Supplementary Materials:
Hierarchical Dense Correlation Distillation for Few-Shot Segmentation

Bohao Peng[1], Zhuotao Tian[4*], Xiaoyang Wu[2], Chengyao Wang[1], Shu Liu[4], Jingyong Su[3], Jiaya Jia[1,4]

[1]The Chinese University of Hong Kong [2]The University of Hong Kong
[3]Harbin Institute of Technology, Shenzhen [4]SmartMore

1. Datasets

Compared with the previous work, we mainly use COCO-20^i [8] for verification in our experiments, which contains a more extensive data size. In this section, we give more research and analysis to prove that when the data is too clean and straightforward, few-shot segmentation will degenerate into foreground segmentation without caring about specific semantic information, and we conclude that COCO-20^i dataset will be a better choice to verify the model's generality.

1.1. Statistical Analysis

The performance gap between the COCO-20^i [8]and PASCAL-5^i [10] datasets was attributed to the class amount and data quantity. However, we found that the image's complexity is also a key factor. We first count the number of pictures with different contained category amounts, and the results are shown in Fig. 1. It can be found that most images of PASCAL-5^{i}have only single foreground. The model only needs to distinguish the significant foreground without semantic support, and few-shot segmentation will degenerate into foreground segmentation tasks.

1.2. Ablation Experiment of the Support Mask

We also ingeniously designed a simple ablation experiment to verify whether the model can extract the supervision information from the support annotations. First, we conduct the experiment following the original few-shot semantic segmentation settings [12], and then we give the model inputs without support supervision by removing the support mask. Tab. 1 and Fig. 4 showthe quantized and qualitative results, respectively. It is noteworthy that the model also can achieve stunning results on PASCAL-5^i dataset, even without the support mask. In Fig. 4, we visualize the performances under the simple and complex scenes. When the images are too straightforward, containing only single foreground with a clean background, few-shot segmentation will degenerate into foreground segmentation and achieve incredible performance even without caring about semantic information.

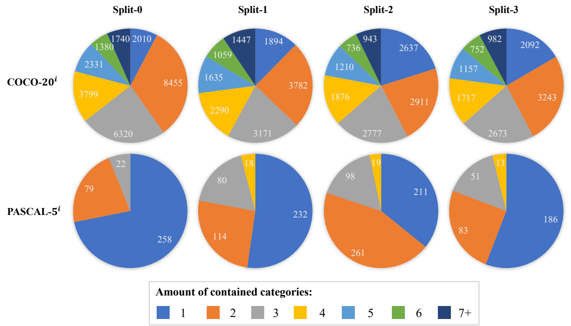

Figure 1. Statistics of pictures containing different category amounts for each fold of COCO-20^i [8] and PASCAL-5^i [10] datasets.

Dataset	Method	mIoU(%)				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean
PASCAL-5^i	w/ $\boldsymbol{M}^s$	71.2	75.4	67.6	63.6	69.5
	w/o $\boldsymbol{M}^s$	62.8	69.4	59.3	53.6	61.3
COCO-20^i	w/ $\boldsymbol{M}^s$	43.8	55.3	51.6	49.4	50.0
	w/o $\boldsymbol{M}^s$	31.0	37.7	31.2	33.8	33.4

Table 1. Ablation studies of the support mask's effects for PASCAL-5^i [10] and COCO-20^i [8] datasets under 1-hot setting.

2. Number of Test Episodes

Few-shot semantic segmentation adopts episode paradigm testing, where each test episode randomly selects the query and support pairs containing the same class objects. Typically, $1k$ episodes are set in the PASCAL-5^i evaluation since there are at most 584 images in each pascal fold. In contrast, we set $10k$ random episodes for each fold evaluation on COCO-20^i dataset and calculate the average mIoU. In this work, we give more experiments

[*]Corresponding Author

results proving that only $1k$ episodes are not sufficient to provide reliable results on COCO-20^i for comparison.

