Supplementary Materials for *Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation*

Jie Liu[1*], Yanqi Bao[2*], Guo-Sen Xie[3,4†], Huan Xiong[4], Jan-Jakob Sonke[5], Efstratios Gavves[1]
[1]University of Amsterdam, Netherlands [2]Northeastern University, China [5]The Netherlands Cancer Institute, Netherlands
[3]Nanjing University of Science and Technology, China [4]Mohamed bin Zayed University of Artificial Intelligence, UAE

Appendix A. Implementation details

We employ ResNet-50 [1] (VGG-16 [2]) pre-trained on ImageNet as our backbone networks. For ResNet-50, the dilation convolution (with stride size = 1) is introduced to ensure that the feature receptive fields of layer2, layer3, and layer4 preserve the same spatial resolution. The backbone weights are fixed except for layer4, which is required to learn more robust activation maps. The model is trained with a SGD optimizer for 200 and 50 epochs on the PASCAL-5^i and the COCO-20^i benchmarks, respectively . The learning rates are initialized as 0.005 and 0.002 with a poly learning rate schedule on PASCAL-5^i and COCO-20^i, respectively. The batch size is set as 8 on PASCAL-5^i and 32 on COCO-20^i. Our entire network is trained with the same learning rate during each epoch, except for layer4 of the backbone network, whose parameters starts back-propagation after training for multiple epochs to ensure a lower learning rate for fine-tuning. Data augmentation strategies in [3] are adopted in the training stage, and all images are cropped to 473×473 patches for two benchmarks. Besides, we leverage multi-scale testing strategy used in the most few-shot semantic segmentation methods for the model evaluation, and the original Groudtruth of the evaluated query image without any resize operations is adopted for the metric calculation. In addition, the window sizes in SAM are set to $\{5 \times 1, 3 \times 3, 1 \times 5\}$, and the kernel sizes in DCM are set as 5×1, 5×5, and 1×5, respectively. We implement our model with PyTorch 1.7.0 and conduct all the experiments with Nvidia Tesla V100 GPUs and CUDA11.3.

Appendix B. Additional results and analyses

Kernel Generation Variants. The dynamic convolution module (DCM), in which we generate dynamic kernels from the support foreground and employ convolution over the query feature, is an essential component in our proposed method. Therefore, here we presents two kernel generation variants: (i) we generate both asymmetric and symmetric kernel weights in parallel (ii) we first generate asymmetric kernel weights, and the asymmetric kernel weights are further used to generate symmetric kernel weights. With a kernel size 5, we term these two variants as $5/25$ (in parallel) and $5 \rightarrow 25$ (serial). As seen in Table A1, these two variants achieve similar performance (66.0 *vs* 66.7), which demonstrates the robustness of the DCM.

Method	1-shot mIoU					FB-IoU
	Fold0	Fold1	Fold2	Fold3	Mean	
5/25	65.9	70.6	66.9	60.5	66.0	77.0
5→25	65.7	71.6	69.1	60.6	66.7	78.0

Table A1. Ablation studies for the kernel generation variants.

Method	1-shot mIoU				
	Fold0	Fold1	Fold2	Fold3	Mean
PANet (Box)	-	-	-	-	45.1
CANet (Box)	-	-	-	-	52.0
Ours (Box)	59.8	70.5	63.2	55.5	62.3
Ours (Pixel)	65.7	71.6	69.1	60.6	66.7

Table A2. Comparison with the existing methods under the bounding box supervision under 1-shot setting.

Method	Backbone	PASCAL-5^i		COCO-20^i	
		MIoU	FB-IoU	MIoU	FB-IoU
w/o MS	VGG16	61.3	72.7	39.2	61.9
w MS	VGG16	61.7	73.7	39.5	62.5
w/o MS	ResNet50	65.7	77.4	41.5	62.7
w MS	ResNet50	66.7	78.0	43.0	63.2

Table A3. Effectiveness of multi-scale testing under the 1-shot setting on the PASCAL-5^i and the COCO-20^i benchmarks.

