Supplementary Materials: Balancing Conservatism and Aggressiveness: Prototype-Affinity Hybrid Network for Few-Shot Segmentation

Tianyu Zou[1] Shengwu Xiong[2*] Ruilin Yao[1,4,5] Yi Rong[3,1*]
[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology
[2] Interdisciplinary Artificial Intelligence Research Institute, Wuhan College
[3] Sanya Science and Education Innovation Park, Wuhan University of Technology
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences
[5] Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences
{zoutianyu, xiongsw, yaoruilin, rongyi}@whut.edu.cn

A. Training Loss

The overall loss function consists of a main loss and an auxiliary loss. The main loss follows the original baseline's objective, while the auxiliary loss leverages the binary cross-entropy (BCE) between the predicted affinity masks and the ground truth. Taking SCCAN [5] as an example, its training loss is defined as:

$$\mathcal{L}_{\text{SCCAN}} = \mathcal{L}_{\text{main}} + \lambda \cdot \mathcal{L}_{\text{aux}}, \tag{1}$$

$$\mathcal{L}_{\text{main}} = \text{Dice}(\hat{M}_q, M_q), \tag{2}$$

$$\mathcal{L}_{\text{aux}} = \left(\frac{1}{N} \sum_{l=1}^{N} \text{BCE}(M^{\text{aff}}_{q,l}, M_q) \right). \tag{3}$$

Here, Dice$(\cdot)$ denotes the Dice loss, $\hat{M}q$ is the final predicted mask, and M_q is the ground truth. The coefficient λ is a balancing hyperparameter, empirically set to 1 following AENet's configuration [6]. The $M^{\text{aff}}_{q,l}$ represents the l-th affinity mask predicted by the l-th Prototype-Guided Feature Enhancement (PFE) module, and N is the total number of such modules (e.g., SCCAN uses 8 PFE modules).

In contrast, HDMNet adopts a different loss formulation that includes a main loss, an auxiliary loss, and a stage-wise distillation loss. The main loss is computed using cross-entropy between the predicted mask and the ground-truth annotations. The auxiliary loss is similar to SCCAN as Equation 3. Additionally, HDMNet introduces a hierarchical knowledge distillation mechanism, which progressively aligns the intermediate correlation maps across stages. Specifically, the softmax-normalized output of a higher-level correlation map serves as the teacher, while the corresponding lower-level output acts as the student. Formally, the overall training loss for HDMNet is defined as:

$$\mathcal{L}_{\text{HDMNet}} = \mathcal{L}_{\text{main}} + \lambda \cdot \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{KL}}, \tag{4}$$

$$\mathcal{L}_{\text{main}} = \text{CE}(\hat{M}_q, M_q), \tag{5}$$

$$\mathcal{L}_{\text{KL}} = \sum_{x \in \mathcal{X}} \phi_t(x) \log \left(\frac{\phi_t(x)}{\phi_s(x)} \right), \tag{6}$$

where $\phi_t(x)$ and $\phi_s(x)$ denote the teacher and student outputs at spatial position x, respectively. To compute the divergence, the teacher map is resized to match the resolution of the student. For the final stage, which lacks a subsequent teacher, the ground truth is used for direct supervision.

B. Experimental Evidence on Conservatism and Aggressiveness

To further support our claim regarding the conservative of prototype learning methods and the aggressive of affinity learning methods, we conduct more quantitative analyses on false positive (FP) and false negative (FN) rates across representative models. Table 1 presents the FP and FN values of seven methods on the PASCAL-5^i dataset under the 1-shot setting. Among them, SCCAN [5], HDMNet [4], and Aff represent affinity learning approaches, while SSP [2], HPA [1], RARE [3], and Pro correspond to prototype learning methods. Note that Pro and Aff are the simplest models in their respective categories. Pro is built by removing the self-supporting module from SSP, and Aff is built from SCCAN by removing its specific designs.

As shown in Table 1, we observe that all prototype learning methods exhibit higher FN values but lower FP values compared to affinity learning methods. This shows similar phenomena as Figure 2 in our paper, where prototype learning methods tend to under-segment (i.e., miss foreground regions) and affinity learning methods tend to over-segment by activating background regions. These results

*Corresponding author

Metrics	SCCAN	HDMNet	Aff	SSP	HPA	RARE	Pro
FP	0.097	0.098	0.118	0.077	0.075	0.065	0.084
FN	0.036	0.039	0.035	0.056	0.048	0.058	0.053

Table 1. False Positive (FP) and False Negative (FN) rates of affinity learning and prototype learning methods on PASCAL-5^i.