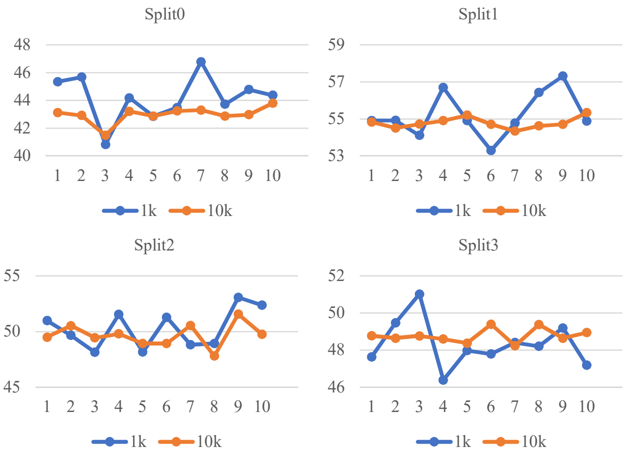

Figure 2. Comparison between $1k$ and $10k$ test samples set on COCO-20^i dataset for each fold.

Temperature	mIoU(%)				
(T)	Fold-0	Fold-1	Fold-2	Fold-3	Mean
0.5	42.1	54.9	48.6	48.6	48.6
1	**43.8**	**55.3**	**51.6**	49.4	**50.0**
2	43.1	55.1	50.0	**49.5**	49.4
5	42.4	54.9	48.2	47.5	48.3

Table 2. Ablation studies of the distillation temperature.

We iterate 10 times evaluations and then plot the line chart of each fold's result for both $1k$ and $10k$ episodes set as shown in Fig. 2. Insufficiency sampling episodes will lead to a precarious and significant fluctuation of test results. For example, the difference between the results is up to 6.0% mIoU in COCO-20^i Fold-0.

3. Implemetation Details

3.1. Decoder Structures

The decoder fuses the matching results $\{\boldsymbol{X}_l \in \mathbb{R}^{c_l \times h_l^q \times w_l^q}\}_{l=1}^{L}$ from coarse resolution to fine grain. The structure of decoder block and classification head is illustrated in Fig. 3. Decoder block inputs feature-match results from the same stage and sequential output from the last stage block following hierarchical paradigm. We use residual connection [4] to alleviate exploding/vanishing gradient problem. Classification head inputs the last stage output with maximum resolution and predicts the dense mask as the final output.

3.2. Experimental Environment

Software and hardware environment:

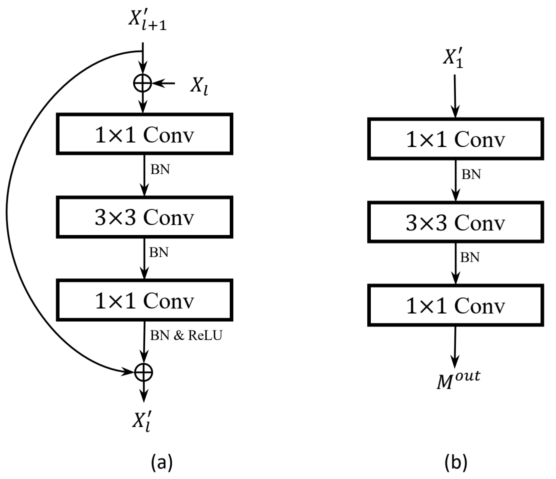

Figure 3. Structures of (a) decoder block and (b) classification head. Specifically, we adopt batch normalization (BN) on single GPU machine and convert it to sync batch normalization (SyncBN) for multi-GPU training.