Experiments with Bounding Box Annotations. Following PANet [4] and CANet [5], we evaluate our model with weakly-supervised annotation (i.e, bounding box instead of pixel-wise annotation) on the support set. As shown

*Equal Contribution.
†Corresponding author.

Methods	1-shot						5-shot					
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
CANet (CVPR19) [5]	53.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1	69.3
CANet (CVPR19)+DCM	64.7	66.8	51.8	51.9	58.8	69.3	65.3	67.2	52.7	52.9	59.5	70.1
PFENet (TPAMI'20) [3]	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
PFENet (TPAMI'20)+DCM	62.2	69.6	59.2	58.0	62.3	73.5	63.1	70.0	60.0	58.5	62.9	73.6

Table A4. Generalization ability of proposed DCM under both 1-shot and 5-shot settings.

Methods	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
Original Groudtruth [5]	VGG16	58.9	69.1	63.2	55.7	61.7	73.7	63.4	70.7	68.1	59.0	65.3	77.2
Non-original Groudtruth	VGG16	59.3	69.5	63.3	55.8	62.0	73.8	64.0	71.2	68.4	59.0	65.7	77.2
Original Groudtruth [3]	ResNet50	65.7	71.6	69.1	60.6	66.7	78.0	70.0	73.2	70.9	65.5	69.9	80.7
Non-original Groudtruth	ResNet50	65.6	71.8	69.2	60.5	66.8	78.0	70.0	73.2	70.9	65.5	69.9	80.6

Table A5. Comparison with Original Groudtruth and Non-original Groudtruth

in Table A2, our method with the bounding box annotation achieves slightly inferior performance than that with expensive pixel-wise annotation. In addition, using bounding box annotation as supervision, our method also significantly outperforms both PANet and CANet. This experiment indicates the potential of our method in the segmentation task with weak supervision.

Multi-Scale Testing. As a post-processing method, multi-scale testing [6] is widely adopted in many semantic segmentation tasks. Many few-shot semantic segmentation methods as well as our proposed method DPCN also use this strategy to improve segmentation performance. We present our results with / without multi-scale testing (termed as w MS and w/o MS, respectively) in Table A3. The scales in our experiments are set as 473 and 641. We can see that (i) our model without the multi-scale testing also achieves state-of-the-art results using VGG16 and ResNet50 backbones, which further demonstrates the effectiveness of our proposed method (ii) multi-scale testing brings 1% mIou improvement for our model with ResNet50 backbone, while a slight improvement when we use VGG16 as the backbone network.

Generalization Ability of DCM. The dynamic convolution module (DCM) can be used as a plug-and-play module to improve current prototype-based few-shot segmentation methods. We merge DCM into CANet and PFENet, and present the corresponding results under both 1-shot and 5-shot settings in Table A4. DCM further improve the performance of CANet and PFENet under both 1-shot and 5-shot settings, which shows the effectiveness of the DCM.

Evaluation using Original Groundtruth. As in PFENet [3], we also evaluate our model with original groundtruth of the query image and the non-original one resized to the same size as training image size (473×473). We can find from Table A5 that our proposed model obtains similar performance with the original or non-original query groundtruth.

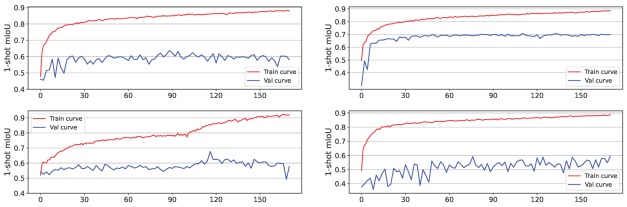

Figure A1. Training and Validation curves (x-axis: epochs, y-axis: 1-shot mIoU) on PASCAL-5^i benchmark.

