Gazemo: Mimicking Human Saccades via Foveal-Peripheral Feature Modeling for Lightweight Semantic Segmentation

Supplementary Material

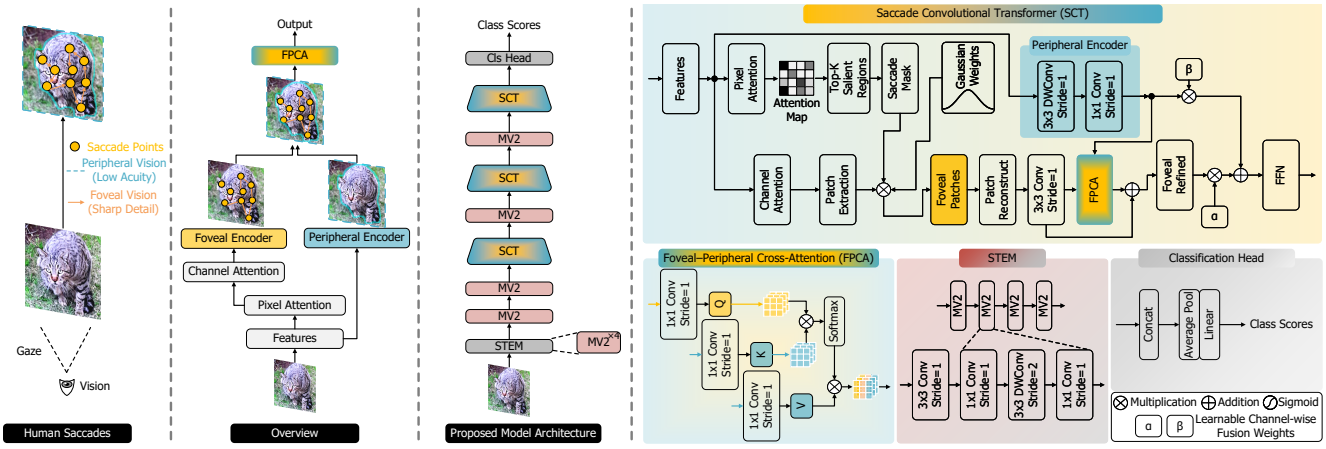

Figure 1. The architecture of the proposed Gazemo model for the task of image classification.

This supplementary material provides additional details on the ImageNet pre-training and classification experiments (Sec. A), visualization of the saccade locations (Sec. B), a detailed description of the model architecture (Sec. C), results of Gazemo with Batch Normalization layers (Sec. D), and the structure of the Feed-Forward Network (Sec. E). It also discusses the model's limitations and potential future directions (Sec. F).

A. ImageNet Pre-training

For ImageNet classification, Gazemo is initialized with weights pre-trained on the ImageNet-1K dataset. As shown in Fig. 1, the classification variant adopts a streamlined configuration, employing an average pooling layer followed by a linear projection to generate class scores from global representations. The Aggregated Feature Unification (AFU) decoder used in segmentation is omitted to maintain architectural simplicity and focus on feature discrimination. All models are trained and evaluated on 224×224 input images. The quantitative classification results are reported in the main paper.

B. Visualization of Selected Saccade Locations

To provide insight into the behavior of Saccade Convolutional Transformer (SCT), we visualize the selected top-k spatial locations on Cityscapes. As illustrated in Fig. 2, the selected points frequently correspond to object interiors, boundaries, and thin structures, supporting the interpretability of Gazemo's saliency-guided feature modeling.

C. Detailed Network Structure

The architectural configurations of the proposed Gazemo variants—Tiny (T), Small (S), Base (B), and Large (L)—are provided in Tab. 1, Tab. 2, Tab. 3, Tab. 4. Each variant is designed to balance computational efficiency and representational capacity under different resource constraints. All models adopt a MobileNetV2 backbone for hierarchical feature extraction, followed by the Saccade Convolutional Transformer (SCT) modules that capture foveal–peripheral interactions. The parameters N and H denote the number of layers and attention heads in each SCT stage, respectively. An input resolution of 512×512 is used for all configurations.

