Supplementary Materials

MSI shows that background information overlooked by existing methods can help to segment a target. By utilizing both STF and SIF, MSI is able to obtain information on the target class from the entire support image while minimizing bias to the target class. Through STF, MSI is able to obtain more fine-grained target information which potentially gets removed by masking background features. By using SIF, it becomes possible to acquire more information about the target that might have been missed by the limited support mask. Furthermore, it utilizes other objects in the background of the support image to either avoid segmenting non-target objects or aid segmenting the target object in the query image. In this supplementary material, we include the following items:

A.1 Correlation map analysis.
A.2 Detailed architecture.
A.3 Training profiles.
A.4 Data augmentation.
A.5 Differences between VAT & VAT + MSI.
A.6 Additional qualitative results.
A.7 Failure cases.

A.1 Correlation Map Analysis

Since FM (Feature Masking) eliminates background features, high correlation values exist only within the target object area (Fig. 11). When the support mask does not accurately cover the entire target object, the target information received by a network is limited (Fig. 14).

The support target features (STF) have high activation around the target object in the support image but not in the query image. Some background areas in the support image features (SIF) are activated because they match the background features of the query image. Although both SIF and STF weakly capture the target object of the query image in their correlation maps, once these correlation maps are fed to the encoder, in the encoded feature maps, we observe strong signals across the target class of the query image (see the feature map in Fig. 11). On the contrary, the FM encoder produces a more scattered activation around the target.

A.2 Detailed Architecture

In Fig. 3 of the main manuscript, the encoder part is shown briefly. In Fig. 12, we reveal more details relevant to the encoder that are specific to each baseline architecture: HSNet [22], VAT [9] and, ASNet [12]. For HSNet, the SCM is divided into 3 convolutional blocks and used in 3 stages of the encoder step-by-step. VAT utilizes query features again in the decoding stage. ASNet uses pooling to reduce the input feature size.

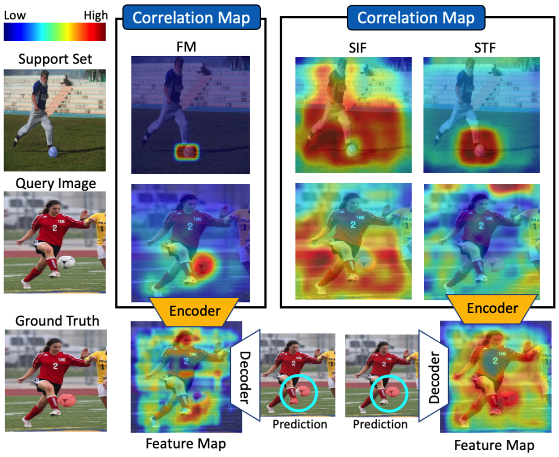

Figure 11. Correlation map and feature map visualization. Although STF and SIF weakly capture the target class of the query image in their correlation maps, once these correlation maps are fed to the encoder, we notice strong signals on the target object.

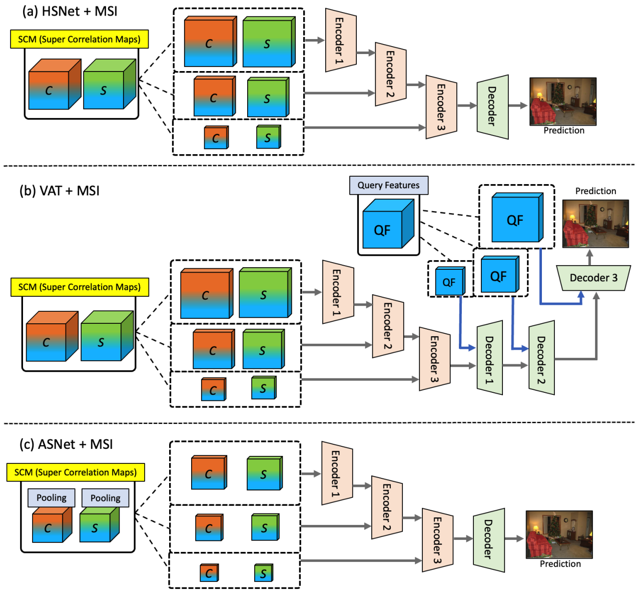

Figure 12. Detailed architectures showing where SCM is used for HSNet [22], VAT [9], and ASNet [12]. (a) HSNet + MSI (b) VAT + MSI (c) ASNet + MSI.