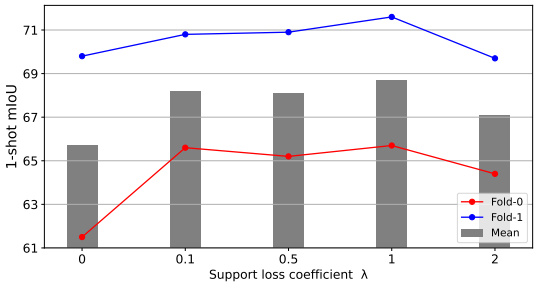

Figure A2. Ablation study for Support loss coefficient λ on fold0 and fold1 of PASCAL-5^i benchmark.

Training and validation performance. We present the performance (mIoU) changing process as the training epoch increases in Fig. A1. As can be seen, the mIoU of the training process are much better than that of the validation in each fold. Besides, the validation mIoU in fold0 and fold1 are relatively stable, while the validation mIoU in fold2 and fold3 fluctuate as the training process goes on.

Support Loss Coefficient λ. During the model training, we use the predicted query mask as a pseudo mask for predicting the support mask, which requires a support loss

$\mathcal{L}_{seg}^{q\to s}$ for supervision. For the support loss coefficient λ, we take its value from $\{0, 0.1, 0.5, 1, 2\}$ to study its influence on our model. The performance of the fold0 and fold1 as well as their mean on the PASCAL-5^i benchmark are used for illustration. As shown in Fig. A2, our model achieves best results when the support loss coefficient is set as 1. And λ is set as 1 in all our experiments.

Appendix C. Additional qualitative results

In this section, we present more qualitative results of our proposed DPCN and its baseline to demonstrate its few-shot segmentation performance. Appearance and scale variations (more obvious in the COCO-20^i benchmark) are the innate difficulty of the few-shot semantic segmentation task. Therefore, we first show some examples sampled from COCO-20^i benchmark with large object appearance and scale variations in the Fig. A3 and Fig. A4, respectively. As can be seen, our model DPCN exhibits great superiority in alleviating appearance and scale variations. Besides, we also sample some examples from PASCAL-5^i benchmark, and the qualitative results are presented in Fig. A5. Furthermore, DPCN occasionally predicts more accurate segmentation than human-annotated ground-truth (Fig. A6), which further demonstrates the effectiveness of our method. Finally, we give some visualization of the support activation maps, initial pseudo mask as well as the refined pseudo mask in Fig. A7. We can see that the support activation maps can capture complementary object details in the query image, the initial pseudo mask gives rough pixel-wise location estimation of object, while the refined pseudo mask can estimate more accurate object location in the query image.

References

[1] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 1

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[3] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2020. 1, 2

[4] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9197–9206, 2019. 1

[5] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[6] Qi Zhao, Binghao Liu, Shuchang Lyu, Xu Wang, and Yifan Yang. A self-distillation embedded supervised affinity attention model for few-shot segmentation. *arXiv preprint arXiv:2108.06600*, 2021. 2

