Unify the Views: View-Consistent Prototype Learning for Few-Shot Segmentation

Supplementary Material

1. Detailed Formulas

Prototype Consistency Loss (Eq. (6)). We provide the detailed formulation of the prototype consistency loss briefly outlined in the manuscript. To encourage structural consistency under viewpoint changes, we minimize the distance between class-level descriptors $\mathbf{p}_S^v$ and $\mathbf{p}_Q^v$ derived from support and query features. The loss is defined as the mean squared error:

$$\mathcal{L}_{\text{proto}} = \text{MSE}(\mathbf{p}_Q^v, \mathbf{p}_S^v) = \frac{1}{C}\sum_{i=1}^{C}\left(\mathbf{p}_Q^v[i] - \mathbf{p}_S^v[i]\right)^2, \quad (1)$$

where C is the feature dimension.

Mask Prediction Loss (Eq. (16)). To supervise the predicted query mask $\hat{M}_Q \in [0,1]^{H\times W}$ against the binary ground truth $M_Q \in \{0,1\}^{H\times W}$, we adopt a compound loss combining pixel-level and region-level objectives.

Binary Cross-Entropy (BCE) focuses on pixel-wise classification:

$$\mathcal{L}_{\text{BCE}}(\hat{M}_Q, M_Q) = -\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\Big[M_Q^{ij}\log\hat{M}_Q^{ij} + (1-M_Q^{ij})\log(1-\hat{M}_Q^{ij})\Big], \quad (2)$$

which emphasizes local accuracy but may suffer under foreground-background imbalance.

Dice Loss (DL) enhances shape and region consistency:

$$\mathcal{L}_{\text{DL}}(\hat{M}_Q, M_Q) = 1 - \frac{2\sum_{i,j}\hat{M}_Q^{ij}M_Q^{ij}}{\sum_{i,j}\hat{M}_Q^{ij} + \sum_{i,j}M_Q^{ij}}, \quad (3)$$

by measuring global overlap and suppressing fragmented predictions. The final mask supervision loss in Eq. (17) is computed as a weighted sum of these two components.

2. Computational Efficiency Analysis

To evaluate the computational efficiency of the proposed method, we compare its parameter count, FLOPs, and inference throughput against the FCP baseline under identical experimental conditions (RTX 3090, batch size = 1, no AMP). As shown in Table 1, both models build upon frozen backbones and image encoders; the reported metrics correspond solely to the overhead introduced by their respective task-specific modules.

Table 1. Computational efficiency comparison. Both models use frozen SAM and ResNet encoders; reported values reflect additional modules introduced atop the backbone.

Model	#Params (M)	FLOPs (G)	Throughput (fps)
FCP [1]	26.0	257.9	21.59
Ours	27.6	283.8	16.53

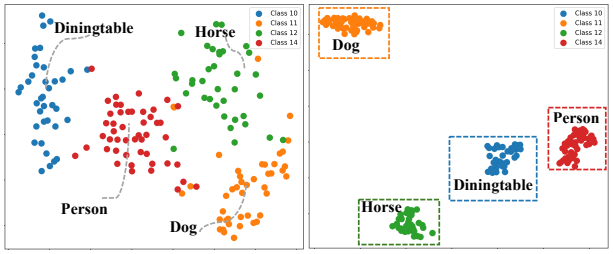

Figure 1. t-SNE visualization of SVGA-enhanced query features on PASCAL-5^i fold 2.

Compared to FCP (26.0M parameters, 257.9 GFLOPs), our VINE introduces a modest increase in model size and computation, reaching 27.6M parameters (+4.1%) and 283.8 GFLOPs (+10.0%). This computational increase leads to a corresponding reduction in inference throughput, from 21.59 fps to 16.53 fps. It is worth noting that the newly introduced SVGA module contributes less than 2% of the total FLOPs. The observed accuracy improvements, as reported in the main text, justify this slight increase in computational cost, demonstrating a favorable trade-off between efficiency and performance enabled by our lightweight architectural additions.

3. Additional Qualitative Results

t-SNE Visualization To complement the t-SNE analysis in the main text, Figure 1 presents the visualization on PASCAL-5^i fold 2. As with the main results on folds 0, 1, and 3, the raw ResNet features exhibit scattered intra-class clusters and inter-class overlap. In contrast, the SVGA-enhanced features form compact, well-separated clusters, indicating stronger class discrimination and improved alignment across geometric variations. This supports the generalizability of SVGA in promoting cross-view consistency across all data folds.