D. Results with Batch Normalization Layers

This section presents the results of our proposed Gazemo model incorporating Batch Normalization (BN) after every layer except the final segmentation layer. Results on ADE20K [33], Pascal Context [17], Cityscapes [4], and COCO-Stuff [1], as well as generalization on ImageNet classification [5], are shown in Tab. 5, Tab. 6, Tab. 7, Tab. 8, and Tab. 9, respectively. Applying BN after almost every layer does not yield better results than applying BN only after convolutional layers. This can be attributed to the fact that batch normalization, when applied repeatedly, can over-normalize activations, diminishing their diversity and

Figure 2. Cityscapes visualization of SCT-selected saccade points and corresponding predictions.

Table 1. Backbone configuration of Gazemo-T for 512×512 input resolution. SCT = Saccade Convolutional Transformer.

Layer Type	Kernel Size	Expand Ratio	Output Channels	Stride	Output Resolution
Stage 1					
Conv	3×3	–	16	2	
MobileNetV2	3×3	1	16	1	256×256
Stage 2					
MobileNetV2	3×3	4	16	2	
MobileNetV2	3×3	3	16	1	128×128
Stage 3					
MobileNetV2	5×5	3	32	2	
MobileNetV2	5×5	3	32	1	64×64
Stage 4					
MobileNetV2	3×3	3	64	2	
MobileNetV2	3×3	3	64	1	32×32
Stage 5					
MobileNetV2	5×5	3	128	2	
SCT Layer	H=4, N=2				16×16
Stage 6					
MobileNetV2	3×3	6	160	2	
SCT Layer	H=4, N=2				8×8

Table 2. Backbone configuration of Gazemo-S for 512×512 input resolution. SCT = Saccade Convolutional Transformer.

Layer Type	Kernel Size	Expand Ratio	Output Channels	Stride	Output Resolution
Stage 1					
Conv	3×3	–	16	2	
MobileNetV2	3×3	1	16	1	256×256
Stage 2					
MobileNetV2	3×3	4	24	2	
MobileNetV2	3×3	3	24	1	128×128
Stage 3					
MobileNetV2	5×5	3	48	2	
MobileNetV2	5×5	3	48	1	64×64
Stage 4					
MobileNetV2	3×3	3	96	2	
MobileNetV2	3×3	3	96	1	32×32
Stage 5					
MobileNetV2	5×5	4	160	2	
SCT Layer	H=6, N=3				16×16
Stage 6					
MobileNetV2	3×3	6	192	2	
SCT Layer	H=6, N=3				8×8

expressive power. Additionally, the batch-dependent nature of BN may not be as effective in hybrid architectures like Gazemo, where convolutional and transformer layers interact differently with normalization.

Table 3. Backbone configuration of Gazemo-B for 512×512 input resolution. SCT = Saccade Convolutional Transformer.

Layer Type	Kernel Size	Expand Ratio	Output Channels	Stride	Output Resolution
Stage 1					
Conv	3×3	–	16	2	
MobileNetV2	3×3	1	16	1	256×256
Stage 2					
MobileNetV2	3×3	4	32	2	
MobileNetV2	3×3	3	32	1	128×128
Stage 3					
MobileNetV2	5×5	3	64	2	
MobileNetV2	5×5	3	64	1	64×64
Stage 4					
MobileNetV2	3×3	3	128	2	
MobileNetV2	3×3	3	128	1	32×32
Stage 5					
MobileNetV2	5×5	4	192	2	
SCT Layer	H=8, N=4				16×16
Stage 6					
MobileNetV2	3×3	6	256	2	
SCT Layer	H=8, N=4				8×8

Table 4. Backbone configuration of Gazemo-L for 512×512 input resolution. SCT = Saccade Convolutional Transformer.