A.3 Training Profiles

Figure 10 (main paper) shows that VAT + MSI provides 3.5x faster convergence compared to VAT [9]. In Fig. 13, we reveal that our MSI with HSNet baseline, HSNet + MSI, enables 4.5x faster convergence in comparison to HSNet [22]. This could be because STF provides a strong prior on the target boundary information to the encoder network.

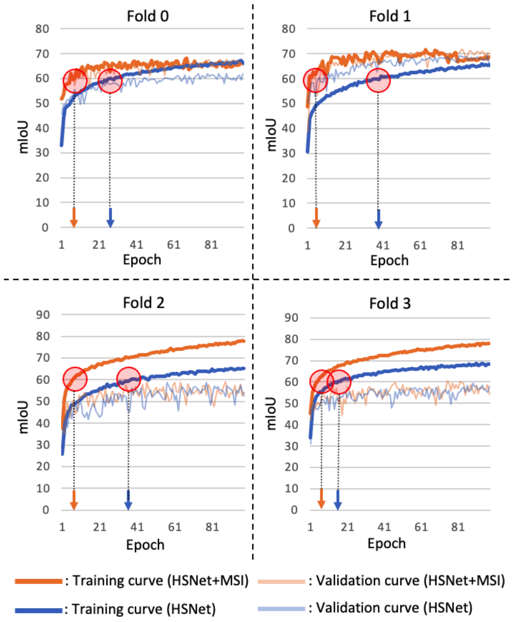

Figure 13. Train. and Val. profiles of HSNet [22] and HSNet+MSI on PASCAL-5^i [6] with ResNet50 [8]. HSNet+MSI provides 4.5x faster convergence (to reach 60% in mIoU) on average than HSNet. Red circles indicate when train accuracy reaches 60% in mIoU.

Method & Backbone	Data Aug.	MSI	mIoU	FB-IoU
VAT [9] (ResNet50 [8])	-	-	65.5	77.8
	✓	-	65.3	77.4
	-	✓	67.2	78.6
	✓	✓	**68.3**	**79.1**
VAT [9] (ResNet101 [8])	-	-	67.9	79.6
	✓	-	67.5	78.8
	-	✓	68.9	79.4
	✓	✓	**70.1**	**82.3**

Table 9. The impact of data augmentation on VAT [9] and VAT + MSI on PASCAL-5^i [6]. Best results are shown in **bold**.

A.4 Data Augmentation.

VAT [10] does not recommend using data augmentation [2, 4] as it causes performance drop. However, when using our proposed method with VAT, the performance improves using CATs data augmentation [4] (see Tab. 9).

A.5 Differences between VAT & VAT + MSI.

VAT performs correlation mapping only between query features (QF) and masked support features. We term the masked support features as feature masking (FM) (Fig. 1). FM loses fine-grained target information such as textures and boundaries because of late-stage masking and masking features with inaccurate support masks (Fig. 8, Fig. 9), which limits VAT. VAT+MSI utilizes both SIF and STF (Fig. 2) to compute correlation maps CS_1 and CS_2 respectively with QF. CS_1 allows the network to learn additional contextual target information in SIF (e.g., in the background), whereas CS_2 uses STF to mask the input and has stronger target information than FM, which helps the network to activate target foreground features while simultaneously deactivating target-irrelevant background features in SIF. This way, by maximizing target information, MSI can boost performance.

A.6 Additional Qualitative Results.

Figs. 15, 16, and 17 show additional qualitative results of VAT + MSI on three benchmarks.

A.7 Failure Cases

Although MSI shows significant gains over strong FSS baselines, we identify some challenging FSS cases where there is still room for improvement. When a very small part of the target object in the support set is given, the target object might not be accurately segmented in the query image (see the fourth row in Fig. 19). Also, when the target object in the query image lies far in the background and appears blurry (see the fifth row in Fig. 18), the segmentation accuracy could worsen. Other challenging cases are when the color or shape of the target object is quite different (see the third row in Fig. 18 and the third row in Fig. 19) and when the object class in the support set and query image are not the same, but are similar in shape and semantics, such as a computer monitor vs. television (see the second row in Fig. 18).