- CUDA version: 11.7
- PyTorch version: 1.12.1
- GPU: NVIDIA GeForce RTX 3090
- CPU: Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz

3.3. Additional Qualitative Results

More qualitative results are provided to validate and analyze our proposed network effectiveness. Fig. 5 visualizes correlation maps and compared the dense predicted mask with or withou correlation distillation on both PASCAL-5^i [10] and COCO-20^i [8] under 1-shot setting. Fig. 6 shows the correlation maps from 1-3 pyramid stages. The correlation maps from the coarse resolution give rough locations of the classes related to the support annotations, and the fine layers provide more detailed features facilitating segmentation.

4. Ablation Experiments

4.1. Distillation Temperature

During the distillation process, we adopt the temperature T, a hyperparameter, to control the distribution. Given flattened correlation maps, we first apply a softmax layer with T to perform the spatial normalization among all positions:

$$\hat{\boldsymbol{C}}_l'(i) = \frac{\exp(\boldsymbol{C}_l'(i)/T)}{\sum_{j=1}^{h_l^q w_l^q} \exp(\boldsymbol{C}_l'(j)/T)}, \tag{1}$$

where l indicates the stage, T denotes the temperature of distillation [5]. In this section, we study the influence of the temperature set through the ablation experiments, and all results are shown in Tab. 2. When T equals 1, we get the best performance, and we keep this through all experiments as a default if not otherwise specified.

Shot	ReLU	mIoU(%)				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean
1	w/o	45.0	54.4	49.4	50.0	49.7
	w/	43.8	55.3	51.6	49.4	50.0
5	w/o	49.0	61.5	55.7	55.6	55.5
	w/	50.6	61.6	55.7	56	56.0

Table 3. Ablation studies of whether using ReLU activation function in the resnet block.

4.2. ReLU Layer from the Residual Block

Few-shot segmentation typically uses a parameter-fixed backbone, such as ResNet [4] and VGG [11], to extract the features of the input images. Setting ResNet as the backbone, some previous methods directly generate the feature from resnet block, and others drop the last ReLU layer since this activation function loses the negative part of the feature. We study the influences of the activation function in this subsection. The results shown in Tab. 3 prove that the ReLU layer brings slightly improved performance.

5. Use of Existing Assets

We gratefully thank the creators of the following assets that are very helpful for our project.

- PyTorch [9]: https://github.com/pytorch/pytorch
- PASCAL-VOC 2012 [2]: http://host.robots.ox.ac.uk/pascal/VOC/voc2012
- SBD [3]: http://home.bharathh.info/pubs/codes/SBD/download.html
- COCO 2014 [7]: https://cocodataset.org
- PSPNet [13]: https://github.com/hszhao/semseg
- BAM [6]: https://github.com/chunbolang/BAM
- PFENet [12]: https://github.com/dvlab-research/PFENet
- HSNet [4]: https://github.com/juhongm999/hsnet
- SegFormer [1]: https://github.com/open-mmlab/mmsegmentation

References

[1] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 3

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[3] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 3

[4] 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 (CVPR)*, June 2016. 2, 3

[5] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[6] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8057–8067, 2022. 3

[7] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[8] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 622–631, 2019. 1, 2, 5, 6

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 3

[10] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 1, 2, 4, 5, 6

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[12] 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 and machine intelligence*, 2020. 1, 3