Layer Type	Kernel Size	Expand Ratio	Output Channels	Stride	Output Resolution
Stage 1					
Conv	3×3	–	32	2	
MobileNetV2	3×3	1	32	1	256×256
Stage 2					
MobileNetV2	3×3	4	64	2	
MobileNetV2	3×3	4	64	1	128×128
Stage 3					
MobileNetV2	5×5	4	128	2	
MobileNetV2	5×5	4	128	1	64×64
Stage 4					
MobileNetV2	3×3	4	192	2	
MobileNetV2	3×3	4	192	1	
SCT Layer	H=8, N=3				32×32
Stage 5					
MobileNetV2	5×5	4	256	2	
SCT Layer	H=8, N=4				16×16
Stage 6					
MobileNetV2	3×3	6	320	2	
SCT Layer	H=8, N=4				8×8

E. Details of Feed-Forward Network

In the Gazemo architecture, the Feed-Forward Network (FFN) follows each Saccade Convolutional Transformer (SCT) layer to refine the transformed features. The FFN adopts a compact yet expressive design composed of a 1×1

Table 5. Results on ADE20K validation set [33]. Batch Normalization layers are used after every layer. GFLOPs are reported for input resolution of 512×512. '-' denotes unreported results. Single-scale inference is used for mIoU evaluation.

Models	Encoder	Year	mIoU	GFLOPs	Parameters	Latency(ms)
UPerNet [24]	SwinTransformer-T [13]	ICCV 2021	46.1	945.0	60.0M	-
PVMamba [26]	PVMamba-T [26]	ICCV 2025	48.2	941.0	54.0M	-
UPerNet [24]	ConvNeXt-T [14]	ICCV 2021	46.7	939.0	60.0M	-
PSPNet [32]	MobileNetV2 [21]	CVPR 2018	29.6	52.2	13.7M	426
FCN-8s [15]	MobileNetV2 [21]	CVPR 2018	19.7	39.6	9.8M	406
Semantic FPN [11]	ConvMLP-S [12]	CVPR 2023	35.8	33.8	12.8M	311
DeepLabV3+ [3]	EfficientNet [23]	ICML 2019	36.2	26.9	17.1M	388
DeepLabV3+ [3]	MobileNetV2 [21]	ECCV 2018	38.1	25.8	15.4M	414
Lite-ASPP [3]	ResNet18 [8]	ECCV 2018	37.5	19.2	12.5M	259
PEM [2]	STDC1 [2]	CVPR 2024	39.6	16.0	17.0M	-
DeepLabV3+ [3]	ShuffleNetv2-1.5x [16]	ECCV 2018	37.6	15.3	16.9M	384
HRNet-Small [30]	HRNet-W18-Small [30]	ECCV 2020	33.4	10.2	4.0M	256
SegFormer [25]	MiT-B0 [25]	NeurIPS 2021	37.4	8.4	3.8M	308
FeedFormer-B0 [22]	MiT-B0 [25]	AAAI 2023	39.2	7.8	4.5M	-
SegMAN [7]	SegMAN-T [7]	CVPR 2025	43.0	6.2	6.4M	-
U-MixFormer [28]	MiT-B0 [25]	WACV 2025	41.2	6.1	6.1M	-
EDAFormer [29]	EDAFormer-T [29]	ECCV 2024	42.3	5.6	4.9M	-
MegaSeg [10]	MegaSeg-T [10]	WACV 2024	42.4	5.5	4.7M	-
OffSeg [31]	OffSeg-T [31]	ICCV2025	44.2	5.3	6.2M	-
VWFormer [27]	VWFormer-B0 [27]	ICLR 2024	38.9	5.1	3.7M	-
Lite-ASPP [3]	MobileNetV2 [21]	CVPR 2018	36.6	4.4	2.9M	94
CGRSeg-T [19]	CGRSeg [19]	ECCV 2024	43.6	4.0	9.4M	-
R-ASPP [21]	MobileNetV2 [21]	CVPR 2018	32.0	2.8	2.2M	71
HR-NAS-B [6]	Searched [6]	CVPR 2021	34.9	2.2	3.9M	-
LR-ASPP [9]	MobileNetV3-Large [9]	ICCV 2019	33.1	2.0	3.2M	51
HR-NAS-A [6]	Searched [6]	CVPR 2021	33.2	1.4	2.5M	-
LR-ASPP [9]	MobileNetV3-Large-reduce [9]	ICCV 2019	32.3	1.3	1.6M	33
LeMoRe [18]	LeMoRe [18]	ICIP 2025	32.7	0.8	1.6M	24
Gazemo (Ours)	Gazemo-T	Ours	33.0	0.6	1.8M	14
Gazemo (Ours)	Gazemo-S	Ours	36.2	1.2	3.9M	17
Gazemo (Ours)	Gazemo-B	Ours	39.3	2.0	8.4M	21
Gazemo (Ours)	Gazemo-L	Ours	41.4	7.1	15.0M	23