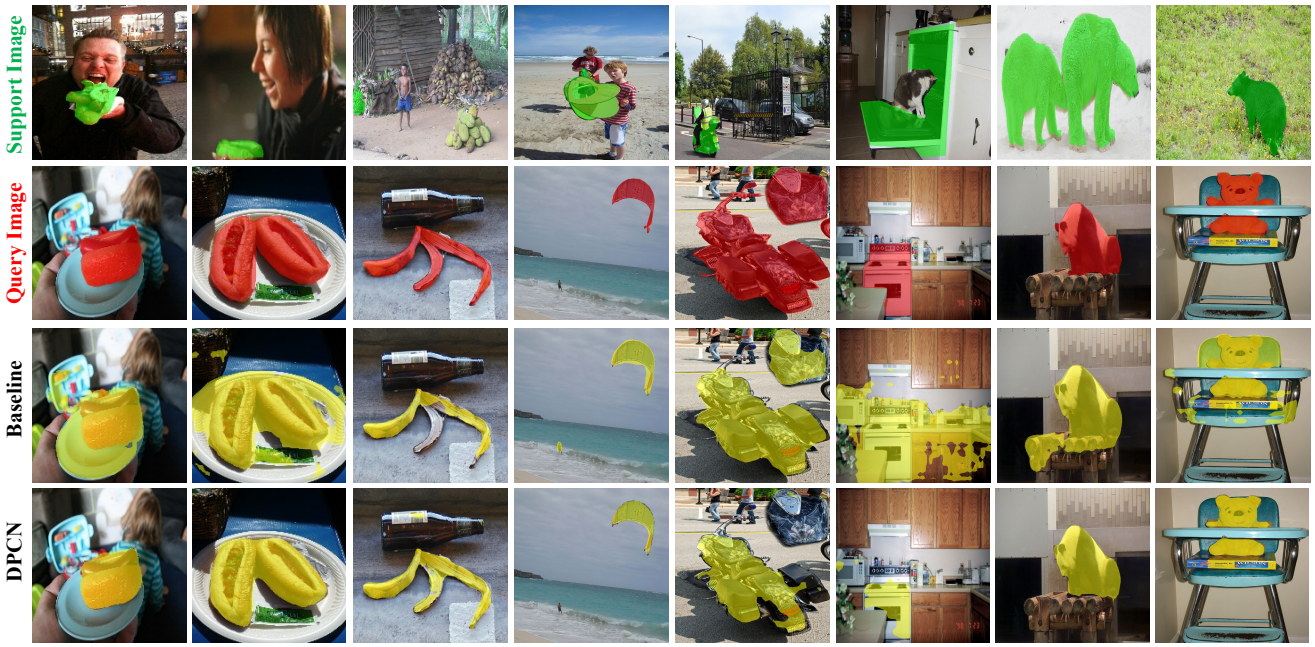

Figure A3. Qualitative results of our method DPCN and baseline model on COCO-20^i benchmark with **large object appearance variations**. Zoom in for details.

Figure A4. Qualitative results of our method DPCN and baseline model on COCO-20^i benchmark with **large object scale variations**. Zoom in for details.