Methods	5^0	5^1	5^2	5^3	Mean
HDMNet [4] (Baseline)	71.0	75.4	68.9	62.1	69.4
PAHNet (HDMNet+HPA)	72.0	75.7	70.0	66.6	71.1
PAHNet (HDMNet+RARE)	72.2	76.7	71.0	66.7	71.6

Table 2. Comparison of various prototype predictors integrated into the PAHNet on PASCAL-5^i.

provide empirical justification for the conservatism of prototype learning methods and the aggressiveness of affinity learning methods. Furthermore, the comparison between SSP and Pro demonstrates that *the conservatism is inherent to prototype learning methods, rather than being induced by specific design components like the self-supporting module.*

To further validate the compatibility and generalizability of our proposed hybrid framework, we integrate different prototype learning methods into HDMNet to form our PAHNet. As shown in Table 2, PAHNet consistently outperforms the HDMNet baseline across all four splits of PASCAL-5^i, regardless of the prototype predictor used. These results indicate that our proposed PAHNet architecture not only achieves a balance between conservatism and aggressiveness, but also serves as a flexible framework that can benefit from various prototype learning methods.

C. Additional Ablation Study

We conduct an additional series of ablation studies to investigate the impact of different components in our method on segmentation performance. Note that the experiments in this section are conducted with the combination of SCCAN and BAM as the baseline on the PASCAL-5^i dataset under the 1-shot setting unless specified otherwise.

Effect of M_q^{aff} and M_q^{pro} in PFE Module. We investigate the individual and combined effects of M_q^{aff} and M_q^{pro} in our PFE module. As shown in Table 3, the baseline mIoU starts at 68.4%. Using only M_q^{aff} results in a slight improvement, with a slight performance drop on Split 2 (-0.5%), probably due to its sensitivity to false positives. In contrast, using only M_q^{pro} results in a consistent mIoU improvement, with an average gain of +1.3% across all splits, benefiting from its conservative that focuses on reliable foreground predictions. When both M_q^{aff} and M_q^{pro} are used together, the mean mIoU further increases by +1.9%, demonstrating the complementary effect of the two predictions in enhancing the final segmentation performance.

Effect of W_{score} and M_{score} in ASC module. We study

M_q^{aff}	M_q^{pro}	5^0	5^1	5^2	5^3	Mean
		69.6	73.2	69.3	61.6	68.4
✓		70.2	73.3	68.8	61.6	68.5
	✓	71.9	74.1	70.1	62.7	69.7
✓	✓	72.3	73.7	71.8	63.3	70.3

Table 3. Ablation study on the effect of M_q^{aff} and M_q^{pro}.

W_{score}	M_{score}	5^0	5^1	5^2	5^3	Mean
		69.6	73.2	69.3	61.6	68.4
✓		72.4	73.7	72.3	62.1	70.1
	✓	72.8	73.3	70.1	63.2	69.9
✓	✓	73.1	74.1	73.1	63.6	71.0

Table 4. Ablation study on the effect of W_{score} and M_{score}.

Method	#Parameters	#FLOPs
HDMNet [4]	$4.2M$	$333.2G$
HDMNet + PAHNet	$4.5M$	$333.8G$

Table 5. Impacts of PAHNet on parameters and FLOPs.

the effect of the re-weight matrix W_{score} and the masking matrix M_{score} in our ASC module. As shown in Table 4, using only W_{score} increases the mIoU to 70.1%, while using only M_{score} improves it to 69.9%. When both W_{score} and M_{score} are applied together, the mIoU further increases to 71.0% (+2.6%), demonstrating the effectiveness of both components. These results confirm that W_{score} and M_{score} provide complementary benefits in enhancing query feature calibration and contribute to overall performance gains.

D. Parameter Amount and FLOPs

Our PAHNet framework is designed as a plug-and-play that can be easily integrated into existing affinity learning methods. For example, it can be incorporated into HDMNet by inserting three PFE and ASC modules. As shown in Table 5, PAHNet introduces minimal overhead in both parameter count and computational cost. Specifically, the number of parameters grows from $4.2M$ to $4.5M$, and the FLOPs increase slightly from $333.2G$ to $333.8G$, demonstrating the efficiency and lightweight design of our approach.

