Rethinking the Correlation in Few-Shot Segmentation: A Buoys View Supplementary Materials

Yuan Wang* Rui Sun* Tianzhu Zhang†
University of Science and Technology of China
{wy2016, issunrui}@mail.ustc.edu.cn, {tzzhang}@ustc.edu.cn

In the supplementary material, we first introduce more details about the support and query feature extraction of the three baseline methods. Then we elaborate on the detailed implementation of our experiments. Finally, we show more qualitative results of our proposed method.

1. More Details of Feature Extraction

We evaluate the effectiveness of the proposed ABCNet on three baselines: PFENet [5], DCAMA [4] and CyCTR [6]. In this section, we present the details of deploying ABCNet on three baselines.

For PFENet, the query and support images are fed into the shared backbones pretrained on ImageNet [3] to obtain image features. High-level features with the shape of $H \times W \times C$ that are from the 5^{th} block of ResNet (total 5 blocks including the stem block) are utilized to generate the prior mask. We flatten the support and query high-level features into $1D$ sequences with the shape of $HW \times C$ as inputs for our ABCNet, that is, F_S and F_Q in Figure 1.

For CyCTR, middle-level features are exploited in the cycle-consistent transformer module [6] to aggregate pixel-level support features into query pixels. We focus on improving the pixel-wise cross-attention module in the CyCTR by the way of suppressing false matches via ABCNet. In specific, the outputs of the 3^{rd} and 4^{th} blocks of ResNet are concatenated and then processed by a 1×1 convolutional layer to obtain the middle-level features. Besides, we also calculate the prior mask and use the masked average pooling on the support features to obtain a global object prototype. Then the expanded prototype, prior mask and middle-level query features are concatenated, transformed using a 1×1 convolutional layer, and flattened to obtain the F_Q in Figure 1. The prototype is also concatenated with the middle-level support features to generate the F_S in Figure 1.

For DCAMA, the cross-attention is conducted in a multi-scale and multi-layer manner. Specifically, collections of multi-scale multi-layer features $\{F_{i,l}^{Q}\}$ and $\{F_{i,l}^{S}\}$ are obtained from the pretrained backbones, where the $i \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ denotes the scale of the feature maps with respect to the original images and $l \in \{1, \ldots, L_i\}$ is the index of all layers of the i^{th} scale. The high-level features are flattened to obtain the F_Q and F_S in Figure 1. We only integrate ABCNet into the cross-attention of high-level features ($i = \frac{1}{32}$) for the following two considerations: (1) The construction of buoys relies on high-level features rich in semantic information. (2) The smaller spatial scale of high-level features is beneficial to improve computational efficiency.

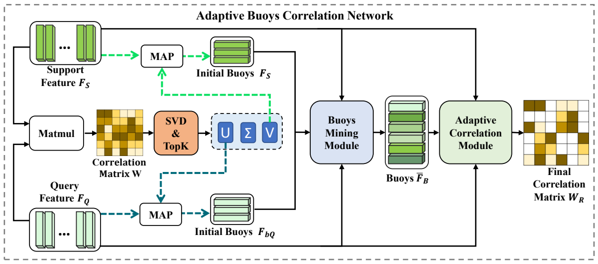

Figure 1. The framework of our proposed Adaptive Buoys Correlation Network (ABCNet). There are two main modules in ABCNet, i.e., the buoys mining module for establishing representative buoys (including the buoys initialization) and the adaptive correlation module for adaptive matching. The ultimate goal of our method is to suppress the false matches that occur in the original pixel-level correlation matrix

2. More Implementation Details

We follow the original training settings of three baseline methods to train the models with ABCNet. Specifically, when we implant the ABCNet into the PFENet, we use the cross entropy loss as the loss function, and the optimizer is adopted as SGD. The momentum and weight decay are set to 0.9 and 10^{-4} respectively. Besides, the ‘poly’ policy is utilized to decay the learning rate. The batch size is set to 4 for PASCAL-5^i and 8 for COCO-20^i with the learning rate 2.5×10^{-3} and 5×10^{-3}, respectively. When training CyCTR with ABCNet, we follow [6] to adopt the Dice loss [2] as the training objective. The transformer blocks

*Equal contribution
†Corresponding author

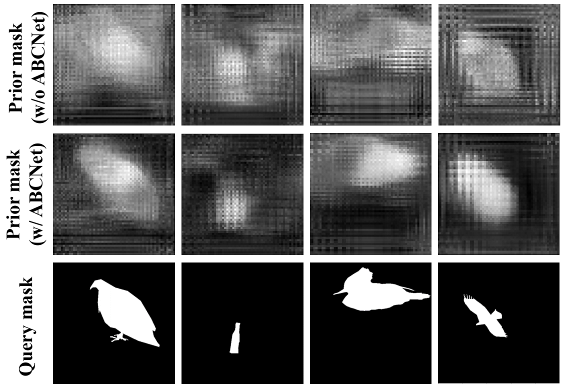

Figure 2. Visualization of prior masks with and without ABCNet. With the assistance of ABCNet, the prior masks can indicate target regions better.

are optimized with AdamW [1], the learning rate is set to 1×10^{-4} and the weight decay is 1×10^{-2}. The rest settings are exactly the same as that in PFENet as described above. When we combine the ABCNet with DCAMA, we employ the mean binary cross-entropy loss to train the model: $\boldsymbol{L}_{BCE} = -\frac{1}{N}\sum[y\log(p) + (1-y)\log(1-p)]$, where the p is the prediction and the N denotes the total number of pixels. We adopt the SGD optimizer, and the learning rate, momentum and weight decay are set to 1×10^{-3}, 0.9, and $1 \times 10{-}4$, respectively. The batch size is set to 48 for both PASCAL-5^i and COCO-20^i by following [4].

3. More Visualization Results

In Figure 3, We show more qualitative results on Pascal-5^i. We can observe that the results from models with ABCNet can segment target objects more accurately. The models without ABCNet usually cannot fully segment targets and are more likely to incorrectly recognize background as foreground objects. We deem that this is ascribed to the extensive false matches existing in the pair-wise pixel-level correlation. In Figure 2, we can observe that the prior masks from the models with ABCNet can localize target objects well.

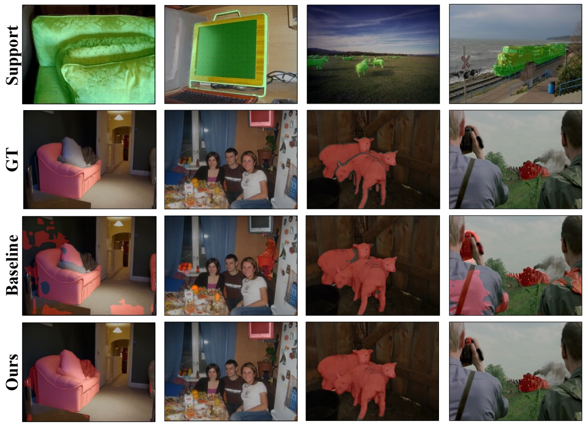

Figure 3. Qualitative comparison with the baseline. Results with ABCNet can achieve more accurate segmentation.

References

[1] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[2] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 1

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[4] Xinyu Shi, Dong Wei, Yu Zhang, Donghuan Lu, Munan Ning, Jiashun Chen, Kai Ma, and Yefeng Zheng. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *European Conference on Computer Vision*, pages 151–168. Springer, 2022. 1, 2

[5] 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

[6] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021. 1