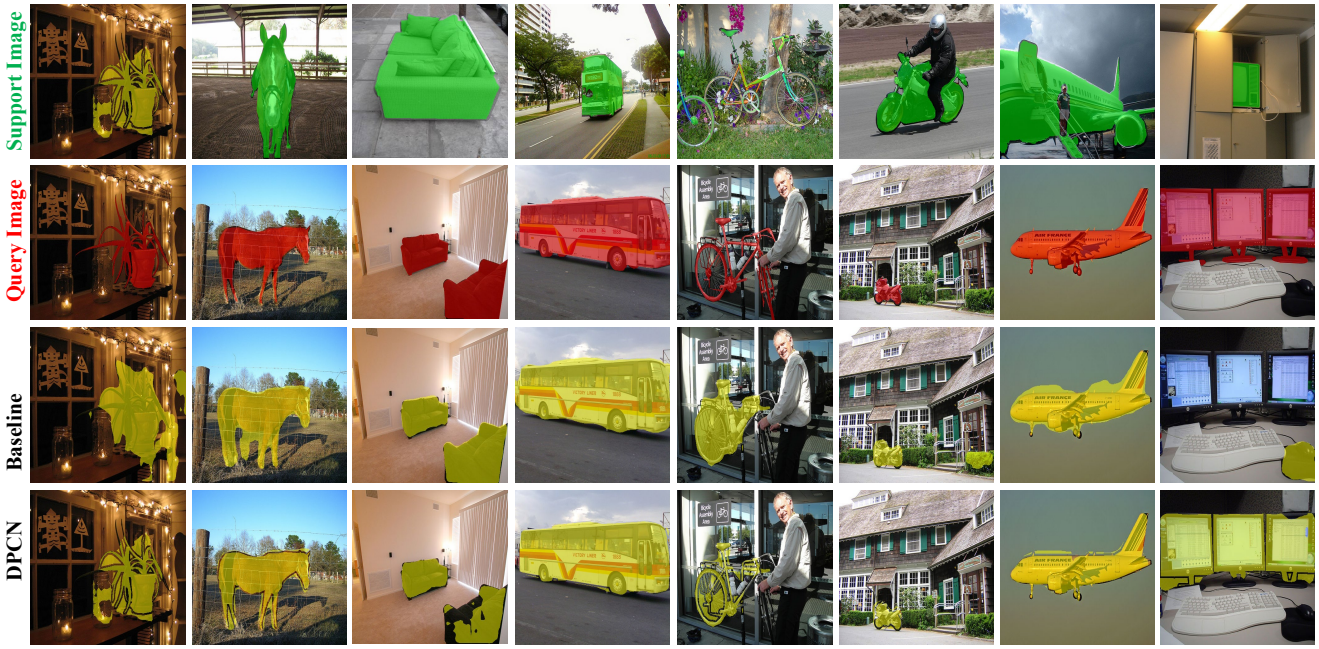

Figure A5. Qualitative results of our method DPCN and baseline model on PASCAL-5^i benchmark. Zoom in for details.

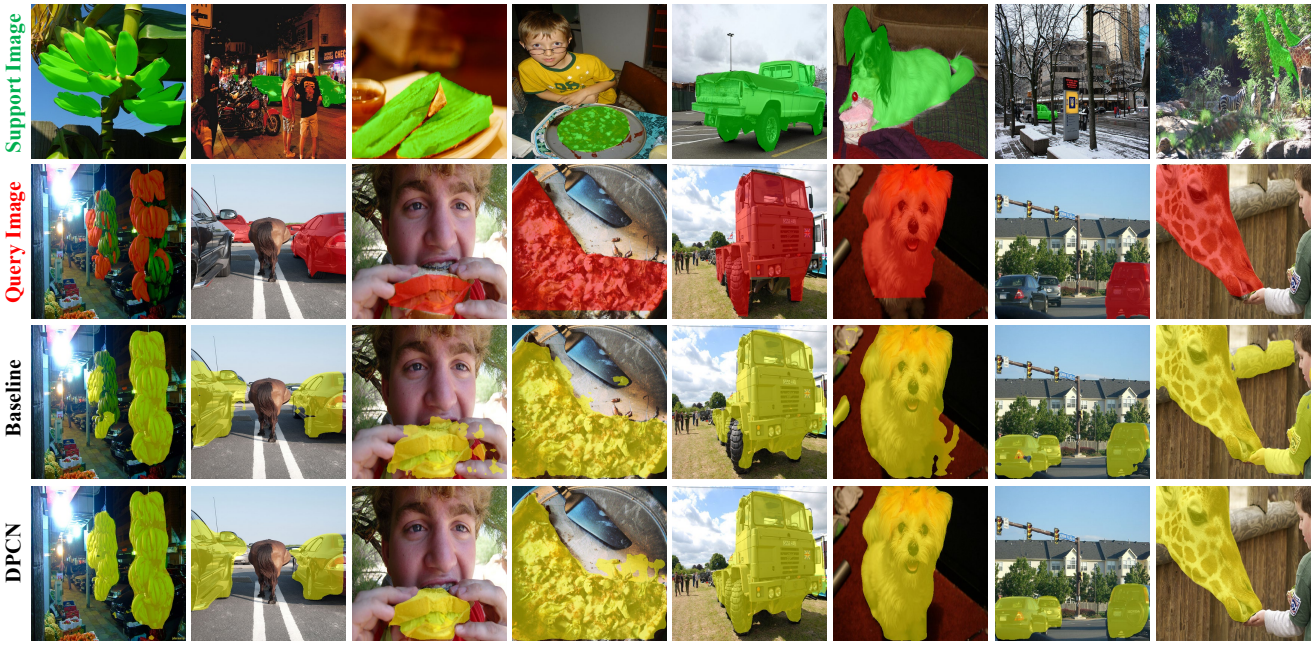

Figure A6. Our proposed DPCN occasionally predicts more accurate segmentation masks than human-annotated ground-truths. Examples are sampled from both PASCAL-5^i and COCO-20^i. Zoom in for details.