Table 6. Results on PASCAL Context test set [17].

Methods	Backbone	mIoU^{59}	mIoU^{60}	GFLOPs
DeepLabV3+ [3]	ENet-s16 [20]	43.07	39.19	23.00
DeepLabV3+ [3]	MobileNetV2-s16 [21]	42.34	38.59	22.24
LR-ASPP [9]	MobileNetV3-s16 [9]	38.02	35.05	2.04
Gazemo (Ours)	Gazemo-T	39.64	34.95	0.50
Gazemo (Ours)	Gazemo-S	42.32	38.26	0.99
Gazemo (Ours)	Gazemo-B	44.37	39.87	1.71
Gazemo (Ours)	Gazemo-L	48.77	44.30	6.25

Table 7. Results on Cityscapes validation set [4].

Methods	Encoder	mIoU	GFLOPs
PSPNet [32]	MobileNetV2 [21]	70.2	423.4
FCN [15]	MobileNetV2 [21]	61.5	317.1
SegFormer [25]	MiT-B0 [25]	71.9	17.7
L-ASPP [3]	MobileNetV2 [21]	72.7	12.6
LR-ASPP [9]	MobileNetV3-Large [9]	72.4	9.7
LR-ASPP [9]	MobileNetV3-Small [9]	68.4	2.9
Gazemo (Ours)	Gazemo-T	66.57	1.17
Gazemo (Ours)	Gazemo-S	69.12	2.16
Gazemo (Ours)	Gazemo-B	71.04	3.74
Gazemo (Ours)	Gazemo-L	74.94	14.02

convolution for channel expansion, a 3×3 depthwise convolution for spatial adaptation, and another 1×1 convolution for projection. This structure enables localized feature interaction while preserving computational efficiency. The detailed layout of the FFN is illustrated in Fig. 3.

Table 8. Results on COCO-Stuff test set [1].

Methods	Encoder	mIoU	GFLOPs
PSPNet [32]	MobileNetV2-s8 [21]	30.14	52.94
DeepLabV3+ [3]	EfficientNet-s16 [23]	31.45	27.10
DeepLabV3+ [3]	MobileNetV2-s16 [21]	29.88	25.90
LR-ASPP [9]	MobileNetV3-s16 [9]	25.16	2.37
Gazemo (Ours)	Gazemo-T	26.82	0.64
Gazemo (Ours)	Gazemo-S	29.78	1.16
Gazemo (Ours)	Gazemo-B	32.51	1.97
Gazemo (Ours)	Gazemo-L	35.00	7.13

Table 9. Image classification performance on the ImageNet-1K benchmark, evaluated at an input resolution of 224×224. *Params* and *L* denote the parameters and inference latency, respectively.