E. Additional Qualitative Results

Qualitative Analysis of Component Modules. As visualized in Fig. 2, the two modules exhibit distinct yet complementary behaviors. The PFE enforces foreground consistency by enhancing the foreground features of support and query, which is reflected in the expanded foreground activation coverage seen in the Class Activation Maps (CAMs). In contrast, the ASC focuses on suppressing background

interference through spatial reweighting of attention scores, resulting in cleaner activation boundaries and fewer background artifacts in the CAM visualizations. While both modules improve baseline performance, the results driven by the second module show superior segmentation precision because of its explicit FG-BG decoupling, especially in complex scenes. These observations align well with the design intent: the PFE enhances the foreground features, while the ASC suppresses the FG-BG feature mismatch.

Qualitative Analysis of M_q^{aff} and M_q^{pro}. Further qualitative results are provided in Fig. 3 to illustrate the effectiveness of the M_q^{aff} and M_q^{pro}. The figure presents CAMs and segmentation perditions under three different settings: using only affinity predictions (M_q^{aff}), only prototype predictions (M_q^{pro}), and their combined fusion. When relying solely on affinity predictions, over-activation in background regions occurs (notably in rows 5 and 6), leading to more false positives in segmentation. By contrast, using only prototype predictions results in less activation of foreground regions (rows 3 and 4), causing incomplete object coverage. Importantly, the combined use of both predictions balances these issues, the prototype predictions constrain affinity attention to semantically coherent foreground areas (rows 7 and 8), achieving both precise localization and full object coverage. This empirical evidence confirms that the effect between conservative prototype guidance affinity learning effectively reduces both false positive and false negative errors, achieving a better balance in few-shot segmentation

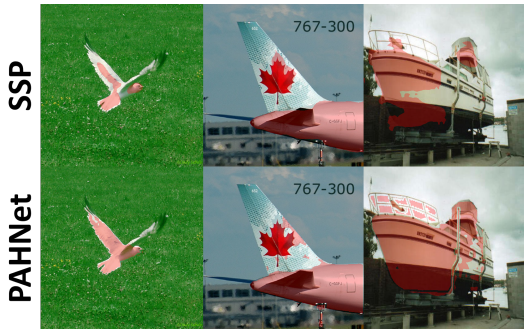

Figure 1. Representative failure cases of PAHNet.

F. Limitation and Future Direction

Despite the effectiveness of our proposed PAHNet framework, there remain several limitations that merit discussion. We illustrate several representative failure cases in Figure 1, where the segmentation results are notably limited by the underestimation of the foreground scope during the prototype prediction stage. The quality of the prototype mask plays a critical role in guiding the subsequent affinity learning module. In complex scenes with large intra-class variations or occlusions, the prototype predictor (e.g., SSP) may generate overly conservative foreground masks. These narrow foreground regions impose overly strong spatial constraints on affinity learning, which in turn hinders the model's ability to recover the complete object area. As a result, the affinity module may fail to propagate semantic cues to the entire foreground, leading to incomplete segmentation. Addressing these limitations in future work may involve enhancing the prototype module's robustness to complex scenes, incorporating uncertainty modeling, or enabling iterative refinement between the prototype and affinity modules.

References

[1] Gong Cheng, Chunbo Lang, and Junwei Han. Holistic prototype activation for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4650–4666, 2022. 1

[2] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 701–719. Springer, 2022. 1

[3] Chunbo Lang, Gong Cheng, Binfei Tu, Chao Li, and Junwei Han. Retain and recover: Delving into information loss for few-shot segmentation. *IEEE Transactions on Image Processing*, 32:5353–5365, 2023. 1

[4] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 1, 2

[5] Qianxiong Xu, Wenting Zhao, Guosheng Lin, and Cheng Long. Self-calibrated cross attention network for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 655–665, 2023. 1

[6] Qianxiong Xu, Guosheng Lin, Chen Change Loy, Cheng Long, Ziyue Li, and Rui Zhao. Eliminating feature ambiguity for few-shot segmentation. In *European Conference on Computer Vision*, pages 416–433. Springer, 2025. 1