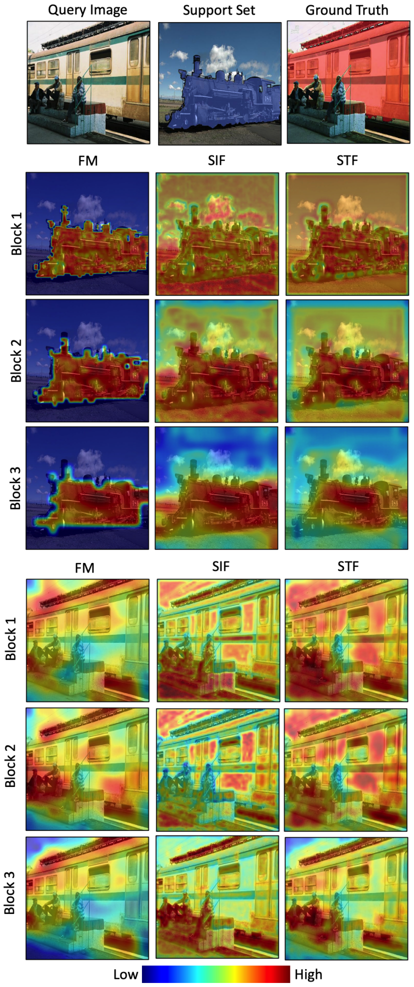

Figure 14. Correlation map visualization. Feature masking (FM) with the inaccurate support mask causes inadvertent loss in the boundary information of the target. On the other hand, Support Image Features (SIF) utilizes the entire image features and the Support Target Features (STF) focuses on the target after removing the background at the input level. Note that in the Block 1, compared to FM, STF has distinct signals at the target class boundary. In the Block 3, compared to FM, in SIF, the correlation is evenly distributed across various areas in the target object.

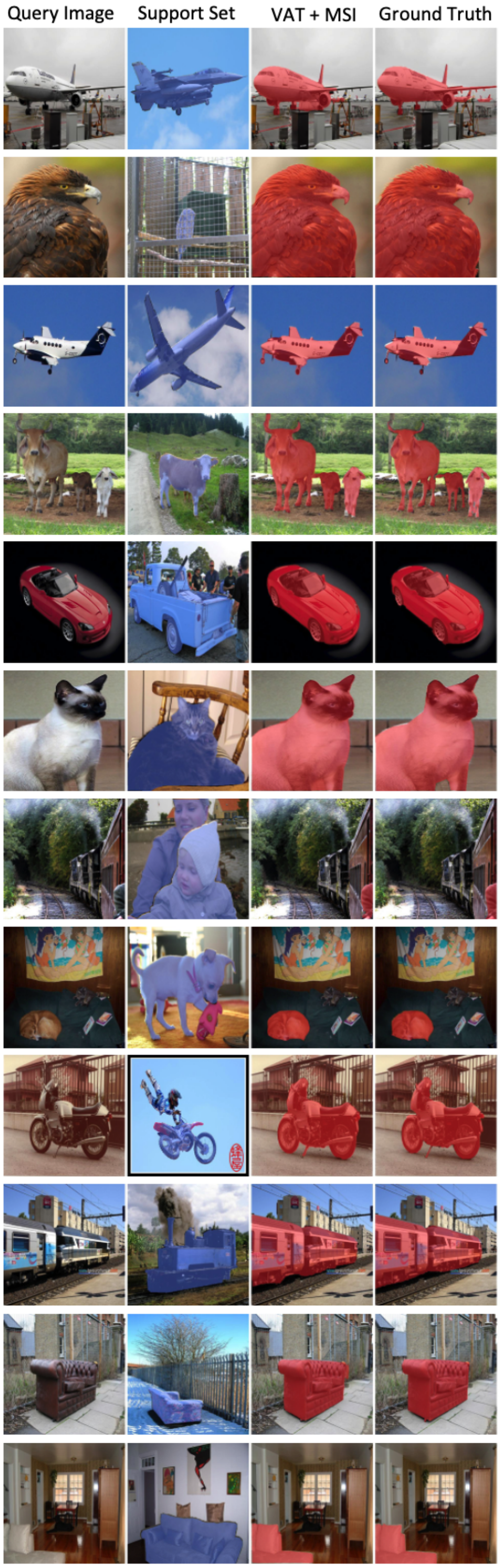

Figure 15. VAT + MSI Qualitative result using ResNet50 [8] on PASCAL-5^i [6].

