Supplementary Material for Dynamic Focus-aware Positional Queries for Semantic Segmentation

Haoyu He[1] Jianfei Cai[1] Zizheng Pan[1] Jing Liu[1]
Jing Zhang[2] Dacheng Tao[2] Bohan Zhuang[1†]

[1] ZIP Lab, Monash University [2] The University of Sydney

We organize our supplementary material as follows.

- In Section A, we investigate the potential for applying our DFPQ to instance segmentation.
- In Section B, we show more comparisons on the visualized cross-attention maps for different positional queries.
- In Section C, we show more qualitative results.
- In Section D, we investigate the effect of different low-resolution features choices in HRCA.
- In Section E, we investigate the effect of starting to employ DFPQ at midway training.

A. Instance Segmentation with DFPQ

To further demonstrate the flexibility of our method, we combine our DFPQ with Mask2former and compare with the baseline instance segmentation methods on COCO `val2017` [6] following the exact settings in [3]. The results are reported in Table I. We observe that our DFPQ consistently improves Mask2former with R50 and Swin-B backbones by 0.3% and 0.2% AP with barely extra parameters and FLOPs. The results suggest the potential to extend our DFPQ to other segmentation scenarios. However, the improvements are not as impressive as in semantic segmentation. Instance segmentation is a very challenging task that requires grouping highly-entangled pixels into groups of instances and is more challenging than semantic segmentation or object detection as recognized by literature [5]. Previous work [1, 4, 5, 9] provides the positional priors by integrating the instance segmentation with a heavy object detection head or branch in a two-stage top-down [1, 4, 9] or a bottom-up style [5] framework. Compared to the literature that employs specific architecture designs to provide the positional priors, we conjecture that our DFPQ has an inferior representational capability that cannot fully encode the required positional priors for the challenging instance segmentation task. However, it is an interesting future direction to additionally encode the instance-level information into our DFPQ, *e.g.*, encoding the bounding boxes or explicitly distinguishing the positional priors among the instance segments.

Table I. Combine our DFPQ with Mask2former [3] and compare with the state-of-the-art instance segmentation methods on COCO `val` [11] with 133 categories. #P and #F indicate the number of parameters (M) and FLOPs (G).

Method	Backbone	Epochs	AP	AP^{s}	AP^{m}	AP^{l}	#P	#F
SOLOv2 [8]	R50	36	37.5	15.8	41.4	56.6	34	-
K-Net [10]	R50	36	38.6	19.1	42.0	57.7	37	-
HTC [2]	R101	36	39.7	22.6	42.2	50.6	80	441
Mask2former [3]	R50	36	42.4	22.1	45.4	64.3	44	225
Mask2former + DFPQ	R50	36	42.7	21.9	46.4	64.4	44	225
Mask2former [3]	Swin-B	36	47.2	27.1	50.8	69.4	115	466
Mask2former + DFPQ	Swin-B	36	47.4	26.8	51.0	70.5	115	466