[13] 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

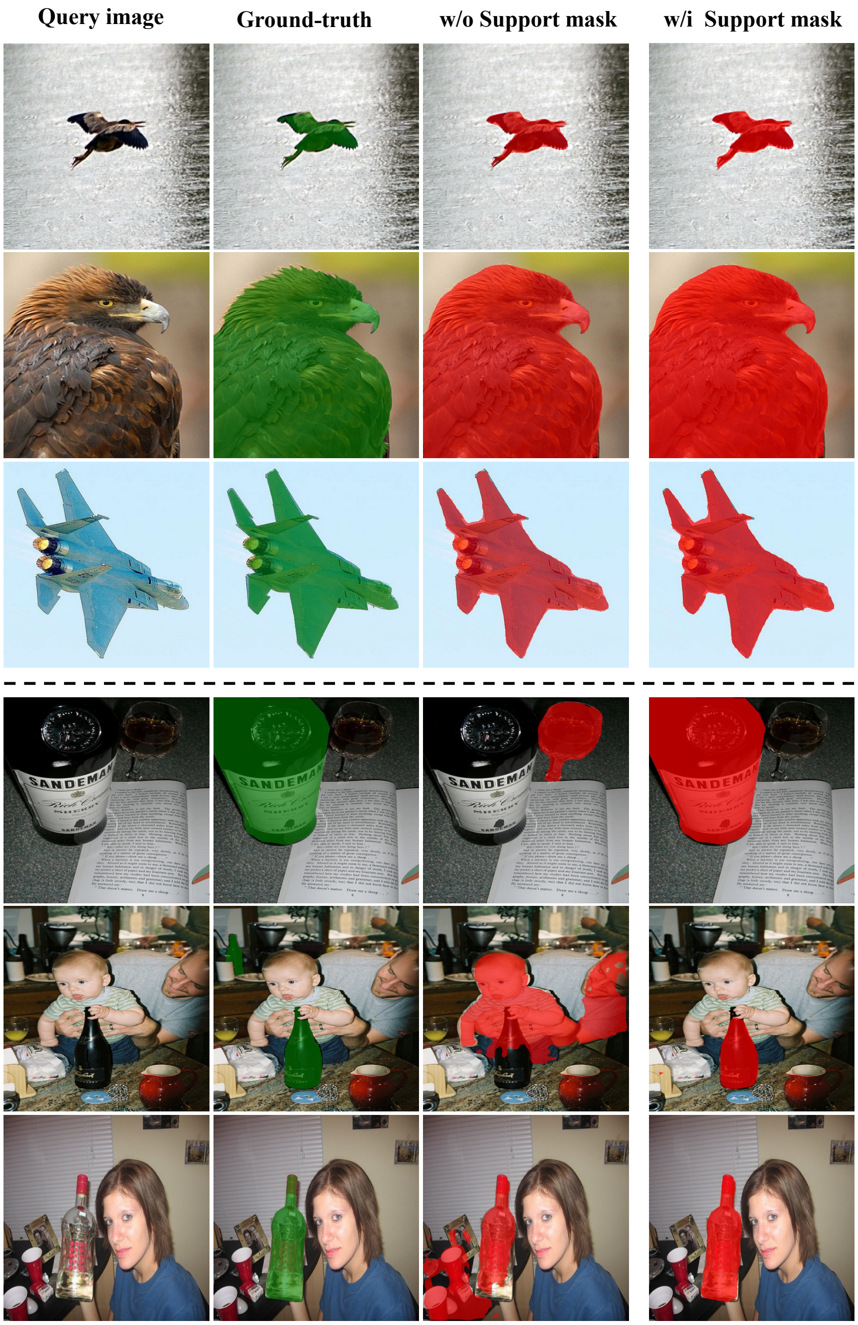

Figure 4. Qualitative results of the ablation studies for the support mask's effects under PASCAL-5^i [10] 1-hot setting. The upper panel shows the query images with only single foreground, while the query images from the below panel contain complex scenes.

.