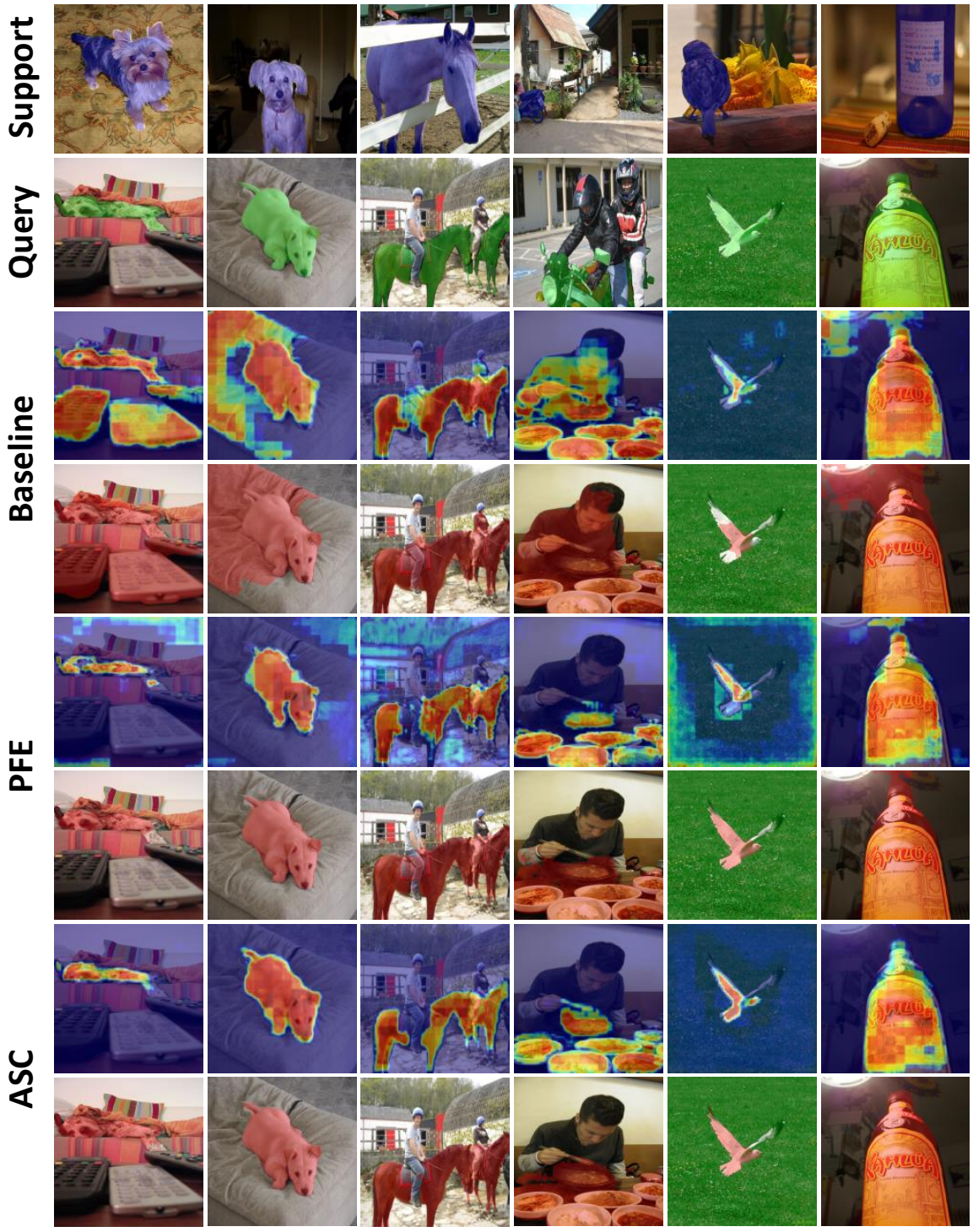

Figure 2. Additional visualization results of CAMs and segmentation perditions on PASCAL-5^i under 1-shot setting. The first and second rows show examples of the support images with ground truth in blue and the query images with labeled masks in green, respectively. The following rows present CAMs from the last stage along with the corresponding prediction results for the baseline model, the baseline enhanced with Prototype-Guided Feature Enhancement (PFE), and the baseline augmented with Attention Score Calibration (ASC). CAMs are visualized as heatmaps to highlight the activation differences across settings.