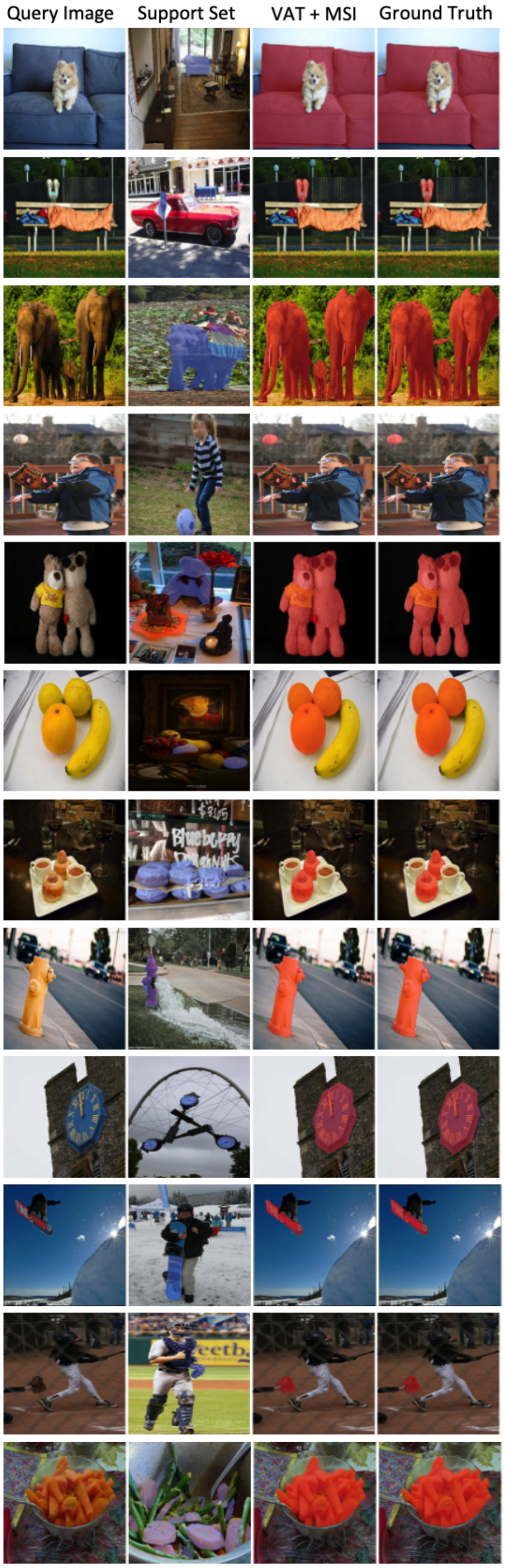

Figure 16. VAT + MSI Qualitative result using ResNet50 [8] on COCO-20^i [16].