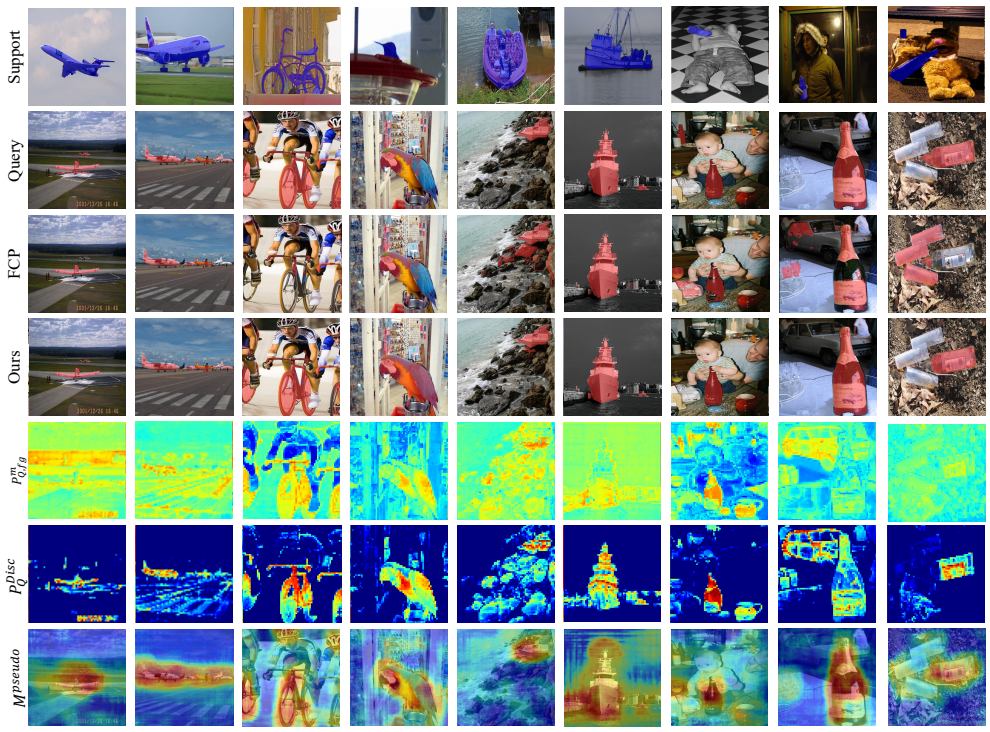

Figure 2. Additional qualitative results on PASCAL-5^i. Our method consistently produces sharper and more complete masks than FCP, even under severe viewpoint variation, occlusion, or incomplete support information.

