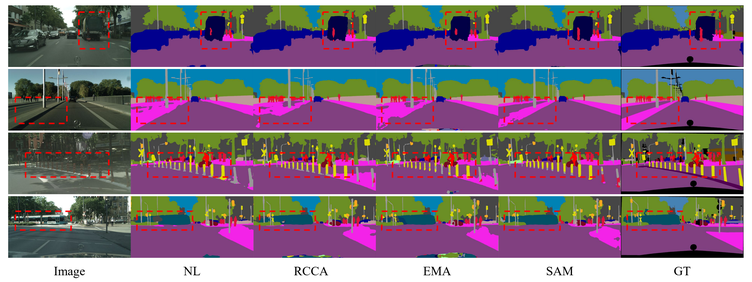
Comparison with other methods on the Cityscapes validation set, where ResNet-18 with aggregation architecture is used as the baseline. GFLOPs and Memory usage are calculated for an input of $1 \times 3 \times 1024 \times 1024$.

Approach	GFLOPs (Δ)	Memory (Δ)	mIOU (%)
Baseline	-	-	72.8
Baseline + NL [4]	3.357	334M	78.1
Baseline + RCCA [2]	0.472	26M	77.7
Baseline + EMA [3]	0.335	12M	75.0
Baseline + AFNB [5]	0.222	14M	77.8
Baseline + Strip Pooling [1]	0.22	18M	75.7
Baseline + SAM-horizontal	0.185	8M	78.3
Baseline + SAM-vertical	0.185	8M	78.5
Baseline + Concat	0.336	10M	73.7
Baseline + GFF	0.271	14M	75.3
Baseline + AFM	0.336	12M	77.8

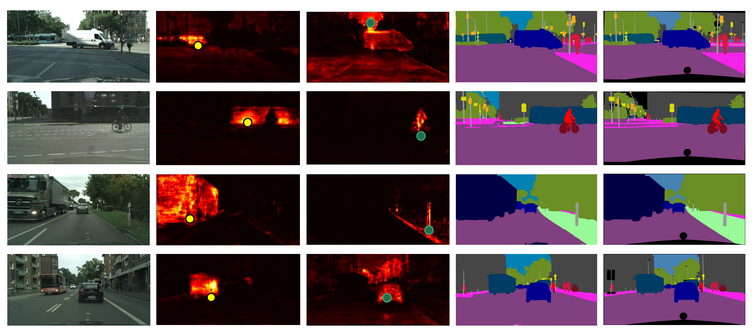
Qualitative Abalation on SAM and AFM



Qualitative Comparison on Attention Modules



Attention Map Visualization



References

[1] Hou, Q., Zhang, L., Cheng, M.-M., and Feng, J. (2020). Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*.

[2] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., and Liu, W. (2019). Ccnet: Criss-cross attention for semantic segmentation. In *CVPR*.

[3] Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., and Liu, H. (2019). Expectation-maximization attention networks for semantic segmentation. In *ICCV*.

[4] Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *CVPR*.

[5] Zhu, Z., Xu, M., Bai, S., Huang, T., and Bai, X. (2019). Asymmetric non-local neural networks for semantic segmentation. In *ICCV*.