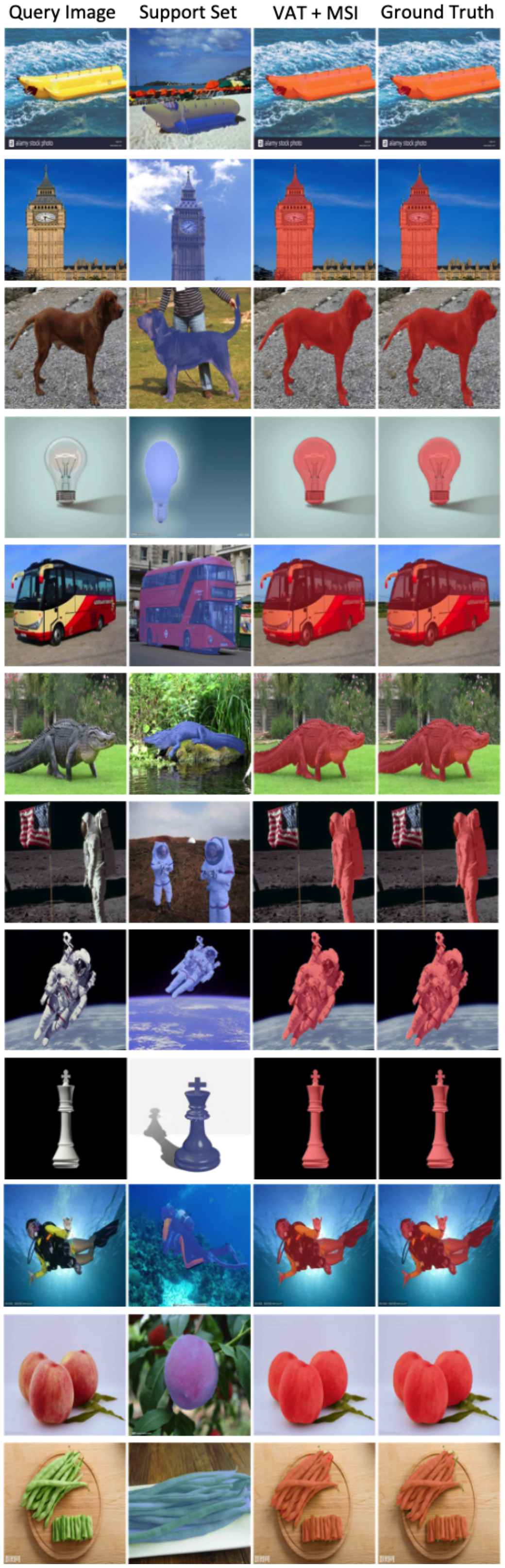

Figure 17. VAT + MSI Qualitative result using ResNet50 [8] on FSS-1000 [15].

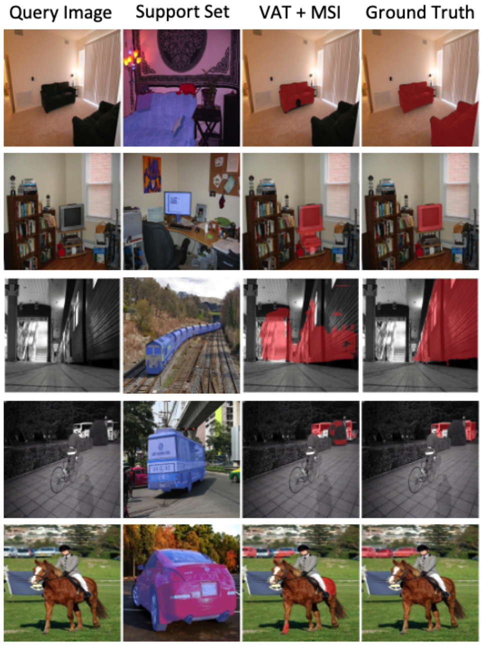

Figure 18. Failure cases of VAT + MSI with ResNet50 [8] on PASCAL-5^i [6].

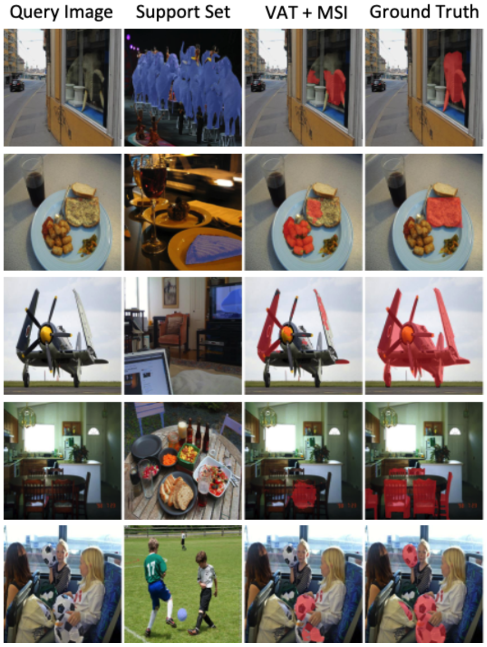

Figure 19. Failure cases of VAT + MSI with ResNet50 [8] on COCO-20^i [16].