Methods	Top-1 Acc.(%)	GFLOPs	Params	L(ms)
Gazemo-T (Ours)	64.8	0.11	1.9M	7.8
Gazemo-S (Ours)	70.1	0.20	4.1M	10.6
Gazemo-B (Ours)	74.4	0.35	8.5M	13.4
Gazemo-L (Ours)	79.7	1.32	15.0M	15.1

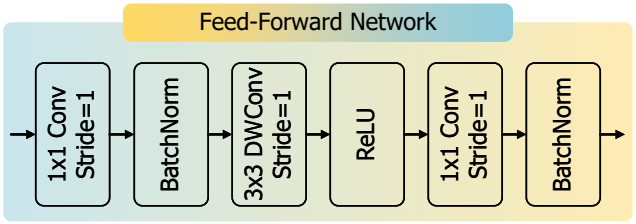

Figure 3. Feed-Forward Network architecture.

F. Limitations and Future Work

The proposed model demonstrates strong performance and efficiency; however, several challenges remain. Like most lightweight architectures, its performance still depends on ImageNet-1K pre-training to achieve optimal accuracy, as training from scratch can limit convergence and representation quality. Although the model has shown promising generalization through image classification experiments, future research will explore improving robustness across diverse datasets and deployment conditions. Additionally, extending the framework toward dynamic multi-saccade mechanisms and optimizing it for real-world, resource-constrained applications remain key directions for advancing practical utility.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1, 3

[2] Niccolò Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. Pem: Prototype-based efficient maskformer

for image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15804–15813, 2024. 3

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 3

[5] 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. 1

[6] Mingyu Ding, Xiaochen Lian, Linjie Yang, Peng Wang, Xiaojie Jin, Zhiwu Lu, and Ping Luo. Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2982–2992, 2021. 3

[7] Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19077–19087, 2025. 3

[8] 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. 3

[9] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 3

[10] Beoungwoo Kang, Seunghun Moon, Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. Metaseg: Metaformer-based global contexts-aware network for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 434–443, 2024. 3

[11] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 3

[12] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6307–6316, 2023. 3

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[15] 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. 3

[16] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 3

[17] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 1, 3

[18] Mian Muhammad Naeem Abid, Nancy Mehta, Zongwei Wu, and Radu Timofte. Lemore: Learn more details for lightweight semantic segmentation. In *2025 IEEE International Conference on Image Processing (ICIP)*, pages 2163–2168, 2025. 3

[19] Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *European Conference on Computer Vision*, pages 239–255. Springer, 2024. 3

[20] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 3

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[22] Jae-hun Shim, Hyunwoo Yu, Kyeongbo Kong, and Suk-Ju Kang. Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2263–2271, 2023. 3

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[24] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3

[25] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 3

[26] Fei Xie, Zhongdao Wang, Weijia Zhang, and Chao Ma. Pvmamba: Parallelizing vision mamba via dynamic state aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10218–10228, 2025. 3

[27] Haotian Yan, Ming Wu, and Chuang Zhang. Multi-scale representations by varying window attention for semantic segmentation. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[28] Seul-Ki Yeom and Julian Von Klitzing. U-mixformer: Unet-like transformer with mix-attention for efficient semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2025. 3

[29] Hyunwoo Yu, Yubin Cho, Beoungwoo Kang, Seunghun Moon, Kyeongbo Kong, and Suk-Ju Kang. Embedding-free transformer with inference spatial reduction for efficient semantic segmentation. In *European Conference on Computer Vision*, pages 92–110. Springer, 2024. 3

[30] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 3

[31] Shi-Chen Zhang, Yunheng Li, Yu-Huan Wu, Qibin Hou, and Ming-Ming Cheng. Revisiting efficient semantic segmentation: Learning offsets for better spatial and class feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22361–22371, 2025. 3

[32] 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. 3

[33] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1, 3