[†] Corresponding author. E-mail: bohan.zhuang@gmail.com

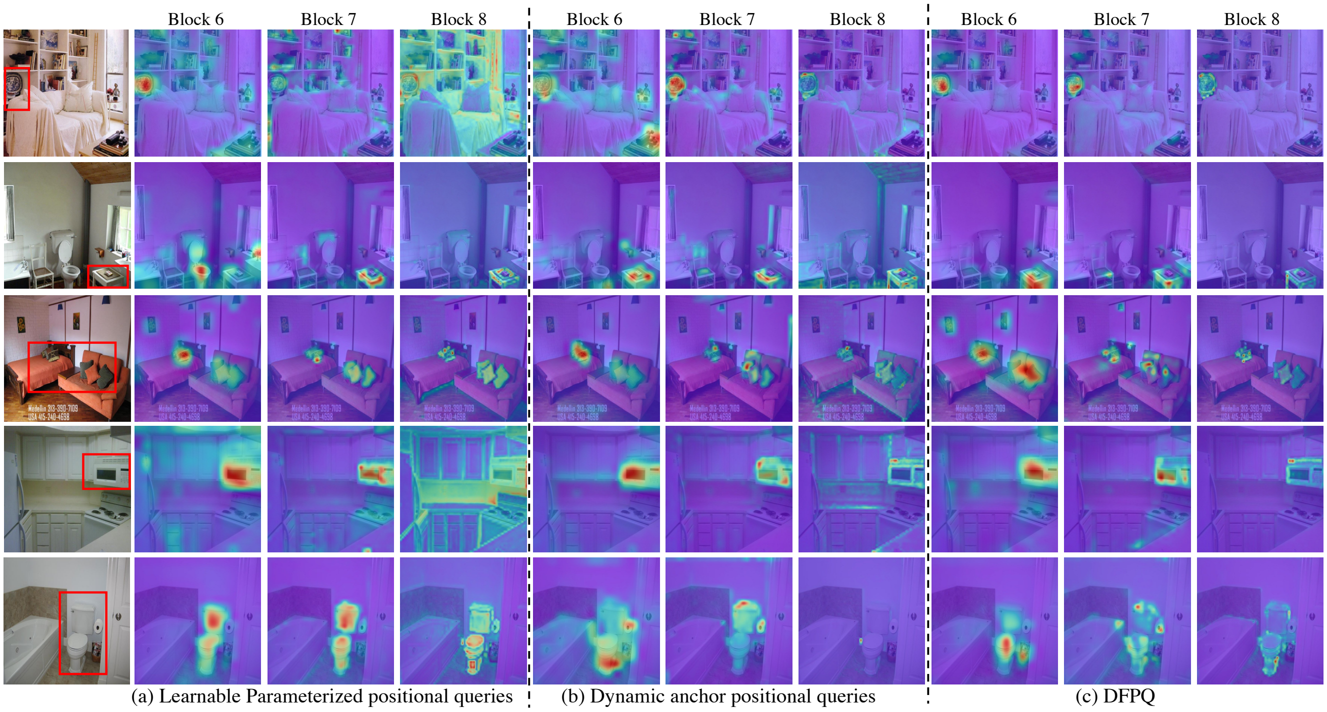
Block 6
Block 7
Block 8
Block 6
Block 7
Block 8
Block 6
Block 7
Block 8
(a) Learnable Parameterized positional queries
(b) Dynamic anchor positional queries
(c) DFPQ

Figure A. Visualizations of the cross-attention maps for learnable positional queries ([1, 3]), dynamic anchor positional queries (alike [7]) and our DFPQ. We show the visualizations for the normalized cross-attention maps in the last three decoder blocks and indicate the target segments in the red boxes. The cross-attention maps with the learnable positional queries and the dynamic anchor positional queries are often scattered without a clear focus and mix up different segments, while the cross-attention maps with DFPQ are more compact and consistent to reflect the target segments.

B. More Comparisons on Visualized Cross-attention Maps

We have discussed and provided both quantitative and qualitative comparisons with other positional queries variants in Section 4.2 of the main paper. We show more comparisons on visualized cross-attention maps with different positional queries in Figure A and observe that the quality of the cross-attention maps for our DFPQ is clearly better than the learnable parameterized positional queries and dynamic anchor positional queries. Our DFPQ progressively refines the cross-attention maps, which become more accurate and compact in the deeper layers. Interestingly, we observe the cross-attention maps with DFPQ can end-to-end learn to localize the boundaries of the target segments in Block 8.

C. More Qualitative Results

We visualize sample predictions of our FASeg model with Swin-L backbone and compare with Mask2former [3] on ADE20K `val` with multi-scale inference in Figure B. We observe that FASeg generates consistent predictions that align with the ground truth. We also present some failure cases in the blue boxes and find that the very small target segments still cannot be localized precisely. We thus take refining the small regions as future work.

D. Effect of Low-resolution Feature scale Choices for HRCA

In HRCA, we identify the most informative pixel positions within the low-resolution features, then map these positions to the high-resolution features and only perform cross-attention on these positions. We keep the number of attended pixels in the high-resolution features the same and investigate the effect of different low-resolution feature scales on ADE20k `val`. The results are reported in Table II. We observe that more coarse-grained features yield better results, where we conjecture that coarse-grained features contain more context information which helps locate the informative pixels. In this case, we use 1/32×1/32 low-resolution features to select the informative pixels by default for the other experiments.