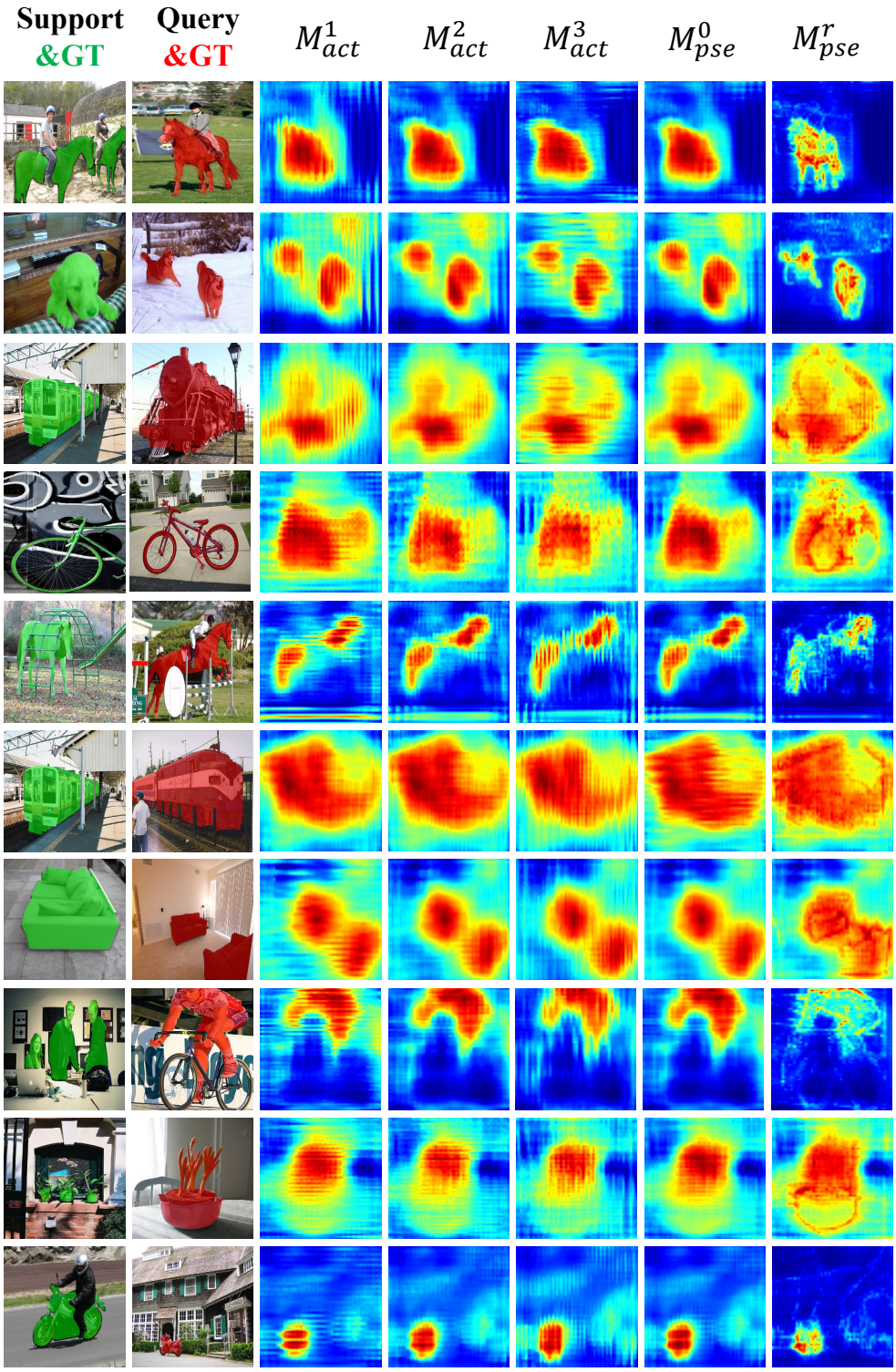

Figure A7. Visualization of the support activation maps $\{M^i_{act}\}^3_{i=1}$ and the initial pseudo mask M^0_{pse} in the support activation module (SAM), as well as the refined pseudo mask M^r_{pse} in the feature filtering module (FFM). Zoom in for details.