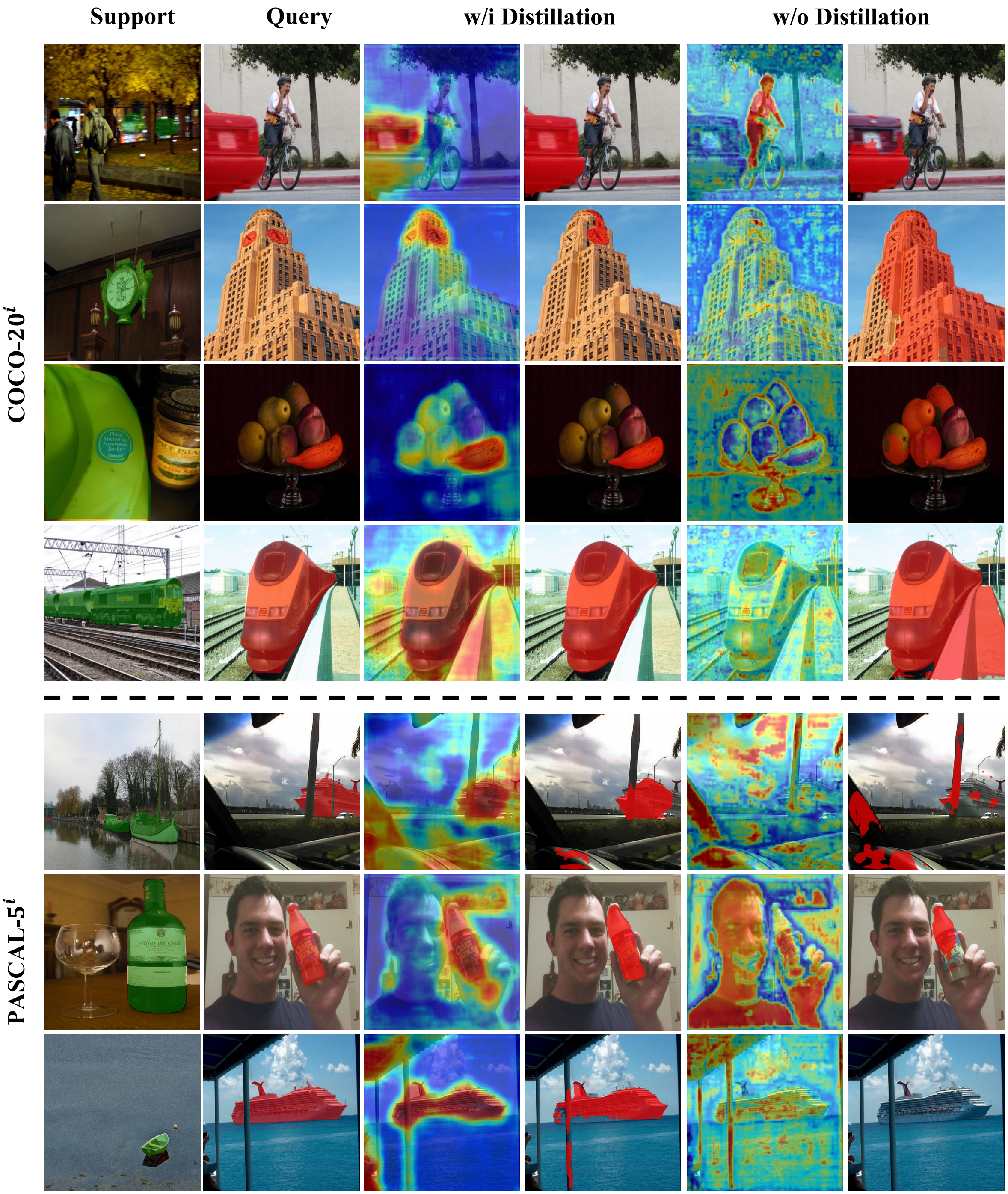

Figure 5. More visualization results of the correlation maps on both PASCAL-5^i [10] and COCO-20^i [8] under 1-shot setting. The first and second columns show examples of the support images with ground truth in green and the query images with labeled masks in red, respectively. Then we show the correlation maps and prediction results within or without distillation, respectively. We select the correlation map from the first stage for a brief introduction and visualize them by heatmaps.