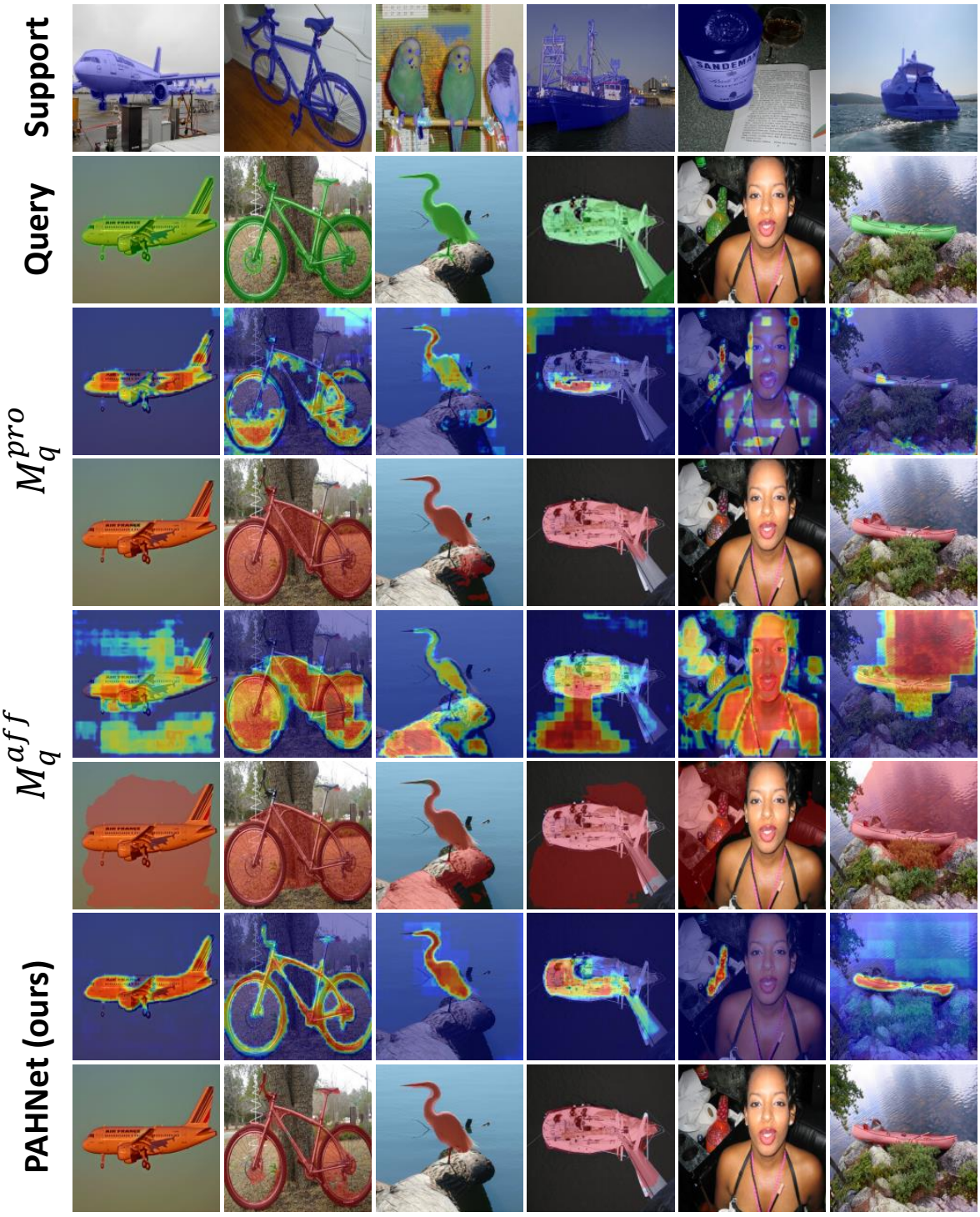

Figure 3. Additional visualization results of CAMs and segmentation predictions on PASCAL-5^i under the 1-shot setting. The first and second rows show support images with ground truth masks in blue and query images with labeled masks in green, respectively. The following rows present CAMs from the last stage and corresponding prediction results under three configurations: using only M_q^{aff}, only M_q^{pro}, and the combination of both. CAMs are visualized as heatmaps to highlight the activation differences across settings.