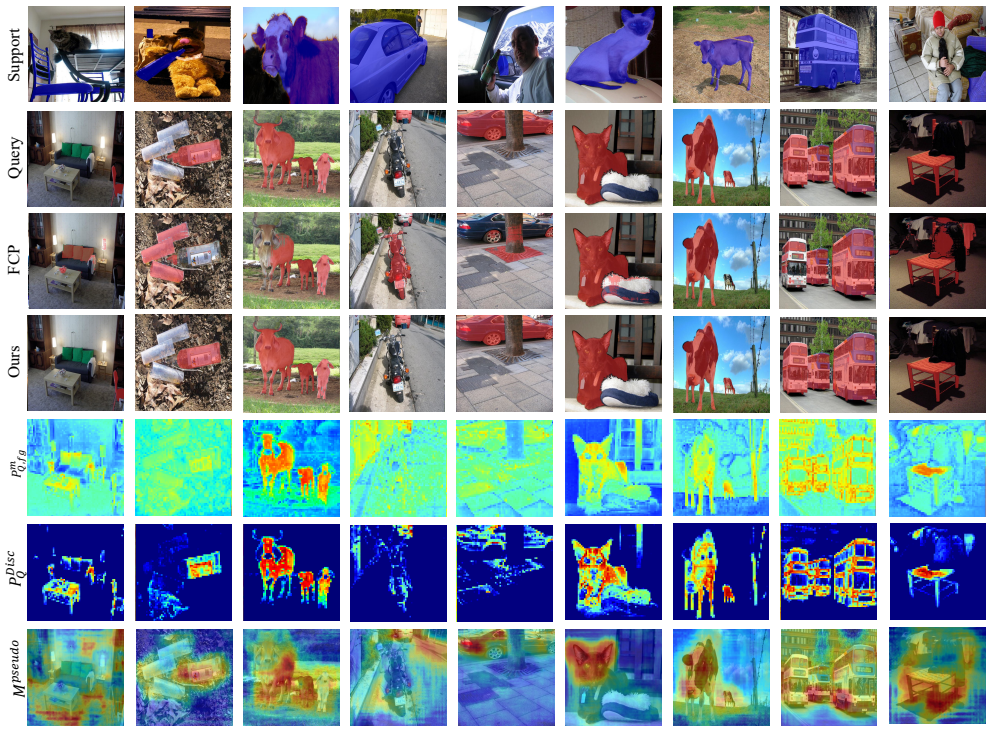

Figure 3. Additional qualitative results on PASCAL-5^i.

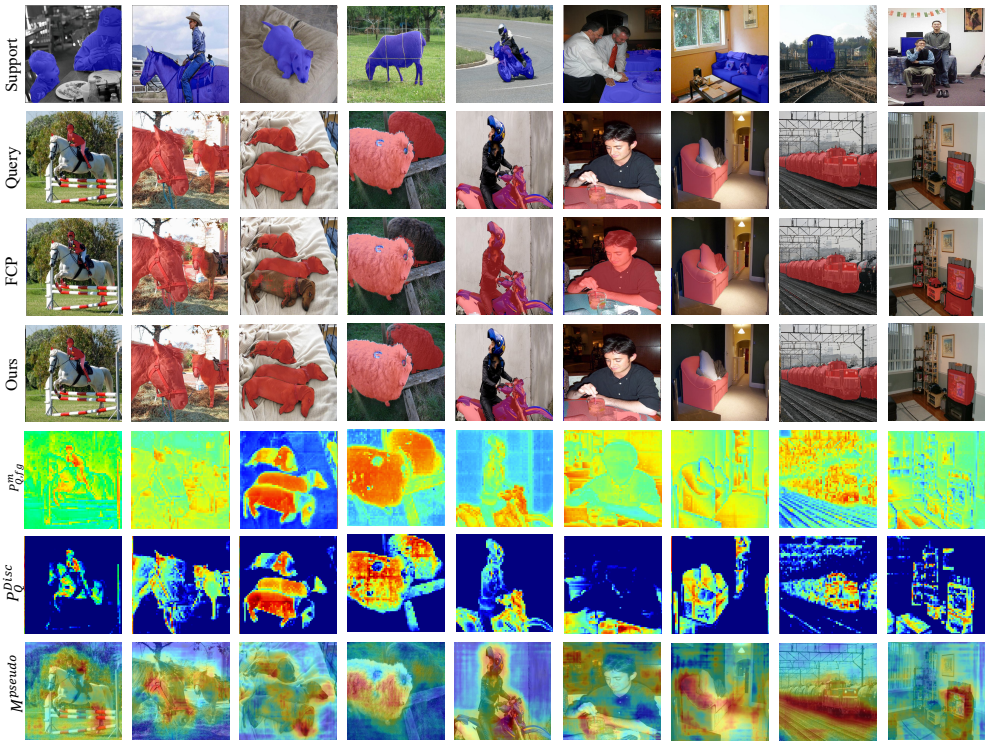

Figure 4. Additional qualitative results on PASCAL-5^i.

More Segmentation Visualizations. Figures 2–4 provide extensive qualitative comparisons across all 20 classes of PASCAL-5^i. These examples showcase challenging scenarios such as severe viewpoint variations, occlusions, and ambiguous support information. Compared to FCP, which often produces fragmented masks or misaligns with the object geometry, our approach consistently generates sharper and more complete segmentations. The integration of the spatial-view graph alignment and the discriminative prior effectively suppresses background interference and reinforces structural coherence, underscoring the robustness of our framework in diverse and demanding few-shot settings.

References

[1] Suho Park, SuBeen Lee, Hyun Seok Seong, Jaejoon Yoo, and Jae-Pil Heo. Foreground-covering prototype generation and matching for sam-aided few-shot segmentation. In *AAAI*, pages 6425–6433, 2025. 1