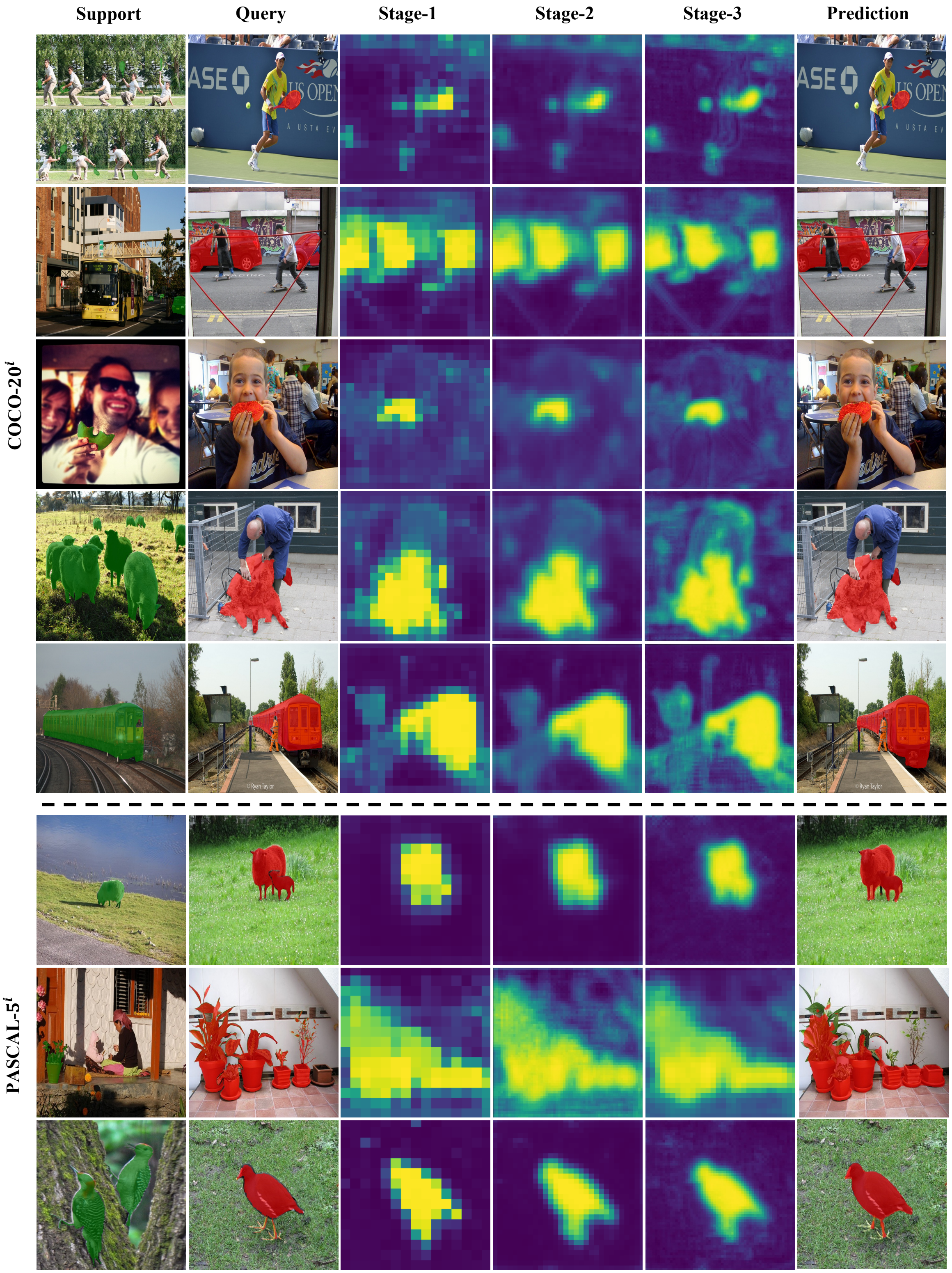

Figure 6. More qualitative results of the correlation pyramid on both PASCAL-5^i [10] and COCO-20^i [8] under 1-shot setting. The first and second columns show examples of the support images with ground truth in green and the query images with labeled masks in red, respectively. The following three columns visualize the correlation pyramid from the first to the third stage of the matching pyramid and the last column is the model's outputs.