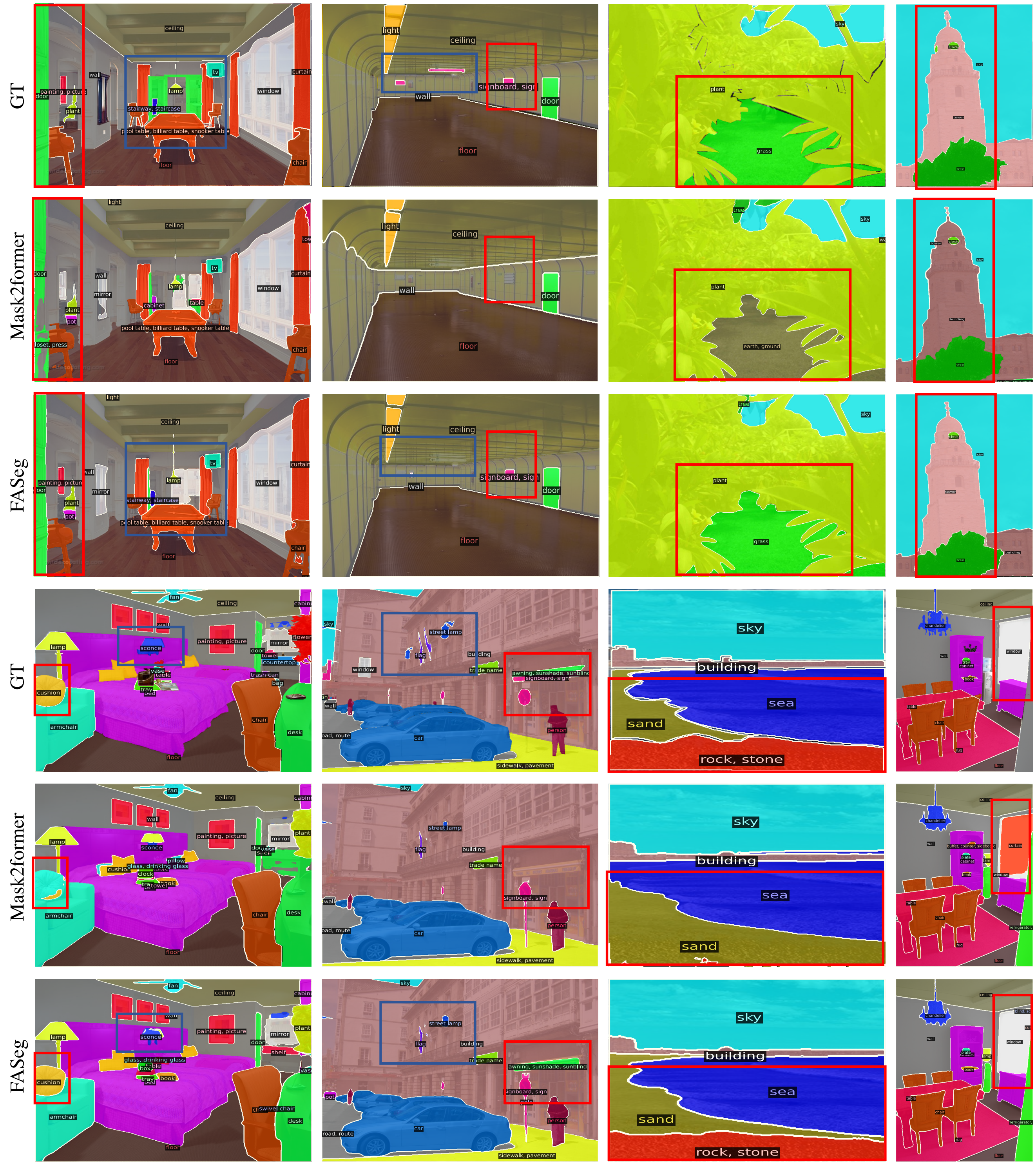

Figure B. Qualitative results on the ADE20K `val` [11]. Compared to Mask2former [3], our FASeg predicts masks with finer details and yields more accurate predictions. The differences are highlighted with red boxes and the failure cases are highlighted with blue boxes. Best viewed in color.

Table II. Effect of low-resolution features for HRCA on ADE20K `val` [11] with 150 categories.

Low-level features	mIoU s.s. (%)
1/8×1/8	47.9
1/16×1/16	48.0
1/32×1/32	48.3

Table III. Effect of starting to employ DFPQ at midway training for FASeg with Swin-B Backbone on ADE20K `val` [11] with 150 categories.

Starting iteration	0	20k	40k	80k
mIoU s.s.(%)	55.0	55.1	55.1	54.7

E. Effect of Starting to Employ DFPQ at the Midway of Training

By default, DFPQ is applied at the beginning of training. We experiment to start employing DFPQ from 20k, 40k, and 80k iterations of the total 160k training iterations for FASeg with Swin-B backbone on ADE20k `val`. The results are reported in Table III. We find that the performance fluctuates within 0.3% mIoU, which suggests that our DFPQ is robust to the starting iteration. We also speculate that our DFPQ is robust to the early-stage training and will eventually learn reasonable positional information.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1, 2

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 1

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 2, 3

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1

[5] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

[7] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*, 2022. 2

[8] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *NeurIPS*, 33:17721–17732, 2020. 1

[9] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, pages 418–434, 2018. 1

[10] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In *NeurIPS*, volume 34, 2021. 1

[11] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